

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.863.2

(07/2022)

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

Methods for objective and subjective assessment of
speech and video quality

**Extension of ITU-T P.863 for multi-dimensional
assessment of degradations in telephony
speech signals up to fullband**

Recommendation ITU-T P.863.2

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

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

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

Recommendation ITU-T P.863.2

Extension of ITU-T P.863 for multi-dimensional assessment of degradations in telephony speech signals up to fullband

Summary

Recommendation ITU-T P.863.2 describes a set of models for predicting perceptual dimensions of degradations linked to the overall speech quality from narrowband (300 to 3 400 Hz) to fullband (20 to 20 000 Hz) telecommunication scenarios. The predictions target user judgements on four perceptual dimensions, as obtained in a subjective test described in an annex.

The models described in Recommendation ITU-T P.863.2 are partially based on internal parameters of the model given in Recommendation ITU-T P.863. Recommendation ITU-T P.863.2 presents a detailed description of all model parts that are not contained in Recommendation ITU-T P.863. A conformity testing procedure is also specified in an annex to allow a user to validate whether an alternative implementation of the models is correct.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.863.2	2022-07-29	12	11.1002/1000/15010

Keywords

Directness, discontinuity, loudness, noisiness, perceptual quality dimensions, quality prediction, spectral colouration.

* 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/software copyrights, 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 appropriate ITU-T databases available via the ITU-T website at <http://www.itu.int/ITU-T/ipr/>.

© ITU 2022

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.....	2
3	Definitions	3
	3.1 Terms defined elsewhere	3
4	Abbreviations and acronyms	3
5	Conventions	4
6	Overview of the models.....	4
	6.1 Model characteristics.....	4
7	Comparison between objective and subjective scores.....	6
8	Speech material.....	6
	8.1 Input or reference speech material	6
	8.2 Degraded speech material.....	7
	8.3 Special requirements for acoustically captured speech material	7
	8.4 Acoustical insertion or capture for loudspeaker phones.....	8
9	Description of the model algorithms	8
	9.1 Colouration model	8
	9.2 Discontinuity model	9
	9.3 Noisiness model.....	9
	9.4 Sub-optimum loudness model	11
	Annex A – Subjective test method for obtaining perceptual dimension scores.....	13
	Annex B – Conformity data and tests	16
	B.1 List of files provided for conformity validation	16
	B.2 Conformity tests	16
	B.3 Digital attachments.....	17
	Appendix I – Reporting of the performance results for the model algorithms based on the correlation, RMSE and RMSE* metrics.....	18
	Appendix II – Test instructions.....	19
	Bibliography.....	23

Electronic attachment: Conformity data described in Annex B.

Introduction

This Recommendation describes a set of four models predicting perceptual dimensions of degradations linked to the overall speech quality in narrowband to fullband telecommunication scenarios. In contrast to the model described in [ITU-T P.863], the aim is not to predict a one-dimensional score for the overall quality, but rather to decompose the perceived quality into four perceptual dimensions, termed colouration, discontinuity, noisiness and sub-optimum loudness. The targeted dimensions reflect user judgements obtained in listening-only tests carried out in accordance with Annex A; in contrast to the seven dimensions listed in [ITU-T P.806], four perceptual dimensions are targeted here.

The models partially make use of internal parameters of the ITU-T P.863 model; thus, the implementation of the models should be performed in accordance with that Recommendation.

Recommendation ITU-T P.863.2

Extension of ITU-T P.863 for multi-dimensional assessment of degradations in telephony speech signals up to fullband¹

1 Scope

This Recommendation describes models whose purpose is to predict perceptual dimensions of degradations linked to the overall speech quality in narrowband (NB) to fullband (FB) telecommunication scenarios. The models provide more detailed information about individual quality dimensions as additional information to the ITU-T P.863 overall mean opinion score (MOS). Perceptual dimensions of degradations may originate from all speech processing components usually considered for telecommunications in clean and noisy conditions. The models predict these dimensions as they are assessed in a listening-only test context, in accordance with Annex A.

In contrast to [ITU-T P.863], the models described in this Recommendation show only one operational mode, in which degraded speech samples are scored against an FB reference signal and predict the perceptual dimension scores on a corresponding scale. The models provide an estimation of the colouration, discontinuity, noisiness and sub-optimum loudness of the degraded speech sample. These four dimensions are not identical to the seven dimensions listed in [ITU-T P.806]. Instead, the four dimensions predicted by the models described in this Recommendation target subjective judgements obtained in a test carried out according to Annex A.

The term telecommunication scenario mentioned in the first paragraph covers all transmission technologies in current:

- public switched networks (e.g., fixed wire public switched telephone network (PSTN), global system for mobile communications, wideband (WB) code division multiple access, code division multiple access (CDMA), voice over long-term evolution and voice over new radio);
- push-over-cellular, voice over Internet protocol (VoIP) and PSTN-to-VoIP interconnections, terrestrial trunked radio; and
- commonly used speech processing components (e.g., coder–decoders (codecs), noise reduction systems, adaptive gain control, comfort noise and other types of voice enhancement devices) and their combinations.

In addition to the commonly used ITU-T and ETSI speech codecs, other coding technologies, as specified by the 3rd Generation Partnership Project 2 and used in CDMA networks, have been considered in the training and selection data. Furthermore, codecs used in broadcasting services with speech-based contents have also been taken into account, e.g., Moving Picture Experts Group-1 audio layer 3 (MP3) or advanced audio coding.

Other technologies or components such as speech storage formats or non-telephony applications such as public safety networks or professional mobile radio connections have not been assessed for the described models, and thus lie outside the scope of this Recommendation.

Tables 1 to 4 of [ITU-T P.863] list test factors, coding technologies and applications to which that Recommendation applies, either in the sense that they have been included in the requirement specification and have been tested accordingly, that they are not intended to be used with, or that further investigation or validation is necessary. Unless specified otherwise in this Recommendation, the limitations in [ITU-T P.863] also apply to the models specified in this Recommendation, as they are partially based on internal parameters of the ITU-T P.863 model.

¹ This Recommendation includes an electronic attachment with the conformity data described in Annex B.

The consideration of the acoustical path to and from (acoustical insertion and acoustical capturing) an actually used terminal may affect the colouration or noisiness of the degraded signal, and is foreseen by the described model. The score targeted by the prediction of the model stems from a diotical presentation of a mono-signal, meaning that the same signal is played out at each ear in the listening context that the model tries to predict.

Dimensions of speech quality that cannot be assessed in a listening-only context, such as conversational aspects and talking quality, lie outside the scope of this Recommendation. The described model considers noises and their influence on perceptual quality dimensions in a listening-only context similar to the one described in [ITU-T P.800] (test cabinet specifications, etc.). The prediction of quality as it can be perceived in a noisy listening environment and the related binaural effects lie outside the scope of this Recommendation.

Non-steady, fluctuating noises can be seen as degradations on the discontinuity scale.

NOTE – Examples of non-steady, fluctuating noises are footsteps and beeps as from an alarm clock. While a human listener can recognize those noises as natural, the models described in this Recommendation recognizes them as interrupted and counts them as degradations on the discontinuity scale. This is an immanent problem of the models described in this Recommendation that have no knowledge about the original background noise.

As is the case for [b-ITU-T P.862] and [ITU-T P.863], the approach of the models described in this Recommendation is called "full-reference" or "double-ended", which means that the quality prediction is based on the comparison between an undistorted reference signal and the received signal to be scored.

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 O.41] Recommendation ITU-T O.41 (1994), *Psophometer for use on telephone-type circuits*.
- [ITU-T P.10] Recommendation ITU-T P.10/G.100 (2017), *Vocabulary for performance, quality of service and quality of experience*.
- [ITU-T P.50] Recommendation ITU-T P.50 (1999), *Artificial voices*.
- [ITU-T P.56] Recommendation ITU-T P.56 (2011), *Objective measurement of active speech level*.
- [ITU-T P.340] Recommendation ITU-T P.340 (2000), *Transmission characteristics and speech quality parameters of hands-free terminals*.
- [ITU-T P.581] Recommendation ITU-T P.581 (2022), *Use of head and torso simulator for hands-free and handset terminal testing*.
- [ITU-T P.800] Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.
- [ITU-T P.806] Recommendation ITU-T P.806 (2014), *A subjective quality test methodology using multiple rating scales*.
- [ITU-T P.830] Recommendation ITU-T P.830 (1996), *Subjective performance assessment of telephone-band and wideband digital codecs*.

- [ITU-T P.851] Recommendation ITU-T P.851 (2003), *Subjective quality evaluation of telephone services based on spoken dialogue systems*.
- [ITU-T P.862.3] Recommendation ITU-T P.862.3 (2007), *Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2*.
- [ITU-T P.863] Recommendation ITU-T P.863 (2018), *Perceptual objective listening quality prediction*.
- [ITU-T P.863.1] Recommendation ITU-T P.863.1 (2019), *Application guide for Recommendation ITU-T P.863*.

3 Definitions

3.1 Terms defined elsewhere

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

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR	Absolute Category Rating
ASL	Active Speech Level
CDMA	Code Division Multiple Access
FB	Fullband
HATS	Head And Torso Simulator
MOS	Mean Opinion Score
MOS-C	Mean Opinion Score- Colouration
MOS-D	Mean Opinion Score-Discontinuity
MOS-L	Mean Opinion Score-sub-optimum Loudness
MOS-N	Mean Opinion Score-Noisiness
MP3	Moving Picture Experts Group-1 audio layer 3
NB	Narrowband
PCM	Pulse Code Modulation
PSD	Power Spectral Density
PSTN	Public Switched Telephone Network
RMSE	Root Mean Square Error
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SWB	Super-wideband
VoIP	Voice over Internet Protocol
WAVE	Wireless Access in Vehicular Environments
WB	Wideband

5 Conventions

None.

6 Overview of the models

The models described in this Recommendation provide estimations of individual perceptual dimensions, which allow diagnostic information to be obtained from speech signals. The four dimensions covered by the models of this Recommendation are:

- 1) colouration (i.e., resulting from frequency response distortions, e.g., bandwidth restrictions and colouration introduced by transducers);
- 2) noisiness (e.g., resulting from additive and multiplicative noise);
- 3) discontinuity (e.g., resulting from time localized and time-varying degradations);
- 4) sub-optimum loudness (e.g., impact of overall play back level).

Each perceptual dimension is related to – but not congruent with – technical causes that might lead to the respective perceptual effect. For example, discontinuity may result from packet losses, but also from time-varying background noises. Similarly, colouration might be caused by a narrowband (NB) codec, but also stem from the send-side or receive-side terminal device.

The four dimensions have been extracted from subjective tests with the aim to cover most degradations present in current and near future telecommunication scenarios, see [b-Wältermann]. By performing two types of multidimensional analyses (similarity judgements with subsequent multidimensional scaling, as well as semantic differential scaling with subsequent principal component analysis), a set of three perceptual dimensions (colouration, discontinuity and noisiness), which have proven to be rather orthogonal to each other, has been extracted. As the corresponding perceptual experiments have been carried out mostly with level-equalized signals, a fourth dimension, sub-optimum loudness, has been added later to account for degradations that stem from a non-optimum listening level. This dimension is not expected to be orthogonal to the other three.

The model for estimating colouration is derived from two ITU-T P.863 indicators, one of frequency combined with a second that quantifies the overall disturbance in the time-frequency plane.

The model for estimating noisiness is based on two indicators that assess the presence of noise in the degraded signal and a third indicator that models the impact of the active speech level (ASL) on the perceived noisiness. The first indicator quantifies the amount of background noise in the degraded signal, the second evaluates the presence of noise on the active speech parts by means of spectral entropy, and the third weights the impact of the previous two indicators on the perceived noisiness according to the deviation of the ASL from the nominal level (–26 dBov).

The model for estimating discontinuity is a combination of three ITU-T P.863 indicators, two of noisiness and another that quantifies the overall added part of the disturbance in the time-frequency plane.

The model for estimating sub-optimum loudness is based mainly on the power in active speech parts in the degraded signal. This information is combined with a second indicator that quantifies the presence of gain variations and models their negative impact on the perceived optimal loudness.

6.1 Model characteristics

6.1.1 Input signal characteristics

Input signals to the models are standard pulse code modulation (PCM) 16-bit linear, Intel byte order, raw format speech files. A sampling frequency of 48 kHz for reference and captured signals is required. Both the reference and the degraded signal have to be mono signals.