

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.

Acoustical recordings have to be done using a head and torso simulator (HATS) with diffuse-field equalization, as specified in Table 1.

Table 1 – Diffuse-field equalization for acoustical recordings

Test condition	Recording	Reference signal	Captured degraded signal	Signal presentation in LOT	Degraded signal for the models
Super-wideband	<i>Electr.</i>	48 kHz mono	48 kHz mono	48 kHz mono over both ears	48 kHz mono
Super-wideband narrowband	<i>Acoust.</i>	48 kHz mono	48 kHz mono	48 kHz mono over both ears	48 kHz mono
Narrowband	<i>Electr.</i>	48 kHz mono	8, 16, 48 kHz mono	48 kHz mono over both ears	48 kHz mono
Wideband	<i>Electr.</i>	48 kHz mono	16 kHz mono (48 kHz mono)	48 kHz mono over both ears	48 kHz mono

The measuring interface for the models of this Recommendation is either an electrical network termination point (or an equivalent) or the acoustical interface of the terminal using an artificial ear.

The receiving terminal is considered to have a flat frequency response. Therefore, for electrically captured signals, the reference receiving model might be a flat response only. Since the acoustical (equalized) captured signal can also be assumed to be "flat", a different input filter for the model between acoustical and electrical recorded material is not necessary. Consequently, the models do not have a switch between electrical and acoustical recordings.

6.1.2 Model output

The output of the models is given in terms of mean opinion scores for colouration (MOS-C), discontinuity (MOS-D), noisiness (MOS-N) and sub-optimum loudness (MOS-L).

6.1.3 Scoring of background noise

Noise at the sending side or inserted in the transmission chain is a perceptual degradation in terms of noisiness and is thus considered in the predictions of the model. The prediction of perceptual quality dimensions as it can be perceived in a noisy listening environment and the related binaural effects lie outside the scope of this Recommendation.

There are test cases where the bandwidth of the background noise exceeds that of the speech signal significantly, and contributes a perceptible amount of energy outside the voice band (i.e., NB voice superimposed with super-wideband (SWB) noise signals).

NOTE – There is an example where in a teleconference audio-bridge a NB signal is transmitted from one far-end terminal to the WB bridge. At the same time, another subscriber is linked by a WB channel to the bridge and inserts a WB noise into the receiver. The observation point at the near end receives an NB signal superimposed on a WB noise.

Corresponding speech samples do not violate the application of this Recommendation, but have not yet been tested.

6.1.4 Scoring of temporal clipping

Depending on the set-up and instructions given in an absolute category rating (ACR) test with speech signals affected by temporal clipping, inconsistent scores may be obtained if speech is partially missing. The scores might be too optimistic in certain cases, e.g., if whole words, phrases or sentences are missed, or more pessimistic, if missing words are not perceivable by the listener. The model scores for discontinuity mostly reflect speech signals where missed voice is clearly perceptible in an ACR context, but still without having the reference signal for comparison (as it is an ACR context).

7 Comparison between objective and subjective scores

Clause 7 of [ITU-T P.863] describes how a comparison between scores that are objective (predicted) and subjective should be performed. However, for the judgement of overall quality and the resulting MOS, the same principles apply here, but for the prediction of perceptual dimensions scores related to colouration, discontinuity, noisiness and sub-optimum loudness, and their corresponding averages MOS-C, MOS-D, MOS-N and MOS-L.

8 Speech material

[ITU-T P.863] specifies the characteristics of speech samples to be structured and used for instrumental (objective) measures when measuring at an electrical interface. With modifications, this method can be also used for the acoustical path.

The temporal structure given in clause 7 of [ITU-T P.863.1] should also be applied to the models specified in this Recommendation. The models described in this Recommendation operate on speech material that follows the rules given in [ITU-T P.863.1] and should put no restrictions on the reference and degraded speech file other than those used in [ITU-T P.863.1].

8.1 Input or reference speech material

Reference speech or a reference signal is an original speech signal without any degradation. This should be recorded and stored in conformity with [ITU-T P.830]. In the case of an acoustical sending path, this signal is used to feed the artificial mouth. This speech signal is used by the model algorithms as a reference against which the effects of the system under test are revealed. The reference signal should fulfil the requirements as defined for FB signals in [ITU-T P.10].

In the case of an electrical insertion, this reference signal is also given to a terminal model, which may reflect NB, SWB or FB behaviour of the terminal to be modelled.

NOTE 1 – SWB signals limited at 14 kHz according to [ITU-T P.10] are considered as sufficient as reference signals, because spectral energies above 14 kHz for speech signals are seen as negligible for speech quality predictions.

For insertion into a proprietary interface, pre-filtering can be applied. It may follow the guidelines described in [ITU-T P.862.3] for those filters.

- The room used for recording reference material must have a reverberation time below 300 ms above 200 Hz (e.g., an anechoic chamber). Recordings must be made using omnidirectional microphones. The distance to the microphone must be approximately 10 cm. Background noise must be below 30 dB_{SPL}(A), where SPL indicates sound pressure level. Speech signals will be band pass filtered to 20 Hz...14 kHz. Directional microphones are allowed on the condition that the frequency response is the same that of the omnidirectional microphones previously mentioned.
- The reference speech signals are sampled at 48 kHz.
- In NB test scenarios, it is allowable to down-sample the reference signals to 8 kHz in a second step for insertion in the test channel (e.g., integrated services digital network/PSTN card). However, for the models described in this Recommendation, the reference has to be made available as a 48 kHz sampled reference as the test scenario is to be scored in an FB context.

The reference signal may be further processed before it is inserted in the transmission channel by either adding noises for testing noisy speech or by individual pre-filtering to achieve proper insertion characteristics for customized or proprietary insertion points such as headset connectors (see also clause 9.1 of [ITU-T P.862.3]). These processing steps are considered part of the system under test, but the reference signal for the models described in this Recommendation remains unprocessed as described in this clause.

NOTE 2 – The consequences of these definitions and proposals are that each captured sequence is measured against a flat and noise-free reference in the case of NB and SWB applications. If the actually used insertion path is not flat, this deviation is considered by the models.

Acoustical insertion applies to handset or hands-free devices. The test setup has to follow the ITU-T P.340 or ITU-T P.581 or other realistic use cases. Potential differences in the presentation level are allowed and are part of the test conditions.

NOTE 3 – Some artificial mouths are not specified above 10 kHz. However, since potential problems in the frequency response or other types of degradations are part of the auditory test and the objective evaluation, this limitation does not affect the evaluation of the models described in this Recommendation.

8.2 Degraded speech material

Degraded speech or a degraded signal is the reference speech that has passed through the system under test and was captured either at the electrical or acoustical interface.

Degraded signals have to be provided in mono and sampled at 48 kHz.

A diotic presentation (both ears receive the same mono signal) of the recordings in the auditory test is required.

All recordings to be used by the models described in this Recommendation must not apply a filter after recording.

It is recommended that speech signals presented in the auditory tests also be used as degraded speech for the objective evaluation. That means signals given to the models for evaluation are the same as those used in the subjective test. For example, if an NB condition is presented in the test, it should be upsampled to 48 kHz before listening or analysing it by the model.

Differences in the duration of active speech (e.g., long muted speech intervals) have to be handled by the model. Disadvantages as described in clause 8.1 of [ITU-T P.862.3] should be avoided by the model.

Weak noise insertions in the capturing path should not require especially pre-processed reference signals as described in clause 7.10 of [ITU-T P.862.3]. Such weak noise insertions also have to be handled properly in the case of noise-free reference signals by the model itself.

Level variations are included as test cases. Those level differences will be restricted to a range of +6 dB to –20 dB relative to a nominal level of the test application (in the case of electrical recording –26 dBov or equivalent). All recordings have to include a digital level as well as the sound level used in the auditory test for each database.

For all test signals delivered for SWB or FB, a digital level of –26 dBov (obtained with [ITU-T P.56]) corresponds to the nominal presentation level (e.g., 73 dB in case of diotic presentation). The actual presentation level can be directly derived from the ITU-T P.56 level of the degraded signal. (e.g., a level of –34 dBov corresponds to a presentation level 8 dB below the nominal level).

The adjustment of the listening devices has to be done accordingly. Here the adjustment can be derived by playing out a calibration signal at –26 dBov (recommended: speech-like babble noise, spectrally shaped according to [ITU-T P.50]) over the listening device and adjusting the sound level at an (A-weighted) ear reference point by means of an artificial ear to the nominal level.

8.3 Special requirements for acoustically captured speech material

Acoustical recordings have to be conducted using diffuse-field equalization for the artificial ear. In the case of recordings in hands-free scenarios using a loudspeaker or a speakerphone, a complete HATS has to be used. The playback device in the FB listening test has to be a diffuse-field equalized headphone in each case.

All acoustically recorded files have to be submitted in mono and sampled at 48 kHz. The signals at the acoustical interface have to be recorded by a diffuse-field equalized artificial ear. Only the signal recorded at one ear is required.

A diffuse-field equalized headphone also has to be used for presentation of such signals in the subjective test. The signal will be presented diotically; this means that each ear of the subject receives the same (mono) signal.

In the case of an acoustically captured signal, the user terminal is part of the recording and is scored as well. Obvious differences between different play-out levels (caused by the real receiving terminal) should be considered for the quality scoring in FB test scenarios. It is highly recommended that the acoustical recordings be presented in the subjective test with the actual sound levels used during recording.

The level differences in a test are restricted to a range of +6 dB to –20 dB relative to the nominal level of the test application. The nominal level is represented by an electrical level of –26 dBov. This nominal level corresponds to 73 dB_{SPL} at each ear.

The actual sound level presented in the listening test should correspond to the sound level during acoustical recording. It may be necessary to record the settings of the acoustical capturing equipment. In a post-processing step, the level of the captured signals can be adjusted so that an electrical signal of –26 dBov is presented at 73 dB_{SPL} at each ear.

The signals delivered into the database pool have to be levelled according to clause 8.2, where it is specified that a digital level of –26 dBov corresponds to the nominal level.

NOTE 1 – The artificial ear is not specified above 12 kHz. However, since potential problems in the frequency response or other types of degradations are part of the auditory test and the objective evaluation, this limitation does not affect the evaluation of the model.

NOTE 2 – Simulated acoustical recordings, where an electrically captured signal is convolved with an impulse response between the terminal (i.e., hands free reference point) and the artificial ear, are allowed. In the same way, simulations of the acoustical insertion can be used.

8.4 Acoustical insertion or capture for loudspeaker phones

Hands-free loudspeaker conditions may be recorded in different types of rooms or cars representative of the telephony situation. Rooms as described in [ITU-T P.340] have to be considered as one type for test conditions considering hands-free loudspeakers. Those recordings may reflect a normal office-type room.

- Room size: The room size should be in a range between $2.5 \times 3 \text{ m}^2$ and $3.5 \times 4 \text{ m}^2$. The room height should be between 2.20 m and 2.50 m.
- Treatment of the room: The playback room should be equipped with a carpet on the floor and some acoustical damping in the ceiling as typically found in office rooms. A curtain should cover one or two walls in order to avoid strong reflections by hard surfaces in the room. The reverberation time of the room should be less than 0.7 s but higher than 0.2 s between 100 Hz and 8 kHz.
- Noise floor: In order to reduce the influence of external noise, the noise floor measured in a room should be less than 30 dB_{SPL}(A).

9 Description of the model algorithms

9.1 Colouration model

The colouration prediction combines two indicators of [ITU-T P.863]: the frequency indicator *predictedMosPureFrq* and a second indicator *d5s0t2* that quantifies the overall disturbance in the time-frequency plane.

The model for estimating the colouration is a linear combination of the two indicators as:

$$\text{MOS-C} = 1.39 + 0.675 * \text{predictedMosPureFrq} - 0.0423 * d5s0t2$$

In the preceding formula, the values for both indicators *predictedMosPureFrq* and *d5s0t2* are limited to avoid extreme values. The maximum value for *d5s0t2* is set to 40, while the minimum and maximum values for *predictedMosPureFrq* are set to 2.0 and 4.5, respectively.

The minimum and maximum values for the colouration prediction are set to 1.0 and 4.75 MOS-C, respectively.

9.2 Discontinuity model

The discontinuity prediction combines three indicators of [ITU-T P.863]: two noisiness indicators (*n011* and *n000mosIntellCorrection*) and another indicator *a0s3t3* that quantifies overall added part of the disturbance in the time-frequency plane.

The model for estimating the discontinuity is a linear combination of the three indicators as:

$$\text{MOS-D} = 4.10 - 0.300 * n011 + 0.354 * n000mosIntellCorrection - 0.111 * a0s3t3$$

Upper limits have been used for two of the indicators used in the preceding formula. The maximum value for *n011* is set to 25, while the maximum value for *n000mosIntellCorrection* is set to 20.

The minimum and maximum values for the discontinuity prediction are set to 1.0 and 4.75 MOS-D, respectively.

9.3 Noisiness model

The noisiness prediction is based on two indicators that are calculated from the time-aligned spectrograms of [ITU-T P.863] and one indicator that models the impact of level deviations from the nominal level (−26 dBov) of active speech parts in the degraded signal. Principally, the first indicator *BGN* represents the background noise level with an ITU-T O.41 weighting. The *SED* represents the noisiness during speech, calculated as the distance between the spectral entropy of the reference and the degraded signal. The second indicator *SEDSTD* is the standard deviation of *SED* over time. The third indicator is the active speech level factor (*ASLF*) and it is derived from *logDISTAP*, which quantifies the power in active speech parts in the degraded signal.

The *BGN* and *SEDSTD* indicators are weighted by the *ASLF* and mapped to the MOS-N using a linear mapping as:

$$\text{MOS-N} = 4.48 - \text{ASLF} * (6.42 * \text{BGN} + 15.86 * \text{SEDSTD})$$

The minimum and maximum values for the noisiness prediction are set to 1.0 and 4.75 MOS-N, respectively.

Clauses 9.3.1 to 9.3.4 describe the steps required to derive the three indicators used in the noisiness prediction.

9.3.1 Active or inactive decision

To differentiate between background noise and noise on speech, the spectrogram frames are firstly divided into active and inactive types. To this end, the short-term signal power is calculated from the reference spectrograms as the sum across all frequency bins. The short-term power is then simply compared to a fixed power threshold to differentiate between active and inactive frames. If no inactive frames are found, the 10 frames with the least energy in the reference signal are marked as inactive (this may happen for non-conforming reference signals that do not contain a silent pause).

9.3.2 Background noise

To estimate background noise, the power spectral density (PSD) is calculated by averaging the degraded spectrograms over time. After that, the ITU-T O.41 weighting is applied to the common

logarithms of PSD values (log-PSDs). The sum over all frequency bins then gives the background noise of the degraded speech signal. The same calculations are then also performed for the reference signal. The difference between both background noise levels finally yields the indicator *BGN*. Additionally, the reference log-PSDs are weighted with the frame-based frequency response of the system, to exclude frequency components that are not included in the degraded signal.

9.3.3 Noise on speech

The noise on speech is estimated using the spectral entropy of the active frames of the reference and degraded spectrograms. The spectral entropy can be interpreted as a measure of disorder or noisiness of the frequency distribution of a PSD. It is based on the Shannon entropy, which indicates the amount of information contained in a stochastic source, based on its probability mass function. To calculate the spectral entropy, instead of a probability mass function, we apply the frequency distribution of a PSD. The frequency distribution $P(t, m)$ is calculated by normalizing the PSD of each frame as follows:

$$P(t, m) = \frac{S(t, m)}{\sum_f S(t, f)}$$

where

$S(t, f)$ is the spectrogram

t is the frame index

f is the frequency bin index of $S(t, f)$

m is the frequency bin index of $P(t, m)$.

Then the Shannon entropy is calculated to yield the per-frame spectral entropy $SE(t)$:

$$SE(t) = \frac{-\sum_{m=1}^N P(t, m) \ln P(t, m)}{\ln N},$$

where N denotes the number of frequency bins. The spectral entropy is normalized with the denominator $\ln N$, which represents the maximal spectral entropy of white noise.

To estimate the noise on speech, the reference and degraded spectrograms are first divided into five frequency bands, ranging in total from 300 to 3 400 Hz. Then, for each band, the spectral entropy is calculated. As a next step, the difference of the spectral entropy between the reference and the degraded signal is taken. Again, to avoid measuring frequency components that are not included in the degraded signal, the per-frame frequency response is applied as a weighting function. Additionally, a second weighting function is applied, which considers the perceptual importance of the individual frequency bands. The average of this weighted difference aggregated over frequency bins and time then yields the indicator *SED*; the standard deviation over time yields the third indicator *SEDSTD*.

9.3.4 Active speech level factor

The *ASLF* models the impact of deviations of the active speech parts in the degraded signal from the nominal level (−26 dBov).

The other two indicators used for the noisiness estimation (*BGN* and *SEDSTD*) quantify the amount of noise in the degraded signal. However, they do not consider the level of the active speech parts in the degraded file. Given a certain fixed amount of noise in the degraded signal, lower ASLs lead to a noisier perception of the degraded signal, because the signal-to-noise ratio (SNR) is lower. In the other direction, the degraded signal is perceived as less noisy when the speech signal level is higher, because the SNR is also higher. The indicator *ASLF* models that behaviour by comparing the ASL in the degraded signal to the nominal level (−26 dBov).

The *ASLF* is derived from the *aAvgActiveDistortedPower* indicator in [ITU-T P.863], which quantifies the power in active speech parts in the degraded signal.

First, the logarithm of the average power in active speech frames in the degraded signal *logDISTAP* is computed as:

$$\logDISTAP = 10 \log_{10}(10^{-8} + aAvgActiveDistortedPower/10^7)$$

Then, the *ASLF* is defined as:

$$ASLF = \frac{17 - \logDISTAP}{30} + 1$$

The indicator *logDISTAP* has a value around 17 in files with ASL close to the nominal level (−26 dBov), its value is > 17 in files where active speech parts are louder and < 17 in files where active speech parts are quieter.

The value of the *ASLF* is then:

- 1 in files with nominal level (−26 dBov);
- < 1 in files with the active speech above the nominal level (*ASLF* is 0.8 for 6 dB amplification of the speech for example);
- > 1 in files with the active speech below the nominal level (*ASLF* is 1.5 for 15 dB attenuation of speech, for example).

This way, a lower level in active speech parts in the degraded signal leads to a worse noisiness rating given the same amount of noise.

9.4 Sub-optimum loudness model

The sub-optimum loudness predictor is based mainly on an indicator of [ITU-T P.863] which quantifies the power in active speech parts in the degraded signal, that is the *aAvgActiveDistortedPower*. This information is combined with a second indicator *gainVarInd* that quantifies the presence of gain variations and models their negative impact on the perceived optimal loudness.

The model for estimating sub-optimum loudness is a linear combination of the *logDISTAP* and the *gainVarInd* indicators as:

$$MOS_L = 2.45 + 0.096 * \logDISTAP - 0.0295 * gainVarInd$$

The *logDISTAP* is derived from the *aAvgActiveDistortedPower* as explained in clause 9.3.4. Upper limits have been used for the two indicators in the preceding formula. The maximum value for *logDISTAP* is set to 20 (corresponding roughly to an ASL of −23 dBov) to avoid the loudness predictions to increase without upper limit for positive amplifications. The maximum value for *gainVarInd* is set to 35.

The minimum and maximum values for the sub-optimum loudness prediction are set to 1.0 and 4.75 MOS-L, respectively.

9.4.1 Gain variation indicator

A new indicator has been developed to model the negative impact of strong gain variations on the perceived sub-optimum loudness. This gain variation indicator *gainVarInd* quantifies the amount of gain variation in the degraded signal with respect to the original reference signal, by integrating the loudness deviations from a central fixed value.

In a first step, the loudness of the time-aligned versions of the reference and degraded signals (*loudRef* and *loudDeg*) is computed for each frame in a frequency band ranging between 250 Hz and 3.5 kHz approximately.

Next, the loudness deviation $loudDev$ between the aligned reference and degraded is computed for each active speech frame with index k as:

$$loudDev(k) = loudDeg(k) - loudRef(k)$$

Then, the fixed loudness deviation $fixedLoudDeviation$, which can be interpreted as the fixed component of the loudness deviation resulting from a gain variation (i.e., resulting from a fixed gain difference between reference and degraded signals), is computed as the median value of the loudness deviation across all active speech frames:

$$fixedLoudDev = median(loudDev(k))$$

Indeed, if there is only a fixed gain difference between reference and degraded signals, the loudness deviation vector $loudDev$ will be more or less constant over the active speech frames, with all values close to $fixedLoudDev$. The more important are the gain variations in the degraded signal compared to the reference signal, the more important will be the deviations with respect to this central value $fixedLoudDev$.

Finally, the gain variation indicator $gainVarInd$ integrates the absolute loudness deviations with respect to this fixed central value as:

$$gainVarInd = \frac{1}{K} \sum_{k=1}^K |loudDevWin(k)|$$

where the loudness deviation from the fixed central value in a window around frame k $loudDevWin(k)$ is defined as:

$$loudDevWin(k) = \frac{1}{W} \sum_{w=0}^{W-1} (loudDev(k+w) - fixedLoudDev)$$

A small smoothing window of $W=10$ frames is considered in the calculation of the loudness deviation, to avoid high values for the gain variation indicator $gainVarInd$ in the case of very fast loudness variations, which are typical of noisy degraded signals.

In addition, to reduce the impact of extreme values on the gain variation indicator, which can be caused for example by frame losses, interruptions or noises, the maximum deviation per active speech frame ($loudDev(k) - fixedLoudDev$) is limited to the ± 10 dB range.

Annex A

Subjective test method for obtaining perceptual dimension scores

(This annex forms an integral part of this Recommendation.)

The perceptual dimension scores predicted by the models described in this Recommendation should reflect as closely as possible the judgements of humans with respect to the individual dimensions given in clause 6. These judgements should be collected in a test carried out in the way described in this annex.

The test procedure described here is not identical to that described in [ITU-T P.806]. In fact, the procedure described in [ITU-T P.806] defines seven rating scales in addition to an overall quality scale. Two of those scales relate to slowly and fast-varying degradations of the speech signal, whereas in the procedure described here there is only one dimension related to discontinuity. Further, two scales described in [ITU-T P.806] describe degradations of low- and high-frequency colouration in the speech signal, whereas the method described here targets only one colouration dimension. In addition, two scales described in [ITU-T P.806] describe degradations due to the level and the variability of background noise, whereas the method described here targets only one noisiness dimension. Finally, the scale defined for loudness in [ITU-T P.806] ranges from much louder than preferred to much quieter than preferred, whereas the scale used here ranges from an optimum loudness level to a non-optimum loudness level. During the development of this Recommendation, both sets of dimensions (the four dimensions described here, as well as the seven dimensions targeted in [ITU-T P.806]) were originally considered. Finally, work on models for estimating the seven dimensions of [ITU-T P.806] was discontinued, and only the four dimensions used here were retained.

The test procedure is similar to a standard ACR overall quality test as it is described in [ITU-T P.800] and [ITU-T P.830]. All deviations from the procedure described in these two Recommendations are given hereafter.

Four descriptive scales are used for measuring the four dimensions colouration, discontinuity, noisiness and sub-optimum loudness. That way, separate scores for the perceptual dimensions present in test conditions containing multi-dimensional degradations can be obtained. The graphical layout of the colouration, discontinuity, noisiness and sub-optimum loudness scales is similar to that of the scales recommended in [ITU-T P.851]. The poles of the scales are labelled with the antonym attributes: continuous – discontinuous (discontinuity dimension); not noisy – noisy (noisiness dimension); uncoloured – coloured (colouration dimension); and optimum level – non-optimum level (loudness), see Figures A.1 to A.4.

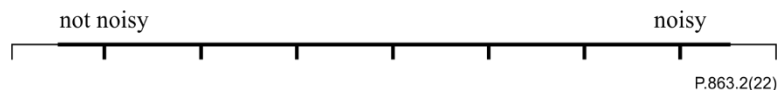


Figure A.1 – Noisiness scale

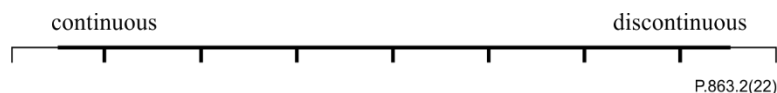


Figure A.2 – Discontinuity scale

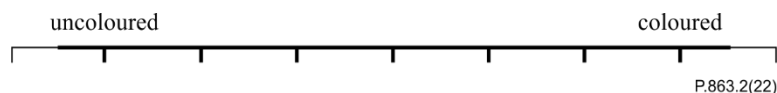


Figure A.3 – Colouration scale

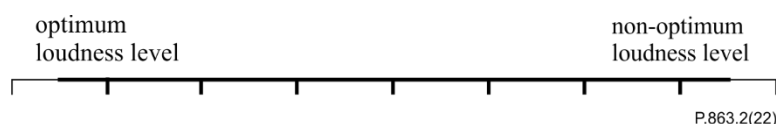


Figure A.4 – Sub-optimum loudness scale

In the dimension assessment experiment, the scales are presented separately, i.e., consecutively for each stimulus. Prior to rating registration, listeners are asked to listen to the entire speech sample. During one trial, they can optionally repeat the playback. The rating scheme for one sample is depicted in Figure A.5 (for three scales).

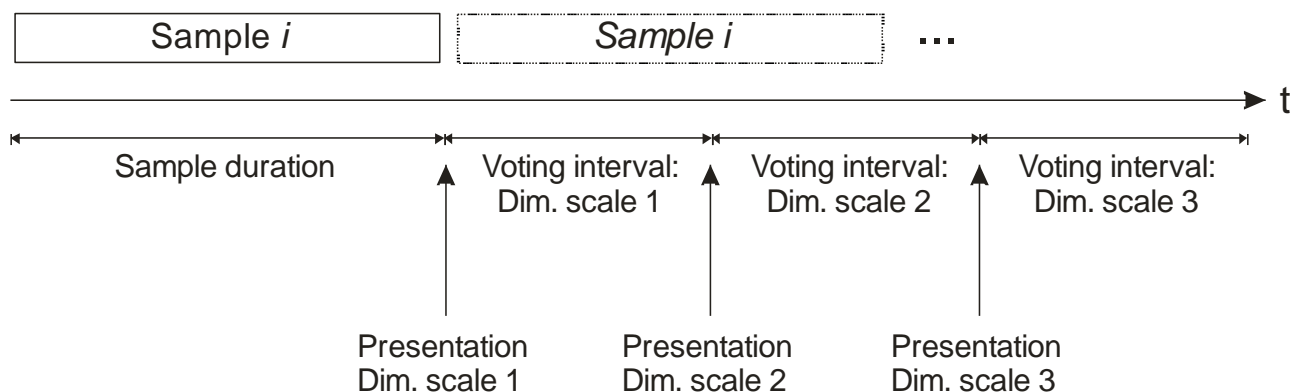


Figure A.5 – Sample presentation and rating (dimension assessment)

The samples are presented in randomized order. For each participant, the order of the scales is permuted, following the scheme tabulated in Table A.1.

Table A.1 – Presentation order of the *discontinuity*, *noisiness*, *colouration* and *loudness* scale

Participant	Dim scale 1	Dim. scale 2	Dim. scale 3	Dim. scale 4
1	dis	noi	col	lou
2	noi	col	lou	dis
3	col	lou	dis	noi
4	lou	dis	noi	col
...

The order is held constant for an individual participant in order to avoid confusion of the scales.

A detailed description of the four-dimension scales is given to the subjects. The instructions start off explaining that in this part of the experiment, the features or characteristics of speech samples are supposed to be judged (i.e., not the quality), and that this evaluation is done by means of four scales. Each scale is labelled with an attribute at each end that describes the characteristic to be judged. Each scale and its usage are separately described in detail, using synonyms to the scale attributes as an aid. In detail, participants are instructed that:

- with the scale in Figure A.1, the noisiness of the sample is supposed to be judged; the labels "not noisy" and "noisy" can be paraphrased with the terms "not hissing" and "hissing", respectively;
- with the scale in Figure A.2, the discontinuity of the sample is supposed to be judged; the labels "continuous" and "discontinuous" can be paraphrased with the terms "regular", "steady", "not chopped", "not bubbling" or "not ragged" and "irregular", "shaky", "chopped", "bubbling" or "ragged", respectively;

- with the scale in Figure A.3, the colouration of the sample is supposed to be judged; the label "uncolored" and "colored" can be paraphrased with the terms "direct", "close", "thick" or "not nasal" and "indirect", "distant", "thin" or "nasal", respectively;
- with the scale in Figure A.4, the loudness level of the sample is supposed to be judged; the label "optimum loudness level" means that the loudness level is neither too high nor too low.

The complete instructions can be found in Appendix II, extended by the inclusion of the loudness scale, see the last list entry in the previous paragraph.

The dimension assessment is divided into training phases where listeners can:

- internalize the meaning of the scales by acoustic examples, and
- familiarize themselves with the usage of the scales.

For the training of the meaning of the scales, exemplary samples for each scale are presented that are distorted in (mainly) one dimension. Therefore, unidimensional anchor conditions specified in Table 4 of [ITU-T P.806] corresponding to the four dimensions can serve as training samples.

The acoustic presentation is done together with the descriptive synonyms by means of a computer screen where the participants can listen to the samples until they confirm that they have understood the meaning of the scales. The understanding is supported by presenting an undistorted sample (direct SWB), stating that this particular sample is completely "not noisy", "continuous", "uncoloured" and of "optimal loudness". A screenshot of the graphical training interface is included in the instructions in Appendix II.

The preceding trials help listeners to familiarize themselves with the practical usage of the scales and the range of degradations to be expected. Therefore, several samples differing in quality and character of the degradation are rated in a brief dedicated training session.

A common transformation rule of the raw scores to dimensional MOS (MOS-C, MOS-D, MOS-L and MOS-N) values has to be agreed. The raw ratings of the discontinuity, noisiness, colouration and loudness scores (ranging from 0 to 6 according to the scale design) can either be linearly transformed to the MOS-range [1;5] or the transformation equation from the extended and continuous scale into the MOS-range [1;5] using absolute categories can be applied:

$$\text{MOS}[1;5] = -0.026 \, 2\text{MOS}_{\text{EC}}^3 + 0.236 \, 8\text{MOS}_{\text{EC}}^2 + 0.190 \, 7\text{MOS}_{\text{EC}} + 1$$

where MOS_{EC} is the score obtained on the extended [1;6] scale.

The resulting values are denoted as MOS-C, MOS-D, MOS-L and MOS-N.

Annex B

Conformity data and tests

(This annex forms an integral part of this Recommendation.)

B.1 List of files provided for conformity validation

The conformity validation process described in this annex relates to the following files, which are provided in the "_Results_PAMD" subdirectory of the electronic attachment:

NOTE – The electronic attachment can be downloaded from <https://www.itu.int/net/itu-t/sigdb/genaudio/AudioForm-g.aspx?val=100015010>

- Test_1_results_ref.txt *file pairs and ITU-T PAMD scores for test 1;*
- Test_2_results_ref.txt *file pairs and ITU-T PAMD scores for test 2;*
- Test_3_results_ref.txt *file pairs and ITU-T PAMD scores for test 3.*

The PAMD_TUB_P501 speech files are in wireless access in vehicular environments (WAVE) format (16-bit linear PCM with WAVE header, little-endian byte ordering, at 48 kHz sampling rate). The SWB_TNO_601_48k and SWB_SQ_48k are in RAW format (16-bit linear PCM, little-endian byte ordering, at 48 kHz sampling rate). These files form an integral part of this annex.

For all conformity tests 1 to 3, there are BAT files prepared. The BAT files assume that the reference executable is called *PAMDmodel.exe* and is called as follows:

PAMDmodel <reference.raw> <degraded.raw> <Sampling Frequency / Hz> <Mode: FB>

B.2 Conformity tests

B.2.1 Conformity data sets

The data sets for the conformity tests are as given in Table B.1.

Table B.1 – Data sets for the conformity tests

Test	Number of file pairs	Data set
1	240	Clause B.1 'PAMD_TUB_P501' as attached (Test 1). – <i>Mandatory</i> –
2	200	Clause B.1 'SWB_TNO_601_48k' as attached (Test 2). – <i>Mandatory</i> –
3	50	Clause B.1 'SWB_SQ_48k' as attached (Test 3). – <i>Mandatory</i> –

B.2.2 Conformity requirements

The test requirements are the confirmation of very narrow distribution of differences to the reference values provided in clause B.1. The requirements refer to all MOS-C, MOS-D, MOS-L and MOS-N scores.

The allowed distribution of differences across all mandatory tests 1, 2 and 3 are summarized in Table B.2. The requirements are based on the absolute difference in the model score between the implementation under test and the reference values given in clause B.1.

Table B.2 – Allowed distribution of differences across all mandatory tests

Absolute difference	Allowed occurrence (%)
> 0.0001	5.00
> 0.001	1.00
> 0.01	0.50
> 0.1	0.05
> 0.3	0.00

For databases other than those specified in this annex, the same error distribution must not be exceeded. For unknown data, a test set of at least 2 000 file pairs – preferably from complete subjective experiments – has to be taken for those statistics.

B.3 Digital attachments

Processing scripts (.bat files) are contained in the electronic attachment to this Recommendation as:

- "Test_1.bat";
- "Test_2.bat";
- "Test_3.bat".

Enclosed speech material is contained in the electronic attachment to this Recommendation in the following folders:

- "PAMD_TUB_P501";
- "SWB_SQ_48k";
- "SWB_TNO_601".

Appendix I

Reporting of the performance results for the model algorithms based on the correlation, RMSE and RMSE* metrics

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

Table I.1 shows the performance numbers of the models described in this Recommendation for each of the four dimensions in the eight databases that are currently available: five were used in the training and three in the validation phase.

Table I.1 – Results of the models described in this Recommendation on available databases

Database <i>Validation</i>	MOS-N			MOS-D			MOS-C			MOS-L		
	r	RMSE	RMSE*	r	RMSE	RMSE*	r	RMSE	RMSE*	r	RMSE	RMSE*
PAMD_SwissQual2	0.910	0.504	0.364	0.669	0.731	0.579	0.858	0.449	0.300	0.483	0.307	0.167
PAMD_TUB1val	0.905	0.396	0.275	0.792	0.539	0.425	0.787	0.464	0.340	0.904	0.365	0.268
PAMD_Orange2	0.892	0.486	0.319	0.853	0.371	0.230	0.683	0.425	0.233	0.601	0.449	0.299

Database <i>Training</i>	MOS-N			MOS-D			MOS-C			MOS-L		
	r	RMSE	RMSE*	r	RMSE	RMSE*	r	RMSE	RMSE*	r	RMSE	RMSE*
PAMD_SwissQual1	0.956	0.300	0.174	0.870	0.315	0.147	0.864	0.390	0.241	0.863	0.337	0.134
PAMD_Orange1	0.928	0.387	0.189	0.802	0.467	0.288	0.841	0.396	0.205	0.714	0.366	0.219
PAMD_DTAG3	0.892	0.328	0.217	0.828	0.341	0.226	0.879	0.336	0.204	0.899	0.306	0.181
PAMD_DTAG2	0.841	0.458	0.230	0.855	0.490	0.342	0.965	0.307	0.183			
PAMD_DTAG1	0.939	0.299	0.158	0.830	0.460	0.256	0.934	0.305	0.130			

NOTE – Root mean square error (RMSE) values are computed after a first order polynomial mapping, while RMSE* values are computed after a third order polynomial mapping.

Appendix II

Test instructions

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

Test instructions read as follows.

"Thank you for attending this experiment. Please take your time to read these instructions. If you have any questions, please address them to the experimenter.

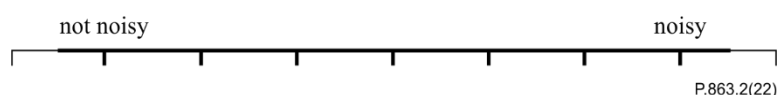
You are taking part in a listening test today where you are going to judge features of speech samples. Each sample was spoken by the same speaker but differs in its characteristic.

The characteristic of each sample should be judged by means of three descriptive scales. As a first step, you are going to get acquainted with them. In the following, the scales and their usage will be described.

Each scale is labelled with attributes at both ends. You are going to give a judgement about how far the characteristic of a speech sample can be described by the attributes.

1) Noisiness

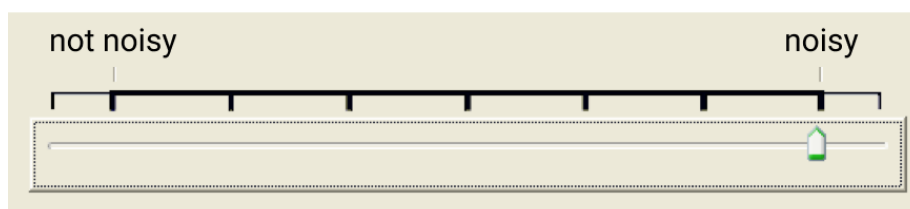
One of the scales is labelled with "noisy" and "not noisy". It is depicted in the following:



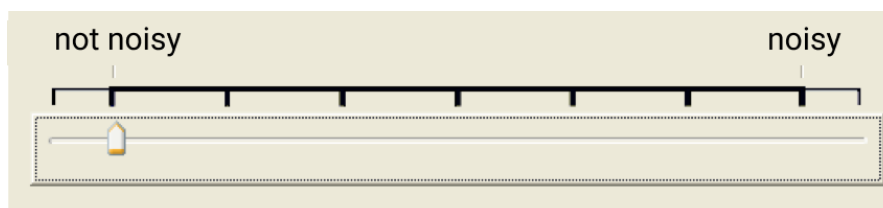
Noisiness scale

With this scale, the noisiness of the sample should be judged. The labels "not noisy" and "noisy" can be paraphrased with the terms "not hissing" and "hissing", respectively.

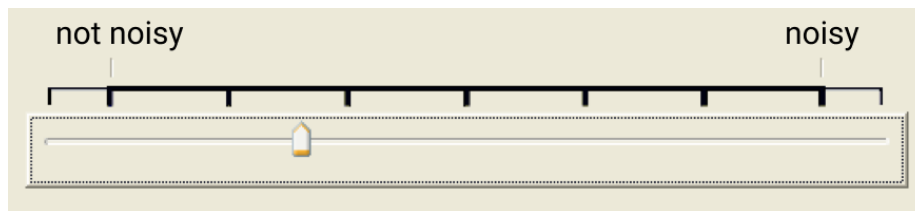
If you think that the speech sample is very noisy, move the software slider to the following position:



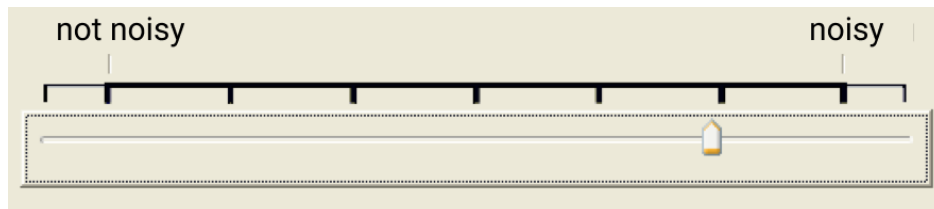
If you do not perceive any noisiness, move the slider to the position "not noisy":



You can make use of the whole range of the scale in order to describe the degree of noisiness. For instance, if you think the degree of noisiness is just moderate, you could move the slider to this position:

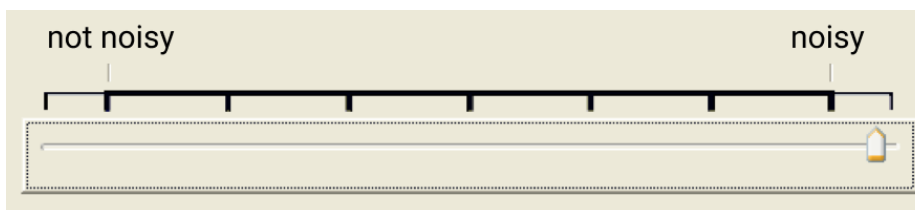


Perhaps, the sample is distinctly noisy, but not extremely noisy, your judgement could look as follows:



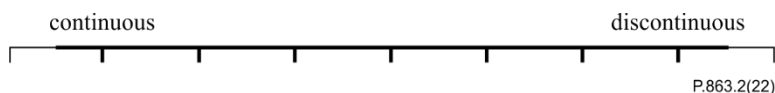
If needed, you can make use of the spaces if you do not want to decide between two tick marks.

In particular, you can use the "overflow areas" beyond the labels if these are insufficient for your judgement, e.g.:



2) Discontinuity

By means of a second scale you are going to judge the discontinuity of a sample:



P.863.2(22)

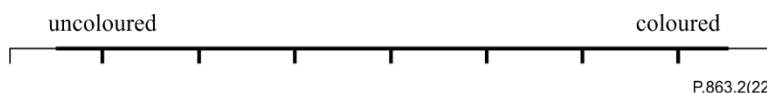
Discontinuity scale

The term "continuous" means that the speech sample is completely regular, steady, not chopped, not bubbling, and not ragged. "Discontinuous" can be paraphrased by the terms "irregular", "shaky", "chopped", "bubbling" and ragged.

Use the scale in the same way as described for the noisiness scale (see previous).

3) Colouration

The third scale serves for describing the colouration of a sample



P.863.2(22)

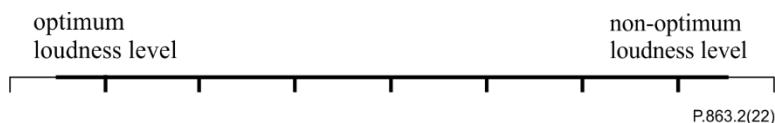
Colouration scale

The term coloured means the sample sounds direct, close, thick and not nasal. In contrast, the term uncoloured means the sample sounds indirect, distant, thin, and nasal.

Again, use the scale in the same way as described for the noisiness scale (see previous).

4) Sub-optimum loudness

The fourth scale serves for describing the loudness level of a sample






The term "optimum loudness level" means that the loudness level is neither too high nor too low. In contrast, the term "non-optimum loudness level" means that the loudness level is either too high or too low.


Again, use of the scale in the same way as described for the noisiness scale (see previous).

Hopefully, the four scales are conceptually already clear. In order to know how the features to be scaled sound like, listen to some typical speech samples that can be associated with a single scale each. Make use of this screen:

Features of speech samples

By means of three scales the features of speech samples should be characterized. The labels of these scales are listed below, together with descriptions and exemplary samples. Please listen to the samples and make yourself clear about the conceptions

- "not noisy" vs. "noisy"
(not hissing vs. hissing); 
- "continuous" vs. "discontinuous"
(regular, steady, not chopped, not bubbling, not ragged vs. irregular, shaky, chopped, bubbling, ragged); 
- "uncoloured" vs. "coloured"
(direct, close, thick, not nasal vs. indirect, distant, thin, nasal); 

Listen to the following sample for conception.  It is completely "noisy", "continuous", and "uncoloured".

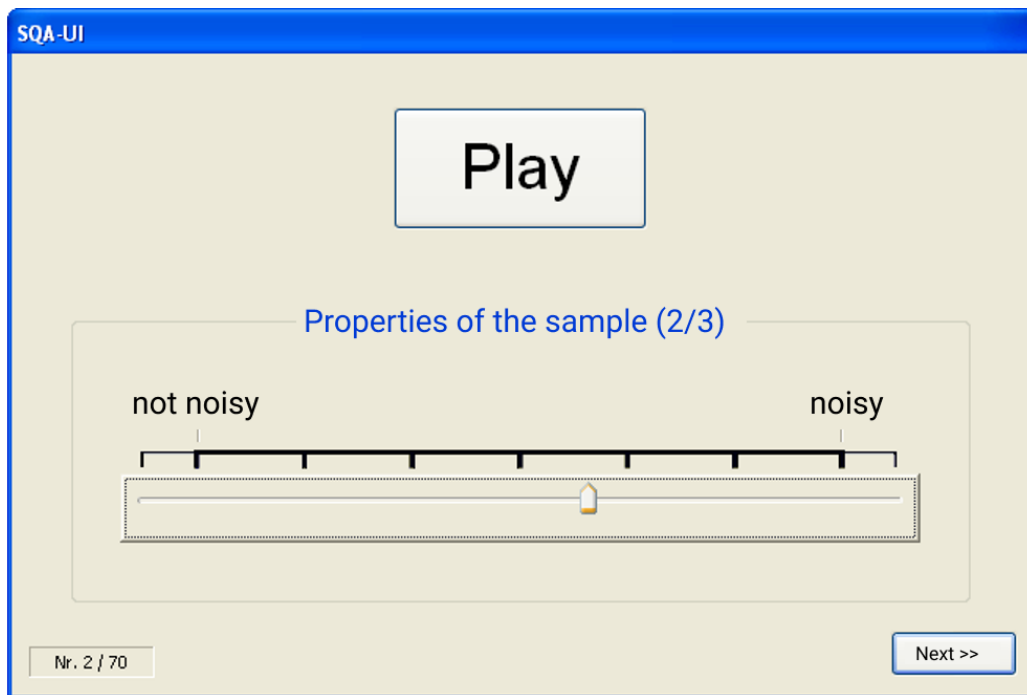
If the meaning of the labels is clear, please proceed reading the instructions.

P.863.2(22)

Figure II.1 – Features of speech samples

Please make yourself familiar again with the scales, the synonyms, and the corresponding acoustic examples. Only if you are aware which feature of a sample is meant by the respective scale, proceed reading. If you have any questions, please address them to the experimenter at any time.

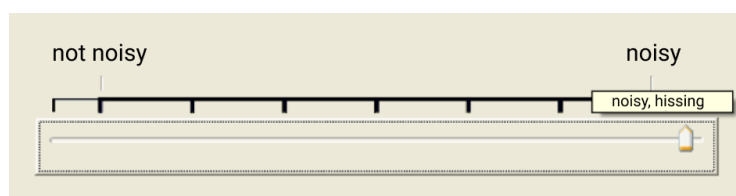
Now the experiment can finally start! The following figure shows a screenshot of the experimental software:



You have already become acquainted with the usage of the scales.

The three scales will be presented subsequently per speech sample. The sample will be played back automatically for the first of the three scales. After registering your vote, please click on "continue". Now, you are going to assess the same sample on the second scale. If desired, you can listen to the sample once again (click "Play"). Press "continue" again after registering your vote and proceed accordingly for the third scale. Once you complete your assessment here, the procedure is repeated for the next speech sample. A training phase takes place prior to the actual experiment in order to provide an overview and for practice in giving your judgements. Thus, you can familiarize yourself with the task first. Halfway through the experiment, the test is interrupted by a short pause.

In order to remember what is meant exactly by the scale labels, place the cursor directly on to the label in order to display its synonyms:



For some samples, you might have the impression that you have already judged them. This, however, is not the case. Thus, assess every sample independently of the others. Do not try to remember how you judged "similar" past samples, but give individual judgements for every scale and every sample.

Please register your vote intuitively and quickly. There are neither right nor wrong answers in this subjective investigation. Your personal impression is important for this investigation exclusively.

Please remember: The test always involves the same speaker. If you are unsure, press "Play" repeatedly and have a look at the descriptions of the scale labels.

If you have any questions, please address them to the experimenter."

Bibliography

- [b-ITU-T P.862] Recommendation ITU-T P.862 (2001), *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*.
- [b-Wältermann] Wältermann, M. (2013), *Dimension-based quality modeling of transmitted speech*. Heidelberg: Springer. 203 pp.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
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	Environment and ICTs, climate change, e-waste, energy efficiency; 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
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
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, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems