

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



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
<b>Methods for objective and subjective assessment of speech quality</b>	<b>Series</b>	<b>P.80</b>
		<b>P.800</b>
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	<a href="http://handle.itu.int/11.1002/1000/12125">11.1002/1000/12125</a>

### 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

	<b>Page</b>
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 Recommendation<sup>1</sup>, 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.

---

<sup>1</sup> 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.

**Table 6-1 – Rating scales used in the ITU-T P.806 subjective testing methodology**

<b>P.MULTI Perceptual Quality Scales</b>		
<b>PQ Scales</b>	<b>Description</b>	<b>Scale Descriptors</b>
<b>S-FLT</b>	Slow-varying degradation in the speech signal	fluttering, babbling, discontinuous
<b>S-RUF</b>	Fast-varying degradation in the speech signal	rough, raspy, harsh
<b>S-LFC</b>	Degradation of low-frequency coloration in the speech signal	dull, muffled, smothered
<b>S-HFC</b>	Degradation of high-frequency coloration in the speech signal	small, distant, thin
<b>B-LVL</b>	Degradation due to the level of background noise	hissing, rushing, roaring
<b>B-VAR</b>	Degradation due to the variability of the background noise	bubbling, intermittent, variable

<b>Overall Scales</b>	<b>Description</b>
<b>LOUD</b>	Overall loudness in the speech signal + background noise
<b>OVRL</b>	Overall quality in the speech signal + background noise

For each of the six PQ rating scales, subjects use a magnitude estimation scale to indicate the amount of the particular PQ that they judge to be present in the sample. Table 6-2 shows an example of the six-category rating scale used by subjects for the PQ rating scales in the ITU-T P.806 test. The bottom category of the scale is labelled 0.0 (zero) to indicate that the specific PQ is not detected in the sample. Furthermore, the numbers associated with the various PQs show one decimal point accuracy (e.g., 0.0, 1.0, ..., 5.0) because the subjects submit their ratings with one decimal point accuracy (e.g., 1.4, 2.7, 4.3).

**Table 6-2 – Magnitude estimation scale used for the PQ ratings**

<b>How would you describe amount of the quality present in the sample?</b>	
Overwhelming	5.0
Somewhat conspicuous	4.0
Very noticeable	3.0
Somewhat noticeable	2.0
Just detectable	1.0
Not detectable	0.0

Table 6-3 shows the scales that are used for the loudness and overall quality ratings. Both of these scales are five-category rating scales with one-decimal point accuracy, i.e., subjects indicate their ratings with one-decimal point accuracy.

**Table 6-3 – Rating scales used for the two overall ratings**

Loudness quality			Overall quality	
Much louder than preferred	5.0		Excellent	5.0
Louder than preferred	4.0		Good	4.0
Preferred	3.0		Fair	3.0
Quieter than preferred	2.0		Poor	2.0
Much quieter than preferred	1.0		Bad	1.0

The following section describes the required methods and procedures for conducting an ITU-T P.806 subjective test.

## 6.1 User interface

Figure 6-1 shows an example user interface used to present the ITU-T P.806 rating scales to the subjects and to collect their responses. In this example, the subject's computer monitor displays the screen shown in Figure 6-1 at the beginning of each trial. The interface includes the computer's mouse and a number of controls and displays on the monitor through which the subject interacts with the system.



**Figure 6-1 – Screen-shot of an example user's interface for the ITU-T P.806 test methodology**

The three large boxes at the top of the screen contain the eight rating scales involved in the test. The box labelled "Perceptual Qualities" contains the six PQ rating scales. Each PQ rating scale is represented by a slider with a range from 0.0 to 5.0 with one decimal-point accuracy. The subject uses the slider to register his or her rating for an individual scale. Subjects enter their ratings on the PQ scales in terms of the amount or degree of the specific quality that he or she perceives is present in the sample, using a rating of 0 = Not detectable, 1 = Just detectable, 2 = Somewhat noticeable, 3=Very noticeable, 4 = Somewhat conspicuous, or 5 = Overwhelming.

Subjects enter a rating for each of the six PQ scales before using the two overall rating scales, loudness and overall quality. The overall scales range from 1.0 to 5.0 with one decimal-point accuracy. Subjects enter loudness ratings as values relative to their own preferred loudness. The overall quality rating scale uses the same scale descriptors as the MOS described in [ITU-T P.800] where 1=Bad, 2=Poor, 3=Fair, 4=Good, 5=Excellent.

## 6.2 Listener instructions and training

ITU-T P.806 is designed to be used for routine testing using naive listeners, rating multiple PQs. All aspects of the test methodology must be standardized so that tests are repeatable, both within an individual subjective test laboratory and across different laboratories throughout the world. One of the key challenges is the specification of training paradigms and instructions for subjects that standardize this training. Instructions and training are meant to ensure that the subject's uncertainty is minimized or, at least, controlled. The ITU-T P.806 methodology is significantly more complicated than any current ITU-T recommended subjective test, so standardized instructions and training are particularly important. They provide listeners with a common baseline as to the nature of the task(s) and are crucial to ensuring that subjects understand exactly what they should be rating for a particular rating scale. It is recommended to use a recorded presentation to provide instructions and training for the ITU-T P.806 test. In addition to providing instructions for taking the test, the presentation should provide example conditions that have been judged to exhibit various degrees of the six PQs involved in the ITU-T P.806 test methodology. An example set of written instructions in English is shown in Appendix I. An example set of written instructions in French is shown in Appendix II.

## 6.3 Designing an ITU-T P.806 test

As indicated in the Scope of this Recommendation, the ITU-T P.806 test methodology has few limitations on the nature and number of test conditions that may be validly and reliably tested in an experiment. There are practical limitations due to the complexity of the task for subjects and the total testing time of an experiment for each subject. Practice has shown that a maximum of around 200 ITU-T P.806 trials, or about three hours of total testing time, is reasonable for a subjective test [b-ITU-T Testing]. The number of trials for a test is a function of the number of test-conditions and the number of talkers involved in the test, where  $\# \text{ trials} = (\# \text{ conditions}) \times (\# \text{ talkers})$ .

### 6.3.1 Talkers

The precision of the results for an ITU-T P.806 test for any given condition increases with increases in the number of talkers since the total # votes/condition is a function of the # listeners involved in the test and the number of talkers per condition. While an increase in the number of talkers will increase the precision for an individual condition, it also reduces the number of conditions that may be evaluated by a subject. Therefore, there is a trade-off between the # talkers and the # subjects to achieve a desired level of precision in the test. A minimum of four talkers (two males and two females) is recommended to provide reasonable sampling of talkers [b-ITU-T Testing].

### 6.3.2 Reference-conditions

Practice has shown that standard reference-conditions should be included in subjective tests (for example, [ITU-T P.800] and [ITU-T P.835]) to provide subjects with a common frame-of-reference within the test for rating the test-conditions. This is particularly important in ITU-T P.806 where subjects are required to rate speech samples on multiple PQ dimensions<sup>2</sup>. ITU-T P.806 introduces the concept of *Exemplar-conditions*, conditions that exhibit a high degree of degradation in a single PQ

---

<sup>2</sup> [ITU-T P.800] tests typically use modulated noise reference unit (MNRU) processing to provide a frame of reference in the overall quality dimension and [ITU-T P.835] uses a combination of MNRU and background noise signal-to-noise ratio (SNR) to provide frames of reference in two dimensions – speech-distortion and noise-intrusiveness.

(i.e., a "pure" exemplar of that PQ). Each ITU-T P.806 test should include one *Exemplar-condition* for each of the six PQs described in Table 6-1 above. These exemplar conditions are designed to provide the subjects with at least one condition within every ITU-T P.806 test where a "pure" example of impairment in the specific PQ is exhibited, i.e., the specific PQ for which the condition is an "exemplar" is the dominant impairment and the other five PQs are absent or relatively less evident. This procedure provides subjects with an opportunity to exercise every PQ scale in every ITU-T P.806 test, regardless of the range or context of the impairments present among the test-conditions. In addition to the six exemplar-conditions, it has become standard practice to include source-conditions among the reference-conditions as an example of a condition with no degradation in any of the six PQs. For tests involving multiple bandwidths, source-conditions are typically included for more than one bandwidth, e.g., superwideband (SWB) and narrowband (NB) source-conditions. Table 6-4 shows example descriptions of the processing that could be used for the six exemplar conditions (R03-R08) and for the two source-conditions (R01 and R02). It is recommended that the reference-conditions (i.e., source-conditions plus exemplar-conditions) be used in the training phase of a test (see clause 6.4.2)

**Table 6-4 – Description**

Cond	Exemplar for	Description of processing
R03	S-FLT	3GPP AMR-WB at 12.65 kbit/s, 8% FER, no DTX
R04	S-RUF	SWB p.50 MNRU at 5 dB
R05	S-LFC	Low-Pass filtering at 500 Hz
R06	S-HFC	Band-Pass filtering 1 kHz-14 kHz
R07	B-LVL	AMR-NB at 12.2 kbit/s + Car noise at 0 dB SNR
R08	B-VAR	AMR-NB at 12.2 kbit/s + Pub noise at 0 dB SNR

Cond.	Source for	Description of processing
R01	SWB	Low-Pass filtering at 14 kHz
R02	NB	Low-Pass filtering at 3.5 kHz

## 6.4 Organization of an ITU-T P.806 test session

For each test subject, an ITU-T P.806 test requires three distinct phases:

- orientation and instruction phase
- preliminary or training phase
- testing phase, including multiple test sessions separated by rest-breaks.

### 6.4.1 Orientation and instruction phase

Listener instructions for the ITU-T P.806 task should be provided to subjects in a standard form, in either written or audiovisual format, so that all subjects have the same baseline for the nature of the task and what is expected from them. Presentation of the standard instructions should provide subjects with all of the information necessary to complete the test without further interaction with or ad hoc instructions from the experimenter. It is preferable that instructions are automated and interactive in order to remove any possible source of confounding of results due to different instruction. An example of written instructions is presented in Appendix I.

At the end of the instructions phase subjects may ask questions to the test administrator, but answers should be limited to guidance in the use of the subject interface and the process of playing samples and entering responses. Under no circumstances should the test administrator offer advice on how the subject should scale his responses on a particular rating scale or what rating should be given to an example condition on any particular scale.

#### **6.4.2 Preliminary or training phase**

During the preliminary or training phase, subjects should be given a standard set of test conditions to practice using the subject interface for listening to the test materials and entering ratings on the rating scales. There should be sufficient training trials to ensure that subjects understand the task and can proceed with the test sessions without further interaction with the test administrator.

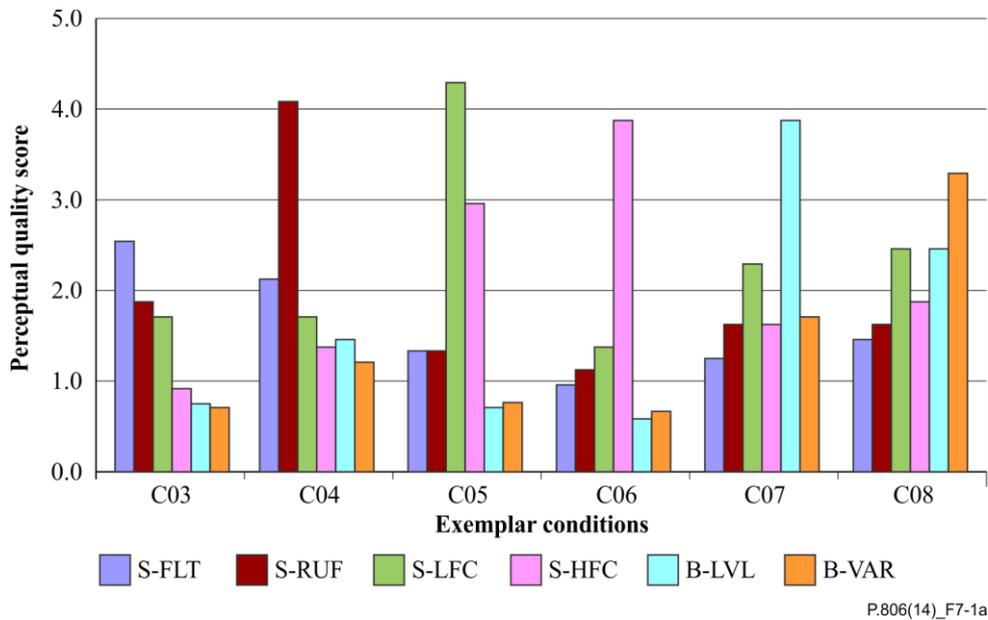
The set of training conditions should give subjects the opportunity to exercise all six of the PQ rating scales and it has been found useful to use the eight reference-conditions described above in clause 6.3.2. Subjects should not be informed that the training or practice conditions will be included during the actual testing sessions.

#### **6.4.3 Testing phase**

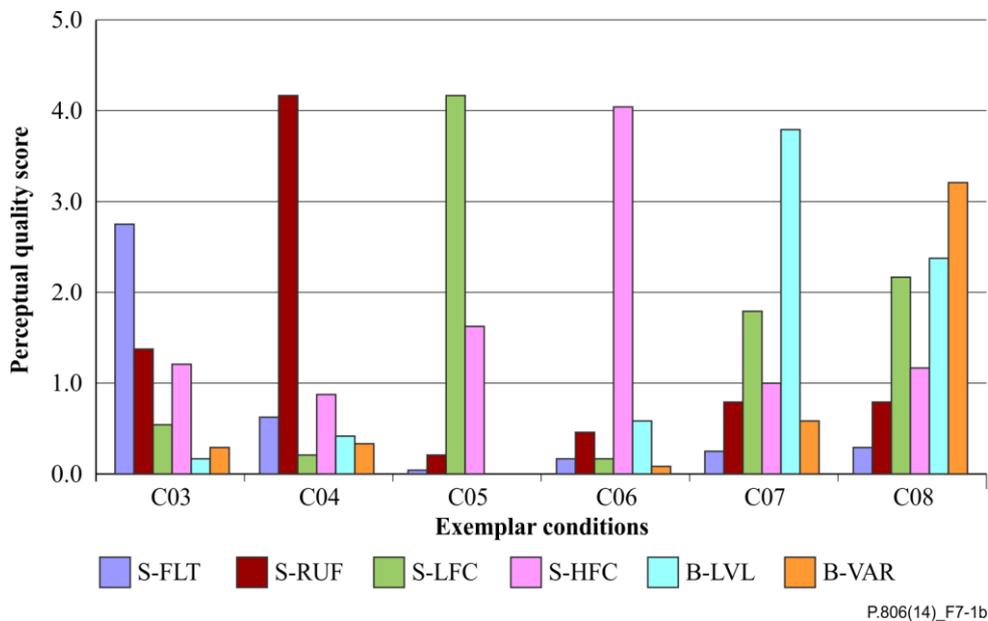
The ITU-T P.806 test methodology requires naive subjects to make ratings on multiple scales where each scale assesses a different PQ dimension of the test stimulus. It places a much higher cognitive load on the subjects and great care should be taken to reduce the effects of fatigue or boredom. To this end, the testing phase should be separated into manageable sub-sessions with mandatory rest-breaks between sub-sessions. Experience has shown that the sub-sessions should be 15 to 20 minutes in duration. Rest-breaks can be of variable duration but should be a minimum of five min. The maximum duration of the total test should be around three hours per subject.

### **7 ITU-T P.806 test results**

The ITU-T P.806 test methodology provides subjective scores for overall quality (OVRL), loudness (LOUD) plus individual scores for six PQs. The pattern of PQ scores provides the test user with a profile of quality degradation scores that may be used to diagnose specific impairments underlying the overall quality score obtained for the test-condition. Figure 7-1a shows an example of PQ profiles for six exemplar-conditions derived from an ITU-T P.806 test conducted in a listening lab in the United States using an English speech database and native-speaking English subjects [b-C110]. Figure 7-1b shows an example of PQ profiles for the same six exemplar-conditions derived from an ITU-T P.806 test conducted in a listening lab in France using a French speech database and native-speaking French subjects. In both figures, the scores are based on the ratings of 16 naive subjects and four talkers (i.e., # votes = 64). The two figures illustrate that subjects at different listening labs in different countries using a different language show a high degree of agreement on the PQ scores for a common set of exemplar conditions. For illustrative purposes, Figures 7-1a and 7-1b only show scores for the PQs but there was very good agreement for the OVRL score as well. Table 7-1 shows the correlation between English and French scores for the sets of 56 conditions (8 reference- and 48 test-conditions).



**Figure 7-1a – PQ Score profiles of an ITU-T P.806 test in English**



**Figure 7-1b – PQ Score profiles of an ITU-T P.806 test in French**

**Table 7-1 – Correlation between ITU-T P.806 scores for 56 conditions tested in two languages by different listening labs**

S-FLT	S-RUF	S-LFC	S-HFC	B-LVL	B-VAR	OVRL
0.870	0.862	0.839	0.764	0.959	0.934	0.940

Typical results for an ITU-T P.806 test should include means and standard deviations for each condition in the test computed over the number of votes per condition (# subjects × # talkers).

Additional statistical analyses could include comparison of pairs of test-conditions by student's t-test or evaluation of experimental factors by analysis of variance.

Procedures and guidelines for the evaluation of individual subject's performance are included in the ITU-T Handbook "Practical procedures for subjective testing" [b-ITU-T Testing]. These procedures should be used to determine if the data for subjects who are unable or unwilling to reliably perform the tasks required in ITU-T P.806 should be removed. Data for such outliers and/or malingerers should be replaced with data from reliable and capable subjects.

## Appendix I

### ITU-T P.806 test instructions in English

(This appendix does not form an integral part of this Recommendation.)

The instructions presented in this appendix are an example of a set of printed instructions.

In this test you will be using a number of PQ rating scales to describe samples of speech that have been processed by various voice communications systems. In each trial in the test you will be presented with a 20 s audio sample which includes five sentences spoken by various talkers. You will now hear an example of the 20 s audio sample.

< subject listens to an audio sample >

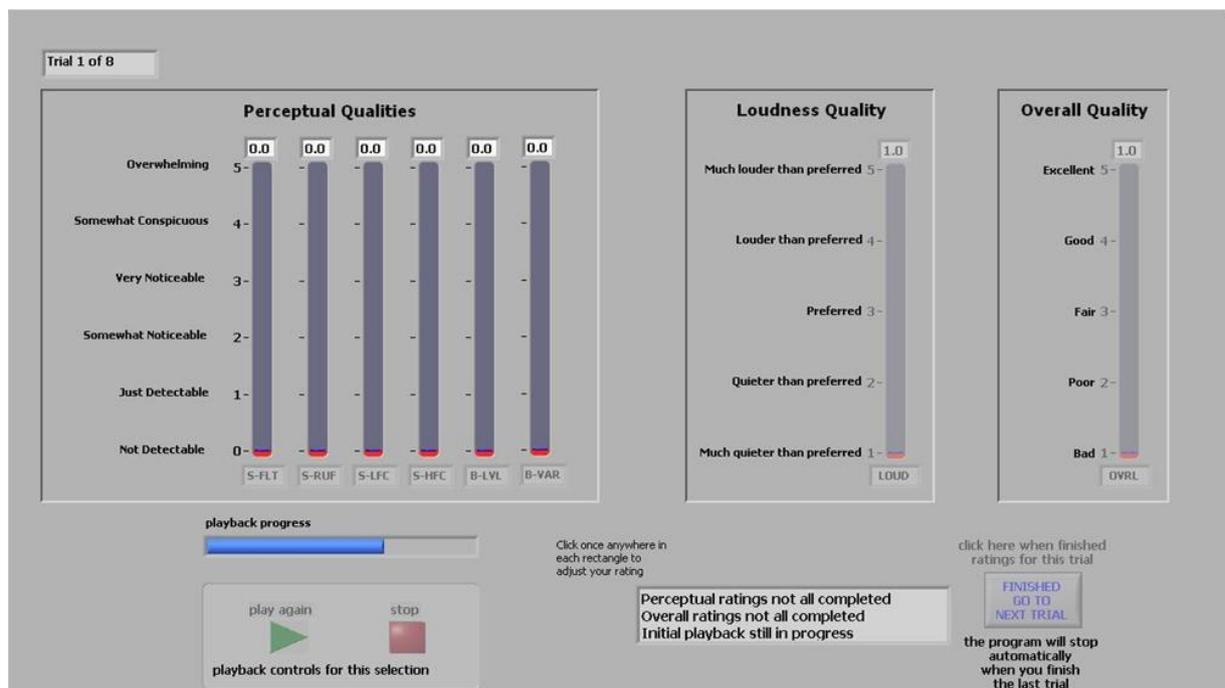
(Refer to <sample>.wav in the electronic attachment to this Recommendation.)

In each trial you will use eight scales to rate the speech sample. Six of these scales are PQ scales where:

- for four of the PQ scales, you will concentrate on the **speech signal** alone while making your ratings: rating scales **S-FLT**, **S-RUF**, **S-LFC** and **S-HFC**;
- for two of the PQ scales, you will concentrate on the **background** alone while making your ratings: rating scales **B-LVL** and **B-VAR**.

For the remaining two scales, you will consider the **overall sample** (the speech signal and the background) while making your ratings: rating scales **LOUD** and **OVRL**.

The rating scales themselves will be described later in these instructions, but first is a description of the computer interface that you will use to enter your ratings. At the beginning of each trial, the screen shown below will appear on your computer monitor.



Carefully review the screen-shot shown above to become familiar with the layout and the controls. The three large boxes in the middle of the screen contain the eight rating scales that you will use to rate the sample. The scales for six PQs are grouped together and must be rated before the two remaining scales become active.

Each rating scale includes:

1. a scale descriptor (for example, S-FLT)
2. labels for the numbered rating categories (e.g., Not Detectable-0, Just Detectable-1)
3. a slider which you will use to enter a rating between 0.0 and 5.0 with one decimal-point accuracy.

At the top of each slider is a box which will indicate the rating you have entered.

At the beginning of a trial, the speech sample will start playing and continue for the entire 20 s sample. You will not be able to access any rating scales until the sample has played for 4 s. At that point, the rating scales in the PQs box will become active. You can access the six PQ scales in any order, but we recommend that you rate all of the scales within one category (either the four scales within the speech signal category or the two scales within the background category) before moving on to the other category. If the sample finishes playing before you have completed your ratings, you can always play the sample again from the beginning. After you make a rating on all six PQ scales you should rate the two overall scales (Loudness Quality and Overall Quality) which will complete the trial.

For four of the PQ rating scales, you are asked to consider only the speech signal while making your ratings. These scales are labelled **S-FLT**, **S-RUF**, **S-LFC** and **S-HFC**. Each of the four PQ scales for the speech signal is described below. You will listen to audio samples that have been judged to exhibit this quality.

- **S-FLT** – the label for this scale is an abbreviation for Signal-Fluttering. Descriptor terms for this scale are *fluttering, babbling, discontinuous*. Use this scale to rate the amount or degree of this quality present in the speech signal of the current sample. You will now hear examples of conditions that have been judged to exhibit the quality described as fluttering, babbling, discontinuous.

**<subject listens to the three audio samples for S-FLT>**

*(Refer to s-flt.wav in the electronic attachment to this Recommendation.)*

- **S-RUF** – the label for this scale is an abbreviation for Signal-Rough. Descriptor terms for this scale are *rough, raspy, harsh*. Use this scale to rate the amount or degree of this quality present in the speech signal of the current sample. You will now hear examples of conditions that have been judged to exhibit the quality described as rough, raspy, harsh.

**<subject listens to the three audio samples for S-RUF>**

*(Refer to s-ruf.wav in the electronic attachment to this Recommendation.)*

- **S-LFC** – the label for this scale is an abbreviation for Signal-Low Frequency Coloration. Descriptor terms for this scale are *dull, muffled, smothered*. Use this scale to rate the amount or degree of this quality present in the speech signal of the current sample. You will now hear examples of conditions that have been judged to exhibit the quality described as dull, muffled, smothered.

**<subject listens to the three audio samples for S-LFC>**

*(Refer to s-lfc.wav in the electronic attachment to this Recommendation.)*

- **S-HFC** – the label for this scale is an abbreviation for Signal-High Frequency Coloration. Descriptor terms for this scale are *small, distant, thin*. Use this scale to rate the amount or degree of this quality present in the speech signal of the current sample. You will now hear examples of conditions that have been judged to exhibit the quality described as small, distant, thin.

**<subject listens to the three audio samples for S-HFC>**

*(Refer to s-hfc.wav in the electronic attachment to this Recommendation.)*

For two of the PQ rating scales, you are asked to consider only the Background in the sample while making your ratings. These scales are labelled B-LVL and B-VAR. When you place the cursor within the slider box for one of these scales, a text box will pop up showing a list of the descriptor terms that apply to that scale. Each of the two PQ scales for the Background is described below along with example conditions that have been judged to exhibit this quality.

- **B-LVL** – the label for this scale is an abbreviation for Background-Level. Descriptor Terms for this scale are *hissing, rushing, roaring*. Use this scale to rate the amount or degree of this quality present in the Background of the current sample. You will now hear examples of conditions that have been judged to exhibit the quality described as hissing, rushing, roaring.

**<subject listens to the three audio samples for B-LVL>**

*(Refer to b-lvl.wav in the electronic attachment to this Recommendation.)*

- **B-VAR** – the label for this scale is an abbreviation for Background-Variability. Descriptor Terms for this scale are *bubbling, intermittent, variable*. Use this scale to rate the amount or degree of this quality present in the Background of the current sample. You will now hear examples of conditions that have been judged to exhibit the quality described as bubbling, intermittent, variable.

**<subject listens to the three audio samples for B-VAR>**

*(Refer to b-var.wav in the electronic attachment to this Recommendation.)*

After you have entered ratings for the six PQ scales you will rate the two overall scales. For these two scales you are asked to consider the overall sample (both speech signal and background) while making your ratings. These two scales are labelled **LOUD** and **OVRL**. The description for each scale is included completely with the boxes displayed on the screen. Note also that the ratings for the overall scales are between 1.0 and 5.0 with one decimal-point accuracy.

- the label **LOUD** is an abbreviation for *Loudness Quality*. You should use this scale to rate the loudness of the sample relative to your preferred loudness;
- the label **OVRL** is an abbreviation for *Overall Quality*. You should use this scale to rate your assessment of the overall quality of the sample.

The test itself will involve X test sessions of Y trials each. At the end of each block of Z trials you will be required to take a break for a minimum of five minutes. You should use this time to leave the booth, stretch your legs, get a drink of water, go to the restroom, or whatever you need to do to help ease fatigue and boredom. The test is self-paced so you can also take additional breaks within a test block if you wish.

You are now ready to start the test. But before you do, you will get an opportunity to practice the rating task in a short session of eight trials.

If you have any questions please ask them to the test administrator before you begin the practice session.

We want to take this opportunity to thank you in advance for your participation in this very important experiment.

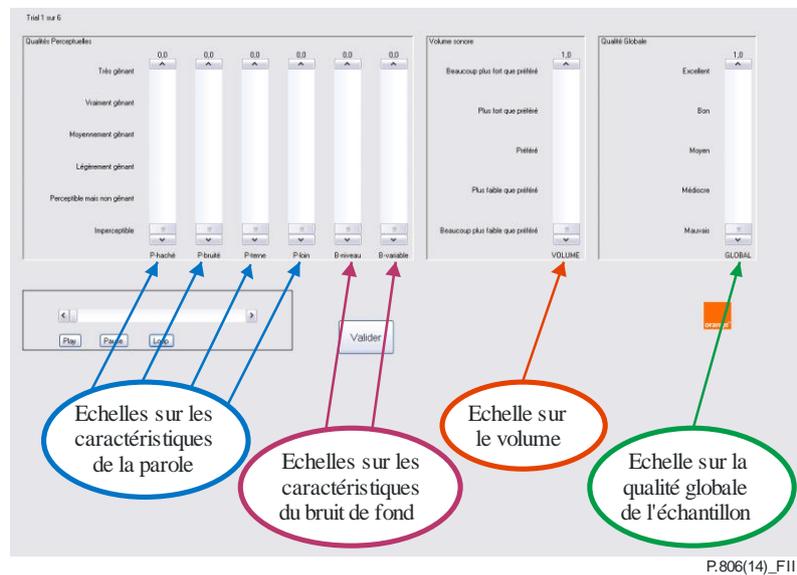
## Appendix II

### ITU-T P.806 test instructions in French

(This appendix does not form an integral part of this Recommendation.)

#### Test multi-échelles

Dans ce test, vous allez utiliser 8 échelles différentes pour décrire les échantillons de parole. Les échantillons de parole que vous allez écouter sont issus de différents systèmes de communications, les échantillons sont d'une durée de 20 secondes et sont composés de 5 phrases.



Pour quatre échelles, vous allez vous concentrer sur les caractéristiques de la parole:

- discontinu, fluctuant
- rauque, bruyé
- assourdi, étouffé, terne
- lointain, faible, voix de robot.

Pour deux échelles, vous vous concentrerez sur les caractéristiques du bruit de fond:

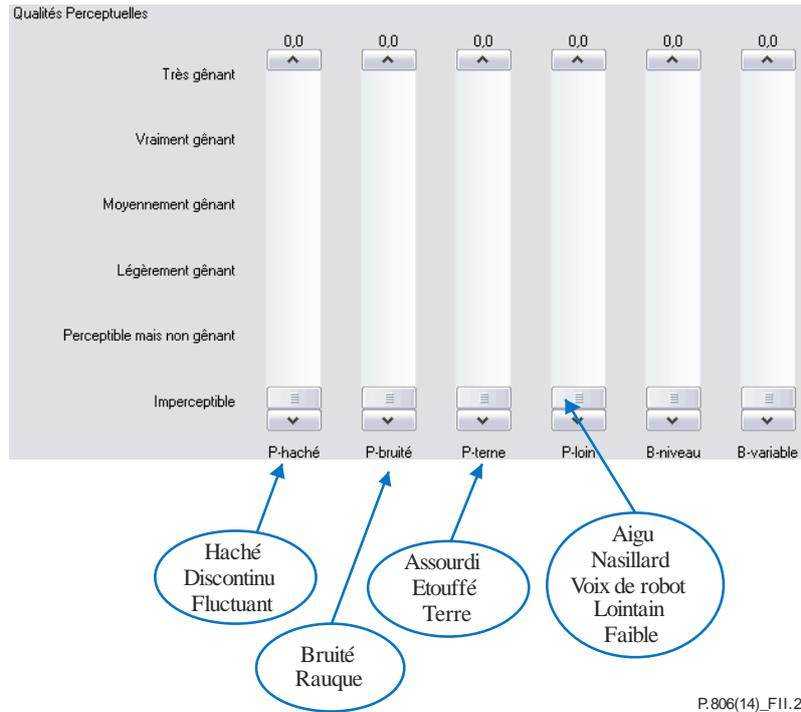
- sifflant, bruyant
- intermittent, variable.

Pour les deux dernières, vous considérez la parole et le bruit de fond pour donner votre note sur:

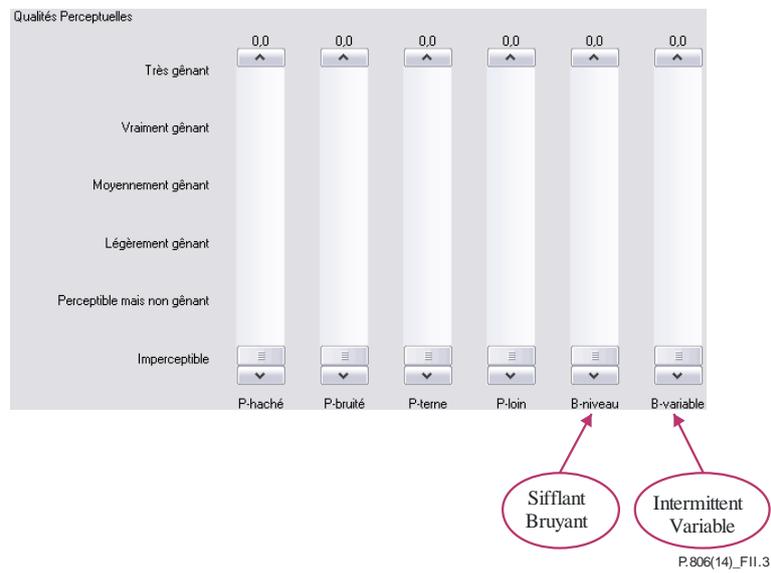
- le niveau sonore (le volume)
- la qualité globale.

Vous pouvez réécouter autant de fois que nécessaire pour décider de la note pour chacune des échelles.

## Caractéristiques de la parole



## Caractéristiques du bruit de fond



### Le niveau sonore (le volume)

Volume sonore

1,0

Beaucoup plus fort que préféré

Plus fort que préféré

Préféré

Plus faible que préféré

Beaucoup plus faible que préféré

VOLUME

### La qualité globale

Qualité globale

1,0

Excellent

Bon

Moyen

Médiocre

Mauvais

GLOBAL

P.806(14)\_F11.4

## Bibliography

- [b-ITU-T Testing] ITU-T Handbook (2011), *Practical procedures for subjective testing*.
- [b-C110] ITU-T T13-SG12-C0110 (2013), *Comparison of Results from P.MULTI testing in two Languages*.
- [b-C228] ITU-T T09-SG12-C0228 (2011), *P.MULTI subjective testing methodology – number of relevant Perceptual Quality scales and additional validation of the test methodology*.
- [b-AH-11-028] AH-11-028, P.MULTI (2011), *A proposed methodology and pilot test*. Q7/12 Rapporteur's meeting, 20-21 June.
- [b-Sen] Sen, D. & Lu, W., *Objective evaluation of speech signal quality by the prediction of multiple foreground diagnostic acceptability measure attributes*, J. Acoust. Soc. Am. (JASA), Volume 131, Issue 5, pp. 4087-4103 (2012); (17 pages).
- [b-Voiers] W.D. Voiers (1977), *Diagnostic acceptability measure for speech communication systems*, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'77.



## SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
<b>Series P</b>	<b>Terminals and subjective and objective assessment methods</b>
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects and next-generation networks
Series Z	Languages and general software aspects for telecommunication systems