International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.863

(09/2014)

SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Methods for objective and subjective assessment of speech quality

Perceptual objective listening quality assessment

Recommendation ITU-T P.863

International Telecommunication Union

ITU-T P-SERIES RECOMMENDATIONS

TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

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

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

Recommendation ITU-T P.863

Perceptual objective listening quality assessment

Summary

Recommendation ITU-T P.863 describes an objective method for predicting overall listening speech quality from narrowband (NB) (300 to 3 400 Hz) to super-wideband (SWB) (50 to 14 000 Hz) telecommunication scenarios as perceived by the user in an ITU-T P.800 or ITU-T P.830 absolute category rating (ACR) listening-only test. Recommendation ITU-T P.863 supports two operational modes, one for narrowband and one for super-wideband. This Recommendation presents a high-level description of the method, advice on how to use it, and some results from a benchmark carried out in the period 2006-2010. All essential parts of the model are described in detail, and are provided in separate pdf-files (see Annex B). These files form an integral part of this Recommendation and shall take precedence in case of conflicts between the high-level descriptions included in the main body of this Recommendation and the corresponding detailed description parts. A conformance testing procedure is also specified in Annex A to allow a user to validate that an alternative implementation of the model is correct.

This Recommendation includes an electronic attachment containing detailed descriptions in pdf format (see Annex B) and conformance testing data (see Annex A).

The 2014 revision of ITU-T P.863 introduces bug fixes and resolves reported issues from ITU-T P.863 field deployments. On average the scores produced by this revised version of P.863 are very close to the values obtained from the previous version (V1.1). Due to the improvements and bug fixes in the revised version, there may however be significant differences for some individual measurements, especially for cases where the previous version failed. It should be observed that there is also a 2014 revision of the companion Recommendation P.863.1.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.863	2011-01-13	12	11.1002/1000/11009
1.1	ITU-T P.863 (2011) Amd. 1	2011-11-09	12	11.1002/1000/11463
2.0	ITU-T P.863	2014-09-11	12	11.1002/1000/12174

Keywords

Listening quality, objective quality, perceptual model, voice quality.

---

* 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 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/.

Table of Contents

Electronic attachment: Detailed descriptions in pdf-format and conformance testing data.

Introduction

Recommendation ITU-T P.863 defines a single algorithm for assessing the speech quality of current and near future telephony systems that utilize a broad variety of coding, transport and enhancement technologies.

The measurement algorithm is a full reference model which operates by performing a comparison between a known reference signal and a captured degraded signal. This is consistent with the algorithms described in Recommendations ITU-T P.861 and ITU-T P.862.

Recommendation ITU-T P.861, published in 1996, was primarily focused on identifying the quality impact of codecs. Subsequent to its release, work on a successor was started to create an algorithm suitable for assessing the additional impact of network impairments. The work resulted in the publishing of Recommendation ITU-T P.862 in 2001. Recommendation ITU-T P.863 (which during its development was known as P.OLQA) incorporates current industry requirements and in particular allows the assessment of super-wideband speech as well as networks and codecs that introduce time warping.

Recommendation ITU-T P.863

Perceptual objective listening quality assessment

1 Scope

This Recommendation[1] defines a single algorithm for assessing the speech quality of current and near future telephony systems that utilize a broad variety of coding, transport and speech enhancement technologies.

Based on the benchmark results presented within the studies of ITU-T, an overview of the test factors, coding technologies and applications to which this Recommendation applies is given in Tables 1 to 4. Table 1 presents factors and applications included in the requirement specification and which were used in the selection phase of the ITU-T P.863 algorithm. It should be noted that the performance of the ITU-T P.863 algorithm under each individual condition in Table 1 is not reflected in this table. Additional and detailed analysis will be undertaken in the characterization phase of the ITU-T P.863 algorithm. Table 2 presents a list of conditions for which this Recommendation is not intended to be used. Table 3 presents test variables for which further investigation is needed, or for which ITU-T P.863 is subject to claims of providing inaccurate predictions when used in conjunction with these. Finally, Table 4 lists factors, technologies and applications for which the ITU-T P.863 algorithm has not currently been validated. Note that the ITU-T P.863 algorithm cannot be used to replace subjective testing.

It should also be noted that the ITU-T P.863 algorithm does not provide a comprehensive evaluation of transmission quality. It only measures the effects of one-way speech distortion and noise on speech quality. The effects of delay, sidetone, echo, and other impairments related to two-way interaction (e.g., centre clipper) are not reflected in the ITU-T P.863 scores. Therefore, it is possible to have high ITU-T P.863 scores, yet poor overall conversational quality.

A characterization phase will follow the approval of this Recommendation. The purpose of the characterization phase is to prove the applicability of the ITU-T P.863 algorithm in real applications and may include new test conditions, new test scenarios and alternate test methodologies.

Table 1 – Factors and applications included in the requirement specification and used in the selection phase of the ITU-T P.863 algorithm

Test factors
Speech input levels to a codec
Transmission channel errors
Packet loss and packet loss concealment
Bit rates if a codec has more than one bit-rate mode
Transcodings
Acoustic noise in sending environment
Effect of varying delay in listening-only tests
Short-term time warping of audio signal
Long-term time warping of audio signal
Listening levels between 53 and 78 dB(A) SPL in super-wideband mode

[1] This Recommendation includes an electronic attachment containing detailed descriptions in pdf format (see Annex B) and conformance testing data (see Annex A).

Table 1 – Factors and applications included in the requirement specification and used in the selection phase of the ITU-T P.863 algorithm

Test factors
Packet loss and packet loss concealment with PCM type codecs
Temporal and amplitude clipping of speech
Linear distortions, including bandwidth limitations and spectral shaping ('non-flat frequency responses')
Frequency response
Coding technologies
ITU-T G.711, ITU-T G.711 PLC, ITU-T G.711.1
ITU-T G.718, ITU-T G.719, ITU-T G.722, ITU-T G.722.1, ITU-T G.723.1, ITU-T G.726, ITU-T G.728, ITU-T G.729
GSM-FR, GSM-HR, GSM EFR
AMR-NB, AMR-WB (ITU-T G.722.2), AMR-WB+
PDC-FR, PDC-HR
EVRC (ANSI/TIA-127-A), EVRC-B (TIA-718-B)
Skype (SILK V3, iLBC, iSAC and ITU-T G.729)
Speex, QCELP (TIA-EIA-IS-733), iLBC, CVSD (64 kbit/s, "Bluetooth")
MP3, AAC, AAC-LD
Applications
Codec evaluation
Terminal testing, influence of the acoustical path and the transducer in sending and receiving direction. (NOTE – Acoustical path in receiving direction only for super-wideband mode.)
Bandwidth extensions
Live network testing using digital or analogue connection to the network
Testing of emulated and prototype networks
UMTS, CDMA, GSM, TETRA, WB-DECT ,VoIP, POTS, PSTN, Video Telephony, Bluetooth
Voice Activity Detection (VAD), Automatic Gain Control (AGC)
Voice Enhancement Devices (VED), Noise Reduction (NR)
Discontinuous Transmission (DTX), Comfort Noise Insertion
NOTE – Individual conditions will be analysed during the characterization phase of ITU-T P.863 and details will be made available.

Table 2 – ITU-T P.863 is not intended to be used with these variables

Test factors
Effect of delay in conversational tests
Talker echo
Sidetone
Acoustic noise in receiving environment
Coding technologies
Applications
Non-intrusive measurements
Two-way communications performance

Table 3 – Test variables for which further investigation is needed

Test factors
Acoustical recordings using free-field microphones without HATS or ear-canal simulators
Coding technologies
Applications
NOTE – The entries to this table may be revised in future version of this standard after contributions investigating these factors

Table 4 – Factors, technologies and applications for which ITU-T P.863 has currently not been validated (For further study)

Test factors
Talker dependencies
Multiple simultaneous talkers
Bit-rate mismatching between an encoder and a decoder if a codec has more than one bit-rate mode
Network information signals as input to a codec
Artificial speech signals as input to a codec
Music as input to a codec
Listener echo
Coding technologies
Coding technologies operating below 4 kbit/s, EVS
Applications
Live VoLTE networks

It has to be noted that relatively to the previous edition of this Recommendation, plain A-Law coding is scored slightly lower (by about 0.1 MOS-LQO) than µ-Law.

It has to be further noted that this Recommendation is more insensitive to very low noise floors in super-wideband mode than in narrowband mode.

The test set of ITU-T P.863 covers the following languages: American English, British English, Chinese (Mandarin), Czech, Dutch, French, German, Italian, Japanese, Swedish, Swiss German. The subjective experiments were conducted in subjective test laboratories located in these countries.

ITU-T P.863 is the next-generation voice quality testing technology for fixed, mobile and IP-based networks. ITU-T P.863 has been selected to form the new ITU-T voice quality-testing standard. This Recommendation was developed between 2006 and 2010 in a competition carried out by ITU-T, in order to define a technology update for [b-ITU-T P.862].

The purpose of the objective ITU-T P.863 model is to predict overall listening speech quality from narrowband (300 to 3 400 Hz) to super-wideband (50 to 14 000 Hz) telecommunication scenarios as perceived by the user. This includes all speech-processing components usually considered for telecommunications in clean and noisy conditions. The term *'listening speech quality'* means the overall speech quality as perceived and scored by human subjects in an absolute category rating experiment according to [ITU-T P.800] or [ITU-T P.830]. In super-wideband mode, ITU-T P.863 scores are predicted on a MOS ACR super-wideband scale; details on the super-wideband experiment design are provided in Appendix II. In narrowband mode, ITU-T P.863 scores are predicted on a MOS ACR narrowband scale. The model outputs in the two modes are referred to as MOS-LQOn and MOS-LQOsw.

As is the case for [b-ITU-T P.861] and [b-ITU-T P.862], the approach of ITU-T P.863 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.

ITU-T P.863 can be applied to signals recorded at an electrical interface (as was the case for [b-ITU-T P.862]) but also to – in case of super-wideband operation mode – signals recorded using an artificial ear simulator. Other technologies or components, such as speech storage formats, or non-telephony applications, such as public safety networks or professional mobile radio connections, were not part of the competition and the selection criteria.

ITU-T P.863 operational modes

It is important to understand and consider the two different operational modes supported by ITU-T P.863:

- super-wideband, and
- narrowband.

Table 5 summarizes the applicability of ITU-T P.863 operational modes to different telecommunication scenarios.

Table 5 – Applicability of ITU-T P.863 operational modes to different signal bandwidths and listening situations

ITU-T P.863 mode	Signal bandwidth	Listening situation	Scale	Output
Super-wideband	Super-wideband wideband narrowband	Flat headphone diffuse field equalized, diotic presentation	MOS ACR super-wideband scale	MOS-LQOsw
Narrowband	Narrowband	IRS receive handset, monotic presentation	Narrowband scale comparable to [b-ITU-T P.862] in conjunction with [b-ITU-T P.862.1]	MOS-LQOn

The main difference between both modes is the bandwidth of the reference speech signal used by the model.

In super-wideband mode, the received (and potentially degraded) speech signal is being compared to a super-wideband reference. Consequently, band-limitations are considered as degradations and are scored accordingly. The listening quality is modelled as perceived by a human listener using a diffuse-field equalized headphone with diotic presentation (same signal at both ear-caps). The prediction uses a super-wideband listening quality scale where the ITU-T P.863 algorithm saturates at MOS-LQOsw = 4.75 for a transparent super-wideband signal. The super-wideband signals were assessed in an ITU-T P.800 ACR listening-only test in the ITU-T P.863 evaluation phase.

In contrast, in narrowband mode the received (and potentially degraded) speech signal is being compared to a narrowband (300 to 3 400 Hz) reference. Consequently, normal telephone band-limitations are not considered as severe degradations and are scored less. This narrowband mode maintains the compatibility to previously developed models such as [b-ITU-T P.862] in conjunction with [b-ITU-T P.862.1]. The listening quality is modelled as perceived by a human listener using a loosely coupled IRS type handset at one ear (monotic presentation). The prediction uses the common narrow-band listening quality scale where the ITU-T P.863 algorithm saturates at MOS-LQOn = 4.5 for a transparent narrowband signal.

NOTE 1 – For the two operational modes, the quality ratings are obtained on two different scales, namely the traditional scale for the narrowband mode and the future oriented scale for the super-wideband mode.

NOTE 2 – Acoustical recordings, as well as the influence of the presentation level, can only be predicted in super-wideband operational mode. The narrowband operational mode is restricted to electrical recordings and a nominal presentation for compatibility with [b-ITU-T P.862] in conjunction with [b-ITU-T P.862.1] application areas.

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 G.191] Recommendation ITU-T G.191 (2010), *Software tools for speech and audio coding standardization.*

[ITU-T P.10] Recommendation ITU-T P.10/G.100 (2006), *Vocabulary for performance and quality of service; plus its Amendments.*

[ITU-T P.56] Recommendation ITU-T P.56 (1993), *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.501] Recommendation ITU-T P.501 (2009), *Test signals for use in telephonometry.*

[ITU-T P.581] Recommendation ITU-T P.581 (2009), *Use of head and torso simulator (HATS) 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.800.1] Recommendation ITU-T P.800.1 (2006), *Mean Opinion Score (MOS) terminology.*

[ITU-T P.810] Recommendation ITU-T P.810 (1996), *Modulated noise reference unit (MNRU).*

[ITU-T P.830] Recommendation ITU-T P.830 (1996*), Subjective performance assessment of telephone-band and wideband digital codecs.*

[ITU-T P-Sup.23] Recommendations ITU-T P-series – Supplement 23 (1998), *ITU-T coded-speech database.*

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:

3GPP2	3rd Generation Partnership Project 2
4G	4th Generation
AAC	Advanced Audio Coding
AAC+	High Efficiency Advanced Audio Coding
ACR	Absolute Category Rating
AMBE	Advanced Multi-Band Excitation
AMR	Adaptive Multirate Codec
AMR-WB	Adaptive Multirate Codec – Wideband
CDMA	Code Division Multiple Access
EFR	Enhanced Full Rate codec
ERP	Ear Reference Point
EVRC	Enhanced Variable Rate Codec
FD	Fractal Dimension
FFT	Fast Fourier Transform
GSM	Global System for Mobile Communications
HATS	Head And Torso Simulator
HD	High Definition
HFRP	Hands-Free Reference Point
IP	Internet Protocol
IRS	Intermediate Reference System
LTE	Long Term Evolution
MOS	Mean Opinion Score
MOS-LQO	MOS – Listening Quality Objective
MP3	MPEG-1 audio layer 3
NB	Narrowband

OVL Overload Level

PCM Pulse Code Modulation

PESQ Perceptual Evaluation of Speech Quality

POTS Plain Old Telephone System

PSQM Perceptual Speech Quality Measure

PSTN Public Switched Telephone Network

rmse Root Mean Square Error

SLA Service Level Agreement

SNR Signal to Noise Ratio

SPL Sound Pressure Level

SQUAD Speech Quality Analysis Device

SWB Super-wideband

TETRA Terrestrial Trunked Radio

UCC Unified Communication & Collaboration

UMTS Universal Mobile Telecommunications System

VAD Voice Activity Detection

VoIP Voice over IP

WB Wideband

WCDMA Wideband Code Division Multiple Access

5 Conventions

This Recommendation does not use specific conventions.

6 Overview of the ITU-T P.863 algorithm

The ITU-T P.863 algorithm compares a reference signal X(t) with a degraded signal Y(t) where Y(t) is the result of passing X(t) through a communications system. The output of the ITU-T P.863 algorithm is a prediction of the perceived quality that would be given to Y(t) by subjects in a subjective listening test.

In a first step, both the reference signal and the degraded signal are split into very small time slices; these are referred to in the following as frames. Then, the delay of each frame of the reference signal relative to the associated frame of the degraded signal is calculated. Based on the delays found, the sample rate of the degraded signal is estimated; if this estimated sample rate differs significantly from the sample rate of the reference signal, whichever signal has the higher sample rate will be down sampled, and the delay is re-determined.

Based on the set of delays that are found, the ITU-T P.863 algorithm compares the reference (input) signal with the aligned degraded (output) of the system under test using a perceptual model, as illustrated in Figure 1. The key to this process is the transformation of both, the reference and degraded signals to an internal representation that is analogous to the psychophysical representation of audio signals in the human auditory system, taking into account the perceptual frequency (Bark) and the loudness (Sone). This is achieved in several stages:

- time alignment,
- level alignment to a calibrated listening level,

- time-frequency mapping,
- frequency warping, and
- compressive loudness scaling.

The ITU-T P.863 algorithm is designed to take into account the impact of the play back level for the perceived quality prediction in super-wideband mode; the playback level is calculated relative to a nominal level of –26 dBov, which represents 73 dB(A) SPL in diotic presentation.

In narrowband operational mode the ITU-T P.863 algorithm is designed to determine the listening speech quality at a constant listening level of 79 dB(A) SPL.

The internal representation level is being processed in order to take into account effects such as local (i.e., rapid) gain variations and linear filtering; such effects may – if they are not too severe – have little perceptual significance. The compensation is only partially carried out and factors which had been calculated by the model for previous frames are taken into account by a moving average strategy; thus, the compensation lags behind the effect. Thus, minor steady-state differences between reference and degraded signal are being compensated. More severe effects, as well as rapid variations, are compensated only partially; this means that a residual impact remains, which contributes to the overall perceptual disturbance.

Furthermore, the ITU-T P.863 algorithm eliminates low levels of noise in the reference signal, while noise in the degraded output signal is also partially suppressed. Operations on the reference signal that change the characteristics of this signal find its justification in an idealization process that subjects usually carry out in their quality judgement. This idealization was identified and modelled on the basis of the subjective test results, where some reference recordings obtained lower subjective scores when the timbre was not optimal or when low levels of noise were present in the recordings. All subjective testing is carried out without a direct comparison to a reference signal (absolute category rating) and consequently, the ideal signal assumption that is made by the individual subject and upon which his or her opinion is based is unknown during the test. Algorithmic operations on the degraded signal, which changes its characteristics before an internal difference function is calculated, are justified by high-level cognitive processes. Well-known examples are the relative insensitivity to linear frequency response distortion and to steady state wideband (WB) noise; such modelling approaches allow one to describe the final quality perception by a small number of quality indicators to be used to model all related subjective effects. In ITU-T P.863, six quality indicators are being computed in the cognitive model:

- a frequency response indicator (FREQ),
- a noise indicator (NOISE),
- a room reverberation indicator (REVERB), and
- three indicators describing the internal difference in the time-pitch-loudness domain.

These indicators are combined to give an objective listening quality MOS. ITU-T P.863 expects always a clean (noise-free) reference signal. The ITU-T P.863 algorithm is based on a further development of the underlying concepts of [b-Beerends 1994], [b-Rix], [b-Beerends 2002], [b-Beerends 2007] and [b-SwissQual].

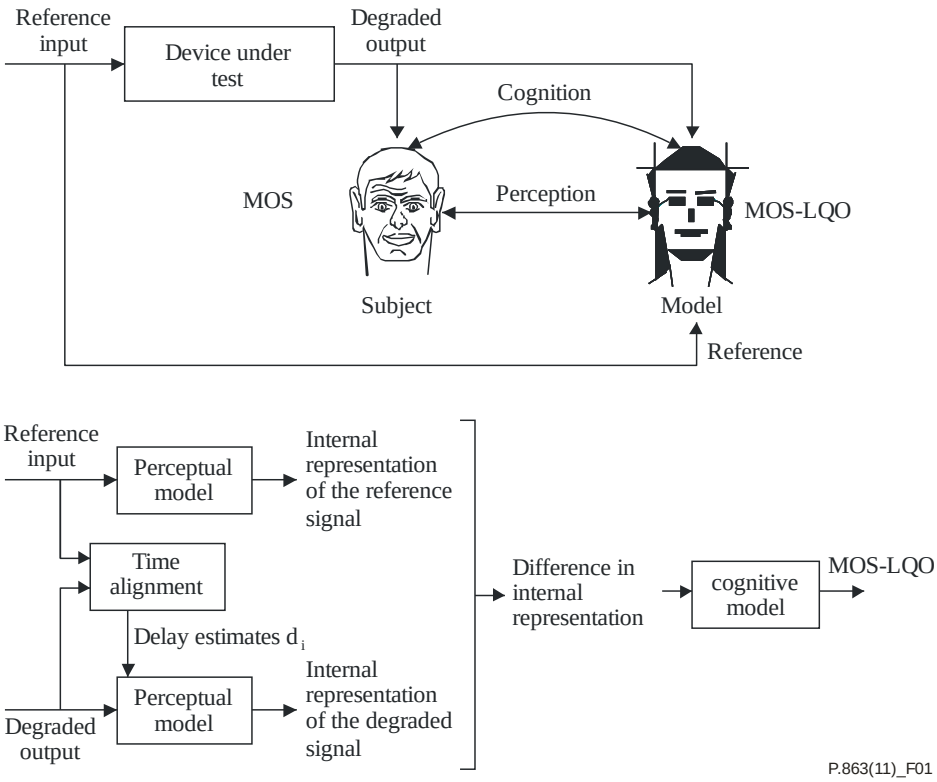


Figure 1 – Overview of the basic philosophy used in ITU-T P.863

7 Comparison between objective and subjective scores

[ITU-T P.800] describes in detail how a subjective listening quality experiment has to be conducted to derive reliable results. However, the results derived in one subjective experiment reflect the relative quality between the included speech samples; the absolute values can vary from experiment to experiment and are dependent on the listeners group and – more importantly – on the design of the subjective test. Subjective votes are influenced by many factors, such as the preferences of individual subjects and the context (the other conditions) of the experiment.

In contrast, an objective measure is independent of the test context and the individual behaviour of the listening panel. It reflects an average test scored by an average group of listeners. An objective model is not able to reproduce exactly the absolute scores of an individual experiment; however, it usually reproduces the relative quality ranking.

For evaluating the accuracy of an objective measure, a comparison to subjective scores is required. To avoid a negative influence of the deviations of individual subjective tests, a scale matching has to be applied before the prediction error of the objective method is calculated. For that reason, an optimal mapping function between the subjective and objective scores is calculated and applied. The sole purpose of this regression is to compensate differences between individual subjective experiments as offsets or individual use of the scale by the subjects. The regression does not change the rank-order, it just adapts the scales of the objective measure and the individual subjective experiment. Therefore, the regression must be monotonic so that information of the rank-order is preserved. The regression normally is used to map the objective ITU-T P.863 score onto the subjective score. It should be noted, however, that this procedure may also compensate for some of the systemic prediction errors caused by the objective measurement method.

A good objective quality measure should have a high correlation with many different subjective experiments after such a regression is performed separately for each experiment. In practice, with the ITU-T P.863 algorithm, the regression mapping is often almost linear, using a MOS-like five point scale. The preferred metric for determining the accuracy of an objective method is the root mean

square error (rmse) between the scores obtained in a subjective experiment and their predictions by an objective method. The mapping function and the rmse calculation are only necessary for determining the accuracy of an objective method compared to a specific subjective experiment. In daily practice, no special mappings to the objective scores have to be applied, since the ITU-T P.863 scores are already mapped to a MOS scale reflecting the average over a huge amount of individual data sets.

8 Speech material

The given restrictions and recommendations are made not to narrow the scope of ITU-T P.863, but rather to guarantee reliable and comparable scores. In general the ITU-T P.863 algorithm will provide similar performance when speech material is used that does not exactly match the requirements in this clause.

8.1 Recommendations on source speech material

Reference speech or *reference signal* is the original speech signal without any degradation. This should be recorded and stored using the information provided in Appendix II. In the case of an acoustical sending path, this signal is used for feeding the artificial mouth. This speech signal is used by the ITU-T P.863 algorithm as a reference against which the effects of the system under test are revealed. Examples of such speech files are to be found in [ITU-T P.501].

The ITU-T P.863 algorithm was tested with human speech material. For a consistent speech quality prediction, active speech parts and speech pauses are required in the speech sample. It is recommended to use typical spoken sentences with typical syllable and word structures. It is not recommended using single word samples only (i.e., counting).

The reference signals should be levelled to –26 dBov. Other signal levels will be accepted by the ITU-T P.863 algorithm too and internally equalized. Independent of the ITU-T P.863 algorithm, signals with different levels than –26 dBov could either be amplitude clipped or affected by lower SNR. For the test speech samples consisting of two male and two female speakers, two sentences each, are recommended.

8.1.1 Sampling frequencies and filtering

It is mandatory that reference and degraded signal have the same sample rate.

Super-wideband operational mode:

In *super-wideband operational mode* the ITU-T P.863 algorithm always requires a super-wideband reference signal. It has to be provided in mono at a sampling frequency of 48 kHz. The reference signal to be used in super-wideband mode has to have no bandwidth limitation and no additional equalization between 50 and 14 000 Hz. The reference signal must not have significant energy outside the super-wideband spectral limits (50 to 14 000 Hz). An example flat band-pass filter is recommended in [ITU-T G.191], 14 kHz.

If the super-wideband mode is chosen and the sample rate of the signals differ from 48 kHz or 8 kHz (e.g., 32 kHz or 16 kHz), the signals will be sample rate converted internally to 48 kHz before processing.

NOTE – The use of reference signals with bandwidth narrower than 14 kHz in super-wideband mode, regardless of sampling rate, will lead to wrong results on the super-wideband MOS-scale.

Narrowband operational mode:

For *narrowband operational mode*, use of a sampling frequency of 8 kHz is preferred. If the narrowband mode is chosen and the samplerate of the signals differs from 8 kHz, the signals will be samplerate converted to 8 kHz. The reference signal to be used in narrowband mode has to have no

bandwidth limitation and no additional filtering in between 100 to 3 800 Hz. It is recommended that there is no significant energy outside this range.

Table 6 – Super-wideband filter definition

Frequency (Hz)	Super-wideband gain (dB)
20	–20 (max)
50	–3
60	0
13 500	0
14 000	–3
15 000	–40 (max)
24 000	–50 (max)

The approximate narrowband filter definition is given in Table 7.

Table 7 – Narrowband filter definition

Frequency (Hz)	Narrowband gain (dB)
20	–20 (max)
100	–3
110	0
3 750	0
3 800	–3
4 000	–40 (max)

8.1.2 Temporal structure and recording requirements

Principle rules for speech material used in the evaluation phase of the ITU-T P.863 algorithm and the recommended use of ITU-T P.863 are:

- At least one silence interval between speech segments (sentences) ≥ 0.5 s.
- For testing background noise conditions: At least one silence interval between speech segments (sentences) ≥1.0 s and ≤ 2.0 s.
- The minimum amount of active speech in each file should be 3 s.
- Reference speech files should have sufficient leading and trailing silence intervals to avoid temporal clipping of the speech signal, e.g., 200 ms of silence each.
- For super-wideband reference speech samples the noise floor of the reference files should not exceed –84 dBov(A) in the leading and trailing parts as well as in the gaps between the sentences.
- 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 omni-directional microphones. The distance to the microphone must be approximately 10 cm. Background noise must be below 30 dB(A)SPL. Speech signals will be bandpass filtered from 20 Hz to 14 kHz. Directional microphones are allowed on the condition that the frequency response is the same as with the before-mentioned, omni-directional microphones.

The ITU-T P.863 algorithm is known to produce accurate results for such signals. Different temporal structures of speech material are subject to further investigation in the characterization phase of the algorithm.

8.2 Insertion of source speech material into the system under test

The reference signal may be further processed before it is inserted in the transmission channel, either by 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. These processing steps are considered part of the system under test, but the reference signal for the ITU-T P.863 algorithm remains unprocessed, as described above.

In the case of an electrical insertion, the reference signal could be given to a model of the sending terminal, which may reflect narrowband or super-wideband behaviour of the terminal to be modelled (i.e., IRS send). For insertion to proprietary interfaces, a pre-filtering can be applied.

Acoustical insertion applies to handset or hands-free devices. The test set up has to follow [ITU-T P.340] or [ITU-T P.581] or other realistic use cases.

NOTE – The consequences of these definitions and recommendations are that each captured sequence will be measured against a flat and noise-free reference in the case of narrowband and super-wideband operational modes. If the actually used insertion path is not flat, this deviation will be taken into account by the speech quality model.

8.3 Recommendations on processed and degraded speech material

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

8.3.1 Sampling filtering and operational modes

In *super-wideband operational mode,* degraded signals have to be provided in mono and at a sampling frequency that allows the representation of the full spectral range of super-wideband (50 to 14 000 Hz, i.e., 48 kHz). A lower sampling frequency for recording is allowed as far as the desired bandwidth of the application is covered (i.e., 8 kHz or 16 kHz for narrow-band channels or 16 kHz for wideband channels). Note that this approach then requires an external up-sampling of the degraded signal to 48 kHz before the comparison can be made to the 48 kHz sampled reference signal.

The degraded signal must not have significant energy outside the super-wideband spectral limits (50 to 14 000 Hz). This can be achieved by applying a band-pass filter according to [ITU-T G.191], 14 kHz.

In *narrowband operational mode,* where 8 kHz samplerate is the preferred choice, the sampling frequency can be lower than required samplerate for super-wideband mode. For all recordings to be scored in narrowband operational mode, a pre-filtering with a lowpass of 3 800 Hz is required if the recording sampling frequency is above 8 kHz.

NOTE – The detailed description of the *ITU-T P.863algorithm* (see Annex B) assumes the same sampling frequency for reference and degraded signal. It considers sampling frequencies of 8 and 48 kHz.

8.3.2 Recording and presentation level

In the *super-wideband operational mode,* differences between the nominal presentation level and the actual presentation level are allowed and are assessed as part of the test condition. The nominal presentation level is 73 dB(A) SPL at the ERP (ear reference point) of both ears. Level differences have to be restricted to a range of +5 dB to –20 dB relative to the nominal level.

For all signals intended to be used for super-wideband measurements, a digital level of –26 dBov (obtained as an aggregated average rmse value across the active speech parts) corresponds to the nominal presentation level (73 dB(A) SPL 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 nominal level for the ITU-T P.863 algorithm in super-wideband operational mode is therefore –26 dBov. The levels of the degraded speech material used to validate the ITU-T P.863 algorithm, were in a range from –21 dBov to –46 dBov.

The *narrowband operational mode* of the ITU-T P.863 algorithm presupposes that degraded speech signals are scaled towards a digital level of –26 dBov (obtained in accordance with [ITU-T P.56]). Those signals are predicted under the assumption of a monotic presentation using an IRS(rcv) handset and 79 dB(A) (SPL) at the ERP. Severe deviations, such as non-optimal presentation levels, from this nominal level are taken into account and were not part of the validation of the ITU-T P.863 algorithm. The evaluation of non-optimal presentation levels in narrowband operational mode is subject of further studies in the characterization phase of this algorithm.

NOTE – It is recommended to scale signals towards a correct relationship between digital level and presentation level in the acoustic domain. It means signals presented (or assumed to be presented) by 79 dB(A) SPL in narrow-band mode or by 73 dB(A) SPL in diotic super-wideband mode should be scaled to –26 dBov.

8.4 Special requirements for acoustical captured speech material

When an acoustical device is tested and a direct relation is needed between the measurement and the device under test as perceived in diffuse field 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 head-and-torso-simulator (HATS) [ITU-T P.581] has to be used. For tests where no direct acoustical match is required, recordings can be made as needed for the application under test. The playback device in the super-wideband 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 or 32 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.

In the case of an acoustically captured signal, the user terminal is part of the recording and its influence will be scored as well. Obvious differences between different play-out levels (caused by the real receiving terminal) should be taken into account for the quality scoring in the super-wideband test scenarios. It is highly recommended that the acoustical recordings be scored by the ITU-T P.863 algorithm with the actual sound levels used during recording.

Level differences in a test are restricted to a range of +5 dB to –20 dB relative to the nominal level of the test application. The nominal level is represented by a digital 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 the 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 a digital signal of –26 dBov is presented at 73 dB (SPL) at each ear.

NOTE 1 – The artificial ear is not specified above 12 kHz. Any potential impairments are seen as part of the device under test and the ITU-T P.863 algorithm will weight them accordingly.

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, (HFRP)) and the artificial ear, can be used with the ITU-T P.863 algorithm. In the same way, simulations of the acoustical insertion can be used.

8.5 Acoustical insertion/capture for loudspeaker phones

- The hands-free loudspeaker conditions may be recorded in different types of rooms or cars, representative of hands-free telephony.

8.6 Technical requirements on signals to be processed by ITU-T P.863

- Standard PCM 16-bit linear, little-endian byte order.
- Sampling frequencies of 48 kHz, 16 kHz, or 8 kHz are accepted. Note the requirements for the two operational modes.
- Evaluating common 16 kHz wideband scenarios requires super-wideband operational mode and 48 kHz sampling frequency. The reference signal has to be provided as a super-wideband signal (sampling frequency 48 kHz, bandpass filtered 50-14 000 Hz).
- *Super-wideband operational mode* requires 48 kHz sampling frequency. Degraded signals can be recorded at 16 kHz or 8 kHz if that is sufficient for the scenario to be tested (i.e., POTS). Before use, such signals have to be up-sampled to 48 kHz. If a narrowband signal is up-sampled to 48 kHz and compared with a super-wideband reference, the resulting MOS-LQO represents the quality of the narrowband signal in a super-wideband experiment.

8.7 Predicted scores by the model

The main result of the ITU-T P.863 algorithm is a single value describing the overall speech quality MOS-LQO [ITU-T P.800.1] on a 1 to 5 MOS scale. The MOS-LQO is linearized to a wide range of subjective tests and associated results.

NOTE 1 – The ITU-T P.863 algorithm scores saturate at MOS-LQO = 4.75 in super-wideband mode and at MOS-LQO = 4.5 in narrow-band mode. It simply reflects the fact that not all subjective test participants will give the highest rating – even for the undegraded reference.

NOTE 2 – The ITU-T P.863 algorithm is able to reproduce the relative quality ranking of individual tests with a small residual prediction error. Details about the expected prediction error are given in Annex A in terms of rmse. The ITU-T P.863 algorithm cannot predict the exact MOS of individual subjective experiments. It predicts on a scale which is a compromise between many subjective tests.

9 Description of the ITU-T P.863 algorithm

Because many of the steps in the ITU-T P.863 algorithm are algorithmically quite complex, a description is not easily expressed in mathematical formulae. The following description is textual in nature and the reader is referred to the detailed descriptions (see Annex B) for further advice.

Figures 2 through 11 give an overview of the algorithm with block diagrams:

- Figure 2 for the overview,
- Figures 3 through 6 for the temporal alignment,
- Figures 7 through 9 for the core of the perceptual model,
- Figure 11 for the masking, and
- Figure 10 for the final determination of the ITU-T P.863 score.
- For each of the blocks a high-level description is given.

9.1 Overview

A general overview on the algorithm is shown in Figure 2. The inputs to the algorithm are two waveforms represented by two data vectors containing 16 bit PCM samples. The first vector contains the samples of the (undistorted) reference signal, whereas the second contains the samples of the degraded signal. The ITU-T P.863 algorithm consists of a sample rate converter, which is used to compensate for differences in the sample rate of the input signals, a temporal alignment block, a sample rate estimator, and the actual core model, which performs the MOS calculation. In a first step, the delay between the two input signals is determined and the sample rate of the two signals relative to each other is estimated. The sample rate estimation is based on the delay information calculated by the temporal alignment. If the sample rate differs by more than approximately 0.5%, the signal with

the higher sample rate is down-sampled. After each step, the results are stored together with an average delay reliability indicator, which is a measure for the quality of the delay estimation. The result from the resampling step, which yielded the highest overall reliability, is finally chosen. Once the correct delay is determined and the sample rate differences have been compensated, the signals and the delay information are passed on to the perceptual model, which calculates the perceptibility as well as the annoyance of the distortions and maps them to a MOS scale.

The corresponding parts of the algorithm (see Annex B) can be found in:

DoCalculateDelayDegPlus(…), in file: OptInterface.pdf

– The link between the perceptual model and the temporal alignment and;

DoAlignmentPlus(…) , in file: OptInterface.pdf,

– The resampling loop.

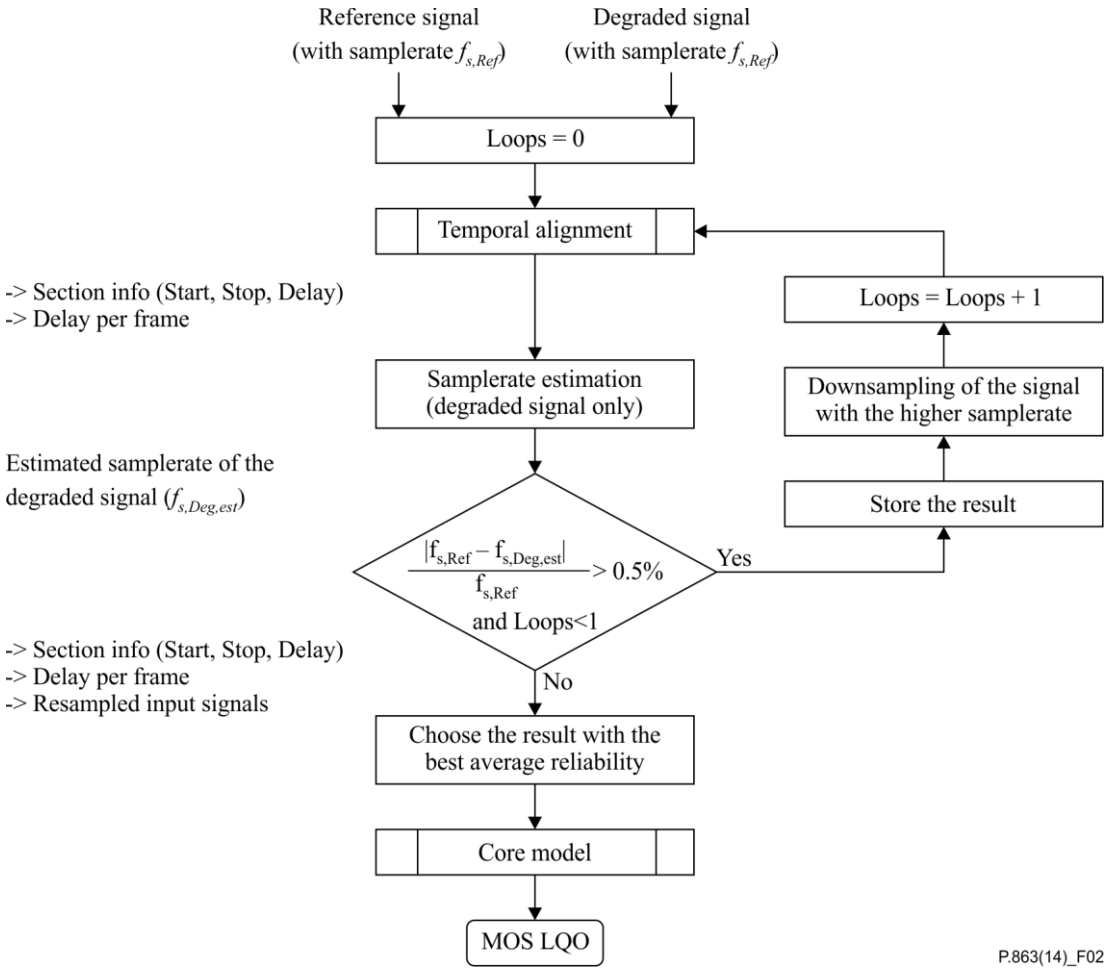


Figure 2 – General overview of the ITU-T P.863 algorithm

9.2 Temporal alignment

The basic concepts of the temporal alignment are:

- To split the degraded signal into equidistant frames and to calculate a delay for each frame. The delay represents the offset in samples at which the best matching reference signal section can be found.
- Whenever possible, the matching counterparts of the degraded signal sections are searched for in the reference signal and not vice versa.

- Stepwise refinement of the delay per frame to avoid long search ranges (long search ranges require high computational power and are critical in combination with time scaled input signals).

The temporal alignment consists of the major blocks filtering, prealignment, coarse alignment, fine alignment, and section combination. The degraded input signal is split into equidistant macro frames, the length of which is dependent on the input sample rate. The delay is determined for each "macro frame". The calculated delay is always the delay of the reference signal relative to the degraded signal (the algorithm does in general perform a search for sections of the degraded signal inside the reference signal). The prealignment determines the active speech sections of the signals, calculates an initial delay estimate per macro frame and an estimated search range required for the delay of each macro frame (a theoretical minimum and maximum delay variation of the detected initial delay). The coarse alignment performs an iterative refinement of the delay per frame, using a multidimensional search and a Viterbi-like backtracking algorithm to filter the detected delays. The resolution of the coarse alignment is increased from step to step in order to keep the required correlation lengths and search ranges small. The fine alignment finally determines the sample exact delay of each frame directly on the input signals, with the maximum possible resolution. The search range in this step is determined by the accuracy of the last iteration of the coarse alignment. In a final step, all sections with almost identical delays are combined to form the so-called "Section Information".

This temporal alignment procedure has the following characteristics:

- There is no hard limit for the static delay.
- It is designed to handle a variable delay of less than 300 ms around the static delay, but no hard limit exists.
- The delay may vary from frame to frame.
- Small sample rate differences (less than approx. 2%) can be handled well, larger differences will be detected (and compensated outside of the temporal alignment).
- Time stretched or temporally compressed signals with or without pitch correction are handled well.
- The alignment works well even under very noisy conditions with an SNR below 0 dB.
- No problems were observed with signal level variations.
- The corresponding parts of the algorithm can be found in:
- CTempAlignment::Run(…), in file: TempAlignment.pdf (see Annex B).

9.2.1 General delay search method

Most of the modules related to the temporal alignment use the same method to find the lag (=delay, offset) between two signals. This method is based on the analysis of a histogram which is created by calculating the cross-correlation function between two signals, entering the found peak value into the histogram, shifting both signals by a small amount and repeating this step again. Once the histogram contains enough values, it is filtered and the peak is determined. The position of this peak in the histogram is equivalent to the delay offset between the two signals.

9.2.2 General delay reliability measure

In most steps of the temporal alignment the simple Pearson correlation is used as a measure of the reliability for a found delay between two signals.

9.2.3 Bandpass filter

Before any further step, both input signals are bandpass filtered. The filter shape is depending on the operating mode of the model (narrowband or super-wideband).

In the super-wideband operating mode, the signals are bandpass filtered to 320 Hz up to 3 400 Hz.

In the narrowband operating mode, the signals are bandpass filtered to 290 Hz up to 3 300 Hz.

NOTE – Those filtered signals are used for the temporal alignment, only. The perceptual model uses differently filtered signals.

9.2.4 Prealignment

The prealignment first identifies reparse points in the degraded signal. Reparse points are positions in the signals where the signal makes a transition from speech pause to active speech. The reparse points mark the beginning of active speech sections, while reparse sections describe the entire active speech segment beginning at a reparse point. For each such reparse point the reparse section information is calculated. This section information stores the position of the beginning and end of the section, as well as an initial value for the delay of said section, an indication of the reliability of the found delay and its accuracy, i.e., an upper and lower bound within which the accurate delay is expected to be found. The general process is depicted in Figure 3, the steps of which are explained in the subsequent chapters.

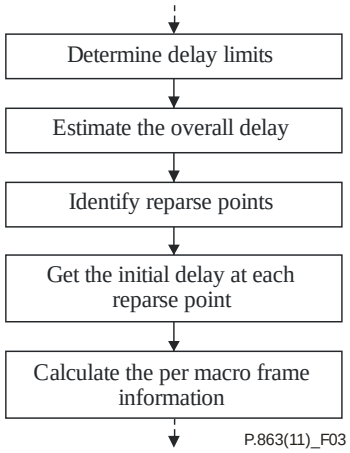


Figure 3 – Overview of the prealignment

The corresponding parts in of the algorithm can be found in:

CTempAlignment::RunPrealignment(…), in file:TempAlignment.pdf (see Annex B).

9.2.4.1 Determine the delay limits

This simple step tries to identify some reasonable upper and lower limit for the overall delay search range. The decision is based on the following assumptions:

- The reference file has at least 40% activity and consists of at least two sentences.
- The total amount of silence is split into at least two sections (typically three).
- Not more than 50% of the silence fall before the start or after the end of the file.
- The active speech part is not cut off at either end due to the delay.

This results in a search range according to the following formulae:

$$DelayHigh = \max\left(2500ms, Startsample_{\mathrm{Ref}}\right) \tag{9-1}$$

$$DelayLow = \max\left(2500ms, -\left(F_{len,\mathrm{Ref}} * 0.2 + F_{len,Deg} - F_{len,\mathrm{Ref}}\right) * 0.8\right) \tag{9-2}$$

With $F_{len,\ \mathrm{Ref}}$ being the length of the reference signal and $F_{len,\ Deg}$ being the length of the degraded signal in milliseconds. Startsample is the detected start of the reference signal; this is usually 0, but may be a later point in the signal, if it starts with a very long silent period.

9.2.4.2 Overall delay estimation

The overall delay is estimated in three steps. First, an attempt is made to match the entire signals using a histogram of cross-correlation functions. This results in the overall delay and the overall reliability. In a second and third step, the same is tried for the first and for the second half of the signals independently, which again results in two more values for delay and reliability. If all three delays are of the same range, then the overall delay is accepted and marked as reliable. The tolerance for acceptable delays is reduced for long signals with poor reliability values.

The entire process operates on the logarithmic energy densities of the signals, averaged over frames with a sample rate depending length as indicated in Table 8. This frame size also defines the best achievable accuracy in this step.

NOTE – For time-scaled signals, the estimation of the overall delay will be very inaccurate and most of the time even unusable. However, if this estimate proves reliable, the efforts for all subsequent alignment steps can be reduced drastically.

Table 8 – Frame sizes used in the overall delay estimation

Sample rate	48 kHz	16 kHz	8 kHz
Frame length [samples]	512	256	128

The corresponding parts of the algorithm can be found in:

CTempAlignment::EstimateOverallDelaySimpleLimits(…), in file: TempAlignment.pdf (see Annex B).

9.2.4.3 Identification of reparse points

In this step, a voice activity detection algorithm (VAD) is used to determine for each macro frame of the reference and the degraded signal if it contains active speech or silence. The beginning of sections of consecutive active macro frames is called a reparse point, since those resemble the points at which the delay measurement is completely restarted. The so-called reparse section information contains for each reparse point the position of the active section's start point, the length of the active section, a coarse delay estimate, a reliability indicator for the detected delay and an estimated range within which the exact delay can most likely be found. The delay determined in this step is simply the difference of the detected reparse section start in the degraded signal minus the startpoint of the according section in the reference signal, and may be very unreliable.

For the reference signal, the active frame detection works very reliably, but especially for noisy signals, the detected active sections may be very inaccurate. Therefore, the plain VAD information is, by far, not sufficient in order to allocate combinations of reference and degraded active sections. Instead, a rather complex method is used. Up to three sets of potential allocations are investigated for each reparse point:

VAD1: A match, plainly based on the VAD information. If the length of the next reference and degraded section does not differ by more than 120 ms for signals with an SNR below 35 dB, or if the length differs by less than 480 ms for signals with better SNR, then the VAD1 section information is marked as valid.

Corr1: The reference section is searched in the degraded signal by using the VAD information of the reference signal only and the overall delay estimate as a hint on where to search. The search is performed twice, once at the beginning of the active section and once slightly delayed. The best of the two results is stored. If sections were long enough and a reasonable match could be found, then the Corr1 information is marked as valid.

Corr2: The reference section is searched in the degraded signal by using the VAD information of the reference signal and the VAD information of the degraded signal as hints on where to search. The

search is performed twice, once at the beginning of the active section and once slightly delayed. The best of the two results is stored. If sections were long enough and a reasonable match could be found, then the Corr2 information is marked as valid.

Now, the best of Corr1 and Corr2 is chosen. If the found match is very reliable and the found degraded section is much longer than the reference section, it will be split and a new degraded section is added to the list of sections that must still be allocated. This result is stored as the default solution.

If neither Corr1 nor Corr2 could be calculated reliably, then the algorithm tries to modify the sections to make them fit by e.g., splitting one large into two small sections. This is inherently dangerous and the result will only be used as a last resort. The matching result will be stored as VAD2.

If this did also not lead to a resolution, a best effort approach is applied, by using the overall delay information.

Now a final check is performed if either the VAD2 match or the overall delay match resulted in a better solution than the correlation based approach and the best matching result is used.

If the best solution so far is still very poor and the VAD1 match was good, that one is taken as a fall back solution.

After this, a final check for overlapping sections, exceeded limits and plausibility is performed. If any issues could not be resolved, the sections are reallocated simply by using the overall delay information, although that delay might be invalid too.

All operations in this clause are performed on the energy densities of the signals, using frames of the same size as in the overall delay estimation (see Table 8).

The corresponding parts of the algorithm can be found in:

CTempAlignment::IdentifyReparsePoints(…), in file: TempAlignment.pdf (see Annex B).

9.2.4.4 Initial delay calculation at each reparse point

So far, the delay of each reparse section is very coarse, since it is mostly based on the information retrieved from the VAD. For each active section this delay is now refined by a multidimensional search (using the two different features log (energy density) and fractal dimension) over two segments from each reparse section and using two different frame sizes. This results in eight proposals for the delay of each section. The proposal with the highest reliability is chosen. If the reliability of a section, as determined by the reparse point identification, is already very high, the search is skipped entirely for this section.

The corresponding parts of the algorithm can be found in:

CTempAlignment::GetInitialDelaysInSamples(…), in the file: TempAlignment.pdf (see Annex B).

9.2.4.5 Determination of active frame flags from the reparse section information

In principle, the VAD information of the degraded signal is used directly in order to mark individual macro frames as active or pause. In addition, however, all frames which are outside any detected reparse section are fixed set to pause, regardless of the VAD information. This avoids the wrong treatment of such sections for very noisy signals, where the VAD information might be misleading. The result is a vector which contains for each macro frame a flag which is set when the frame is active and it is cleared when the frame is a speech pause.

The corresponding parts of the algorithm can be found in:

CTempAlignment:: SetActiveFrameFlags (…), in the file: TempAlignment.pdf (see Annex B).

9.2.4.6 Creation of per frame information from the reparse section information

This process is rather simple. All it does is copy the information from the reparse section information to the according per frame information, making sure that delay changes occur in the middle of the pause between two active sections. This step generates the following vectors:

DelayPerFrame, contains for each macro frame the estimated initial delay.

ReliabilityPerFrame, contains for each macro frame an indication of the reliability of the delay estimate.

SearchRangeLow, contains for each macro frame the lower bound of the range in which the exact delay is expected.

SearchRangeHigh, contains for each macro frame the upper bound of the range in which the exact delay is expected.

All operations after the prealignment perform a stepwise refinement of those vectors.

The corresponding parts of the algorithm can be found in:

CTempAlignment::ReparseSections2DelayVector(…), in file: TempAlignment.pdf (see Annex B).

9.2.5 Fast prealignment method for signals with fixed or piecewise fixed delays

The ITU-T P.863 algorithm includes an alternative prealignment method which allows for fast alignment of signals with fixed or piecewise fixed delays. This method is used once during the first iteration through the time alignment loop. The processing steps can be summarized as follows:

a) Apply a further bandpass filter on the input signals with lower and upper passband edges of 700 and 3 000 Hz, respectively. The signals are then downsampled to 8 kHz sampling frequency.

b) Compute the average active speech level *speechLev* and average noise level *noiseLev* of both signals after filtering. The signals are then scaled to an average active speech level of –26 dBov. The values *speechLev* and *noiseLev* of each signal are scaled accordingly.

c) Compute a threshold level *thr* for active speech. The threshold level is defined by

$$thr = \min\left(\frac{-26\,\text{dBov} + 3\cdot\max\left(noiseLev_{ref}, noiseLev_{\text{deg}}\right)}{4}, -29\,\text{dBov}\right) \qquad (9\text{-}3)$$

d) Compute an estimation of the overall delay by cross-correlating the signal envelopes, where each envelope frame represents the signal level in a 180 ms piece of the respective signal, and the range of calculated cross-correlations is selected such that the degraded signal overlaps with at least 1/4th of the length of the reference signal. The overall delay estimation is then derived from the maximum cross-correlation value found.

e) Perform segment-wise matching of reference signal parts with corresponding degraded signal parts. A segment is defined as a consecutive piece of signal with an envelope level exceeding a set threshold. The matching process starts with the reference segment that has the highest total level, then continues with the next-highest, etc. until all consecutive pieces with an envelope level exceeding the set threshold have been matched. The threshold is then lowered to the value *thr* determined in c) and the matching continues until a delay has been determined for all active speech in the reference signal. The delay of speech pauses is then estimated from the delays of surrounding active speech parts.

f) Use the result from step e) to derive the reparse section information, as well as the *DelayPerFrame*, *ReliabilityPerFrame*, *SearchRangeLow* and *SearchRangeHigh* vectors.

A more detailed description of steps e) and f) is provided in the following paragraphs.

The corresponding parts of the algorithm can be found in:

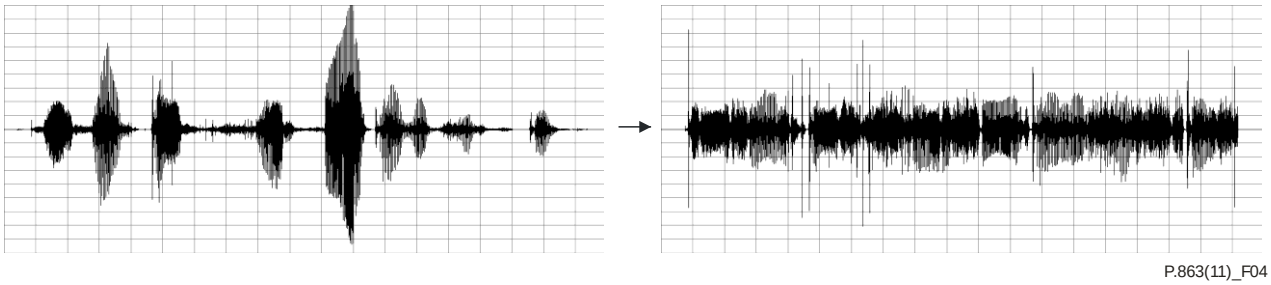
SQTimeAlignment::SQTimeAlignment(…), in file: SQTimeAlignment.pdf (see Annex B).

9.2.5.1 Segment-wise matching of reference signal parts

This step attempts to determine a fixed delay for all segments in the reference signal.

Note that in the context of the fast prealignment, the term envelope refers to the signal envelope of a bandpass filtered and downsampled copy of the reference or degraded signal. After computation of the envelopes, both signal copies are further processed using sliding window power normalization.

This processing utilizes a sliding window of 26.625 ms length as well as the active speech threshold level *thr* determined earlier. The normalization process sets a signal sample to zero if the signal level of the samples in the sliding window, with its centre positioned at the current sample, is equal or inferior to *thr*. Otherwise, the value of the signal sample is normalized (divided) by the root mean square (RMS) value of the samples in the window. A hysteresis is used to preserve an additional 70 ms of samples in the processed signal after the signal level in the window drops to or below *thr*. After normalization, the processed signals are scaled back towards an average active speech level of –26 dBov. Figure 4 shows an example of the effect of the sliding window power normalization:



NOTE – After processing, the speech signal has an almost flat signal envelope during active speech; speech pauses and silent periods longer than 70 ms are clipped to digital silence.

Figure 4 – Sliding window power normalization

Once the normalized signals have been computed, the segment-wise matching loop is entered. The purpose of this loop is to generate a piecewise correspondence list of reference with degraded signal parts. For each entry in the list, the position, length and delay of the signal part is given, along with its type: Assigned, Unsure, Missing or Pause. Figure 5 provides an overview of the operation of the segment-wise matching loop.

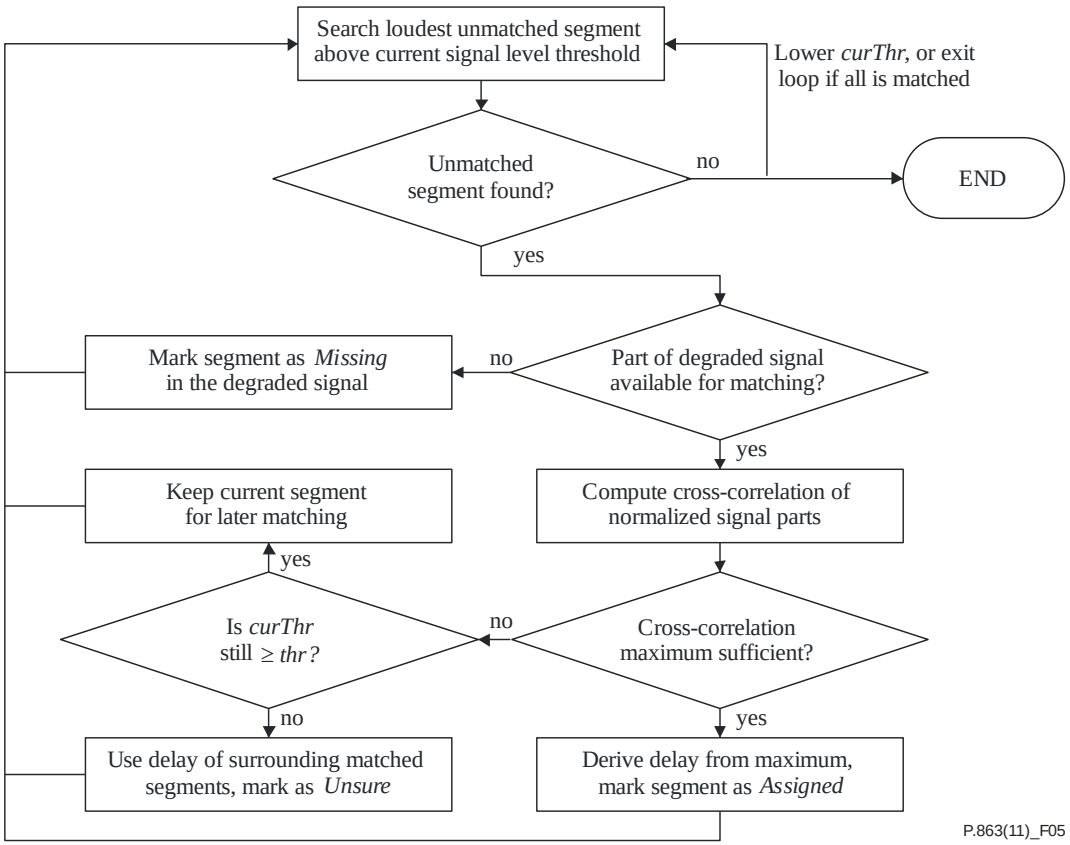


Figure 5 – Overview of the operation of the segment-wise matching loop

The loop starts with a threshold level for segments *curThr* that must be greater than *thr*:

$$curThr = \max\left(thr + 1,\, 0.4 \cdot speechLev_{ref} + 0.6 \cdot noiseLev_{ref}\right) \tag{9-4}$$

The envelope of the reference signal (computed before sliding window power normalization) is then used to select the segment to match. The selection consists in finding a consecutive piece of the reference signal where all envelope frames in the piece have a level ≥ *curThr* and the summed levels of these envelope frames is the highest. For each successfully matched segment, the corresponding envelope frames are set to –80 dBov to prevent their reselection in the following loop iteration.

The operations a) to f) are carried out with each selected segment:

a) Determine the exact sample start and end positions of the current segment in the reference signal. If necessary, the current segment is shortened to a maximum length of 1.5 seconds.

b) Determine the exact sample start and end positions of the part of the degraded signal that is available for matching with the current segment. Any part of the degraded signal may only be matched once with a corresponding part in the reference signal. Additionally, temporal monotonicity must be preserved, i.e., the beginning of the degraded signal is not available for matching with a given reference signal part, if a signal part occurring earlier in the reference was previously matched with a part of the degraded signal after the beginning. If the available degraded signal part is shorter than the current segment length in the reference, matching is only performed for the available length, and the unmatched excess segment length is marked as *Missing* in the correspondence list. If there is no available degraded signal

part at all, the entire segment length is marked as *Missing* in the correspondence list, and the algorithm jumps to the next loop iteration directly.

c) Derive an estimation of the position of the current segment in the degraded signal by using the delay of surrounding, previously matched segments. If no segments were matched yet, the overall signal delay is used. If the surrounding matched segments have different delays, their delays are averaged and weighted with their respective distance to the current segment in the reference signal.

d) Cross-correlate the current segment in the reference with the available degraded signal, using the previously normalized versions of both signals. The available degraded signal is offset such that the estimated position of the current segment in it is located at the zero lag position. The number of computed lags in the cross-correlation is determined by the length of the current segment and available degraded signal, but may not exceed the sample equivalent of 2 seconds (i.e., 8 000 samples in each direction at 8 kHz). This limit may be increased for very long signals, or for signals with large measured overall delay:

$$maxLags = \max\left(8 \cdot \left|overallDelay\right|, 0.25 \cdot duration_{deg}\right), maxLags \in \left[2.0, 6.0\right] \quad (9\text{-}5)$$

e) Derive the actual position and hence delay of the current segment in the degraded signal by searching the maximum absolute value in the obtained cross-correlation vector. Using the absolute value avoids problems with 180 degree phase shifts in the degraded signal. Before the maximum absolute value is searched, the cross-correlation vector is weighted such as to assign a lower weight to values for lags far from the previously selected zero lag position:

$$weight_{lag} = 0.5 + \frac{(1-0.5)}{2} \cdot \left(1 + \cos\left(\frac{\left|lag\right| \cdot \pi}{0.25 \cdot F_s}\right)\right) \quad (9\text{-}6)$$

f) Store the delay determined for the current segment, or keep it for later matching. If the selected maximum absolute value in the weighted cross-correlation vector is deemed high enough and the segment length is at least 64 ms, the current segment is marked as *Assigned* and its delay is stored in the correspondence list. If these conditions are not met, and the segment threshold level *curThr* is greater than the active speech threshold *thr*, the current segment is kept for later matching and no delay is stored. If *curThr* is at or below *thr*, the previously determined estimated delay is stored and the current segment is marked as *Unsure* in the correspondence list. In any of the three cases, the envelope frames of the reference signal corresponding to the current segment are set to –80 dBov to prevent its reselection in the next loop iteration.

Once no signal segments above *curThr* are left in the reference envelope, *curThr* is lowered to *thr* and the loop continues using the lower threshold. Thus, quieter signal parts of the reference are selected for matching. Since surrounding segments with higher signal level have already been matched, the range of possible matches is constrained and the risk of mismatches hence decreased.

Once no signal segments above *thr* are left in the reference envelope, the loop matches consecutive pieces of non-zero samples in the normalized reference signal for which no delay information is stored in the correspondence list. This includes segments previously kept for later matching, as well as very quiet or short parts of the reference signal.

Finally, consecutive pieces of samples equal to zero in the normalized reference signal are selected. Their delay is estimated from the delays of surrounding, previously matched signal parts. This delay is stored and the corresponding entry in the correspondence list is marked as *Pause*.

The correspondence list generated by the segment-wise matching loop is post-processed to detect and correct erroneous matches. This is achieved by splitting the list into two parts corresponding to both halves of the reference signal. In each part, the statistical lower and upper outer fences of the delays of all *Assigned* list entries are calculated. Any list entry for active speech in the reference signal with a delay outside of the calculated fences is then flagged as erroneous match. For each flagged entry, the stored delay is replaced by an estimated delay using the surrounding non-flagged list entries, and the flagged entry is marked as *Unsure*.

9.2.5.2 Computation of reparse section and delay vector information

The correspondence list is used to generate the reparse section information. Each reparse section simply consists of consecutive list entries that do not contain any entry marked as *Pause* longer than 0.1 seconds. Any *Pause* entry longer than this will be skipped and an additional reparse section will be created from the next non-*Pause* list entry on.

Thus, the start point and length of each section is determined by the start point of the first list entry and the total length of all list entries used in the section, respectively. The coarse delay estimate of a section is computed from the weighted average of the delays of the entries used in the section, where the delay of each entry is weighted by its signal length.

Similarly, the reliability indicator of the section is computed using the same weighted average calculation, where the reliability of each *Assigned* list entry is 1 and the reliability of *Unsure* entries is taken from the cross-correlation value obtained in the segment-wise matching loop. *Missing* list entries in the section are not used for the section reliability calculation. Finally, the estimated range for the exact delay of the section is given by the minimum and maximum value of all delays of non-*Missing* list entries used in the section.

The correspondence list is also used to generate the *DelayPerFrame*, *ReliabilityPerFrame*, *SearchRangeLow* and *SearchRangeHigh* vectors. For this, the degraded signal is traversed stepwise in increments of the used macro frame size. For each macro frame, the entry in the correspondence list covering the current position in the degraded signal is found, and the delay vector information for the current macro frame *m* is calculated as follows:

- *DelayPerFrame[m]* is set to the delay of the found list entry.
- *ReliabilityPerFrame[m]* is set to 1 if the found list entry is of type *Assigned*. If the found list entry is of type *Pause* the reliability is set to 0. Otherwise, it is taken from the cross-correlation value obtained in the segment-wise matching loop.
- *SearchRangeLow[m]* and *SearchRangeHigh[m]* are both set to 0 if the found list entry is of type *Assigned*. This dramatically reduces the number of calculations and thus the processing time in the following Coarse Alignment step. If the found list entry is of type *Unsure*, the search range for the exact delay is given by the delay difference of neighbouring *Assigned* entries before and after the current list entry. Additionally, if the current *Assigned* entry directly precedes or follows a list entry of type *Pause*, the search range is widened to include the entire range of delays present in the correspondence list. This is because some speech transmission systems use speech pauses to compensate accumulated delays. If the found entry is of type *Pause*, the search range is set to cover the entire length of the speech pause, i.e., the beginning of a pause in the degraded signal may be matched at most with the end of the corresponding pause in the reference signal and vice versa.

Since the correspondence list only contains information on how to match reference speech parts to the degraded signal, there may be inserted speech parts in the degraded signal for which no delay information is available in the list.

If the position of the current macro frame in the degraded signal is not covered by any entry in the correspondence list, then the delay for the current frame is given by the average delay of the neighbouring entries in the list before and after the current frame. Given that the neighbouring degraded signal parts before and after such an insertion are assigned to adjacent reference signal parts, an inserted part in the degraded signal has no real corresponding part in the reference. Thus the search range must be kept as small as possible. In the ITU-T P.863 algorithm, the search range is set to ±half a macro frame size around the junction of the neighbouring list entries in the reference signal. Additionally, the reliability for the inserted frame is set to 0.

9.2.5.3 Exclusion criteria for the fast prealignment method

The described method provides a computationally efficient and reliable delay computation for signals with fixed or piecewise fixed delays. For signals with highly variable delay or in case of resampling, however, trying to match speech segments by cross-correlation may not be accurate enough. Two criteria are used to verify the applicability of the fast prealignment:

- *Overall match quality:* This metric is calculated at the end of the segment-wise matching loop. It is defined as

$$\frac{totMatchedLen}{totSpeechLen} \tag{9-7}$$

  where totSpeechLen is the total length of non-Pause list entries, and totMatchedLen is the total length of Assigned list entries. List entries with type *Unsure* are also counted towards totMatchedLen if they are directly adjacent to Assigned entries and their delay does not differ by more than 5 ms.
- *Delay drift*: This additional metric detects degraded signals that were resampled with constant pitch by checking for near-constant delay increases or decreases from one macro frame to the next over the entire length of the degraded signal.

In case of detected resampling with constant pitch, or if the overall match quality is inferior to 75%, the results of the fast prealignment are discarded and the regular prealignment method is used as described in clauses 9.2.4.3 to 9.2.4.6. The fast prealignment is also skipped if the temporal alignment is entered a second time due to detected resampling. This is because the basis of the matching process, the cross-correlation of signal segments, may be impaired by even a slight imprecision of the resampling factor estimation.

9.2.6 Coarse alignment

The coarse alignment performs a stepwise refinement of the delay per frame. This is implemented by subdividing each signal into small subsections ("feature frames") and by calculating one characteristic value ("feature") for each of those subsections. The resulting vectors are called feature vectors. Feature frames are again equidistant and their length is reduced from iteration to iteration. Their length is independent from the macro frame length and typically shorter than the length of a macro frame. The iterative length reduction increases the accuracy of the estimated delay with each iteration, but at the same time, the search range, which can be used reliably, is reduced. Multiple feature vectors are calculated and for each macro frame, the feature, which is best suitable, is used to determine the final delay value for the current frame.

The result of the coarse alignment is a vector with the delay per macro frame, expressed in samples, with an accuracy, which is dependent of the feature frame length used in the final iteration.

In detail, the coarse alignment works as follows:

Starting with the lowest resolution (i.e., the longest feature frame length), all feature vectors are calculated for the active sections of both, the reference and the degraded signal. The features used are the energy per frame and the fractal dimension per frame. Now, the so-called correlation matrix is computed for each feature. This matrix is organized in correlation vectors per macro frame. The

correlation vectors contain for each macro frame the correlation between the reference and the degraded feature vector for all possible time lags between SearchRangeLow and SearchRangeHigh around the best delay per frame computed so far. The resolution of the correlation vectors is identical to the resolution of the feature vectors. This means, that with each iteration, the resolution of the correlation vectors is increased. The resulting matrix is of the format N_{cv} x N_{mf}, with N_{cv} being the number of possible lags tested in each correlation vector and N_{mf} being the number of macro frames. Next, the correlation matrices for all features are combined by selecting for each macro frame the correlation vector from the feature, which yields the maximum correlation for this frame. The position of the maximum correlation in the correlation vectors is equivalent to the optimum correction of the delay per frame required to achieve a better match between the two signals. Figure 6 illustrates the relation between macro frames, feature frames, and the correlation vectors.

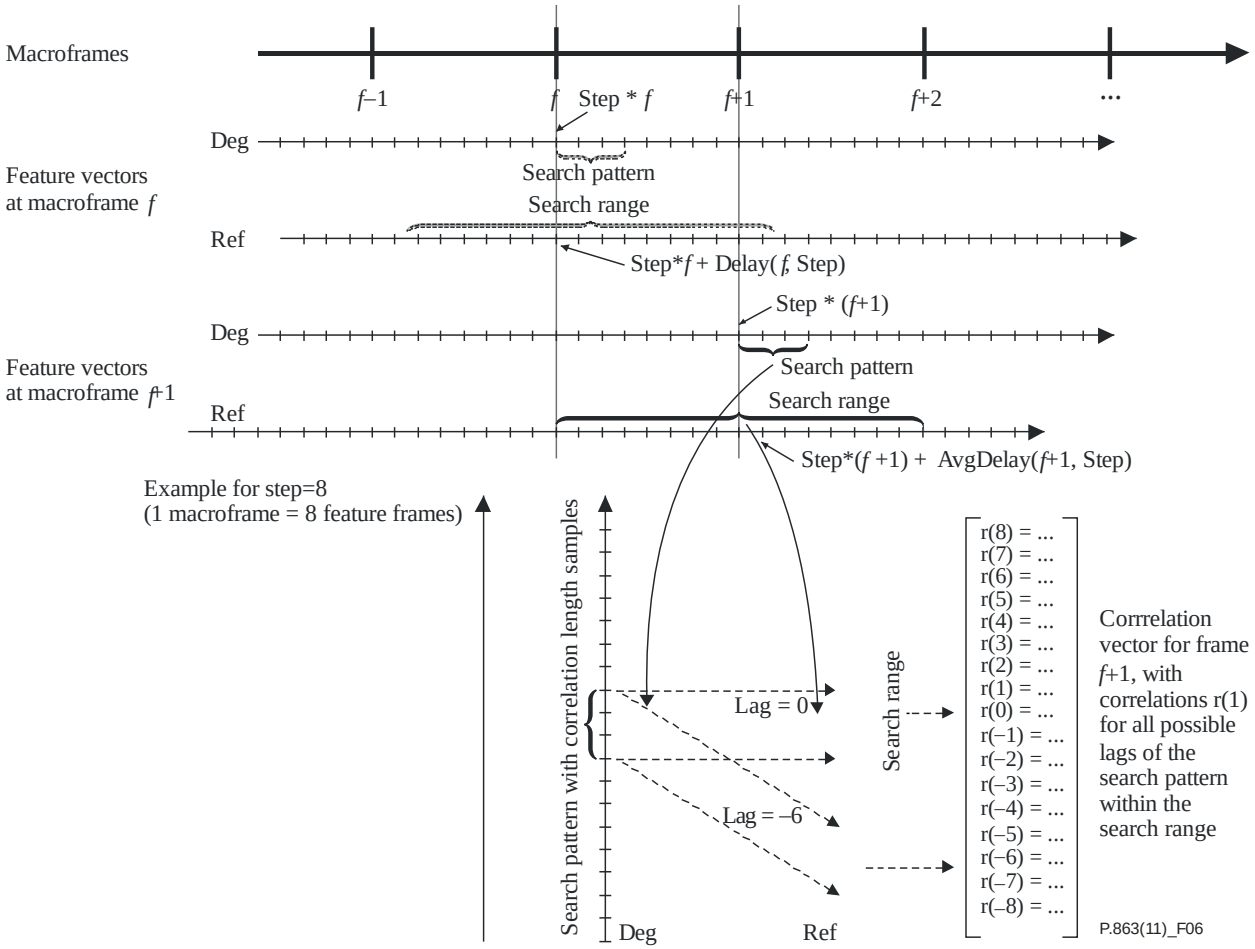


Figure 6 – Relation between macro frames, feature frames and the correlation vectors

However, simply searching the position of the maximum correlation would often lead to wild delay variations, since speech signals are often periodic and in some cases, e.g., when packet loss occurs, a wrong delay would lead to a better correlation than the correct delay. Therefore, a backtracking algorithm similar to Viterbi's algorithm is used to find the best possible path through the resulting combined correlation matrix. This algorithm starts with the last macro frame and traces the ideal path back to the first macro frame. For each frame, the correlations for all lags are weighted with a penalty factor, which depends on the history of the backtracking. This penalty factor is used to penalize larger delay variations.

The coarse alignment is the step requiring the highest computing power of the entire temporal alignment. Therefore, an efficient implementation is essential. To speed up calculations, the following optimizations are used:

a) The correlation vectors are calculated for active frames only.

b) The number of correlations which have to be calculated is limited to [SearchRangeLow:SearchRangeHigh] elements. All other elements in the correlation vector are set to 0. The values of SearchRangeLow and SearchRangeHigh are determined by the prealignment.

c) If the last delay estimate is almost constant and the maximum correlation is very high, then the correlation matrix is calculated for one feature only and the search range of frame *f*+1 is limited to a few elements around the position of the maximum found for the previous frame *f*. Only if this leads to a low correlation maximum for frame *f*, are the remaining elements of the correlation matrix recalculated.

d) If for any frame the optimum delay estimate is very close to the maximum or minimum of all allowed lags, then the feature vectors for this frame are shifted in a way, that the detected delay is centred in the correlation vector, and the correlation matrix is recalculated. While this increases the processing time in the rare cases when it really happens, it saves a huge amount of calculations for most cases, where the search range can remain rather small.

To better deal with some types of distortions and with the properties of some specific speech signals, the following artifices are used before the backtracking algorithm is applied:

- If a new active section is encountered, the first correlation vector belonging to that section is used for the second half duration of the preceding inactive part as well.
- The first correlation vector of the first active section is also used for all preceding frames; this avoids delay changes before the first active frame.
- The further away a macro frame is from the last active section, the lower will be the penalty factor used by the backtracking algorithm; this allows for quick delay adaptations after longer speech pauses.
- If the maximum correlation in a correlation vector is very high, then the penalty factor used by the backtracking algorithm will be reduced significantly.
- During the first ten frames of the speech signals, no penalty factor is used in the backtracking algorithm; this allows for almost instant adaptations of the delay.
- After each iteration of the coarse alignment the resulting, improved delay vector is cleaned up by:
    - removing sporadic delay changes larger than 2 ms, which are reverted to the original delay within 300 ms of the first delay variation,
    - eliminating delay changes larger than 50 ms, which are again modified within a duration that is smaller than the delay variation itself,
    - using the delay of the last valid frame for the first 50% of an invalid section, and the delay of the first valid frame after this invalid section for the second half of the invalid section for sections which are determined as invalid (e.g., inactive frames).

The corresponding parts of the algorithm are found in:

CTempAlignment::CoarseAlignment(…), in file: TempAlignment.pdf (see Annex B).

9.2.6.1 Fractal dimension

The fractal dimension of a signal can be seen as a measure for the signal's complexity. Very noisy signals will show a high fractal dimension (FD) per frame, while a sine tone will result in a very low *FD* value per frame.

In the ITU-T P.863 algorithm, Sevcik's formula is used to calculate the FD_f of each feature frame f:

$$Dist_i = (Sample_i - Sample_{i+1})^2 \tag{9-8}$$

$$L_f = \sum_{i=0}^{N} \sqrt{Dist_i + \left(\frac{1}{n-1}\right)^2} \tag{9-9}$$

$$FD_f = 1 + \frac{\ln(L_f) + \ln(2)}{\ln(2*(N-1))} \tag{9-10}$$

where N is the number of samples in the frame f. The final feature vectors, which are based on the fractal dimension, are DC filtered in order to avoid problems with the subsequent computations [b-Goh].

The corresponding parts of the algorithm can be found in:

CFDFeature::Sevcik(…), in file: FDFeatureModule.pdf (see Annex B).

9.2.6.2 Backtracking algorithm

The backtracking algorithm, which is used to determine the optimal path through the correlation matrix, is very similar to Viterbi's algorithm. In the ITU-T P.863 algorithm, it is assumed that the correlation in each element $R_{m,f}$ of the correlation vector is similar to the probability of a delay-offset f of the macro frame m. All elements of the correlation matrix are first converted to a value, which can be interpreted as the logarithmic probability:

$$p_{m,f} = -\log_{10}(1 - R_{m,f}) \tag{9-11}$$

The challenge is now to find the optimal path through that matrix, which yields the highest overall probability, without having to calculate all possible combinations. To do so, the algorithm starts with the probability vector of the last macro frame m and searches the index of the element with the highest probability, $p_{m,f}$, giving a first path probability pp_m. Next, a penalty is added to all elements of the probability vector P_{m-1}. This penalty is weakest for delay offsets, which would result in the same absolute delay as that for macro frame m and it is strongest for large delay variations. This penalty reduces the likelihood of larger delay variations. Now, the element from P_{m-1} is chosen that results in the highest combined probability pp_{m-1}:

$$pp_{m-1} = pp_m + p_{m-1,f} + Penalty(f) \tag{9-12}$$

For each step, the index f of the chosen optimum is stored. This index is equivalent to the offset of the best delay at the current feature resolution which has to be added to the last optimal delay value for each frame [b-Barkowsky].

The corresponding parts of the algorithm can be found in:

Viterbi (…), in file: TimeAlign.pdf (see Annex B).

9.2.7 Fine alignment

The fine alignment operates directly on both, the reference and the degraded signal at the maximum possible resolution and it determines the exact delay of each frame, expressed in samples. The required search range is drastically limited due to the previous alignment steps. Therefore, it is possible to predict the accurate delay values using very short correlations without compromising the accuracy of the prediction.

The result of the fine alignment is the sample accurate delay value of each macro frame.

In detail, the fine alignment works very similar to the coarse alignment. The main difference is that no iterations are used and that it operates directly on the squared value of the speech signal rather

than on features derived from them: In the first step a correlation matrix is calculated, which contains one correlation vector per macro frame. These correlation vectors contain the correlation between the reference and the degraded signal for all possible delay values within the search range.

NOTE – To avoid problems with 180 degree phase shifts, the correlation is calculated on the squared sample values.

The search range is determined by the frame size used by the last iteration of the coarse alignment. In principle, it would be enough to limit the search range to plus and minus the last feature frame size used, but to be on the safe side, this range is increased by a factor of two. The same backtracking algorithm as in the coarse alignment is applied to find the optimum path through this correlation matrix.

After applying the backtracking algorithm, the resulting delay vector is cleaned up. During this cleanup, each frame is checked whether the delay of the previous or the subsequent frame yields a better correlation than the delay determined so far. If this is the case, the best delay value determined will be used. This is repeated three times for all frames.

The corresponding parts of the algorithm can be found in:

CSpeechDelaySearch::FineAlign(…), in file: SpeechTempAlign.pdf (see Annex B).

9.3 Joining sections with constant delay

In this step, all sections with identical delay are combined, which means one set of information (delay, reliability, start, stop, speech activity) is stored for the entire section.

In a second step, each section $n+1$ is combined with section n:

- if section $n+1$ contains active speech and if the delay for both sections differs by less than 0.3 ms, or
- if section $n+1$ consists of a speech pause and if the delay for both sections differs by less than 15 ms.

The resulting section information is passed on to the psychoacoustic model.

The corresponding parts of the algorithm can be found in:

CTempAlignment::CreateUtteranceVectorsDeg(…), in file: TempAlignment.pdf (see Annex B).

9.4 Sample rate ratio detection

The sample rate ratio detection is required to compensate for perceptually irrelevant differences in the playout speed of both, the reference and the degraded signal. Such differences may have various reasons and they may be intentional (e.g., time scaling due to jitter buffer adaptation) or not intentional (e.g., due to unsynchronized A/D or D/A converters in partly analogue equipment). The resulting effect in any case is the same and can be described as a difference in the sample rate of two signals in the range of very few percent. It is important to note that this is not about the nominal sample rate, rather than the effective sample rate relative to another signal.

The detection of this effect as implemented in the ITU-T P.863 algorithm is based on the delay per frame vector and the detected active sections of the speech signals, as determined by the temporal alignment. The theory behind the algorithm is that sample rate differences will lead to delay changes, which are proportional to the ratio of the effective sample rates. The ITU-T P.863 algorithm performs this by verifying if the delay variations during active speech follow a linear model. If this is the case, the slope of the linear function will indicate the ratio of the sampling rates.

The parameters of the linear function '$y = ax + b$' that best fits the delay function over time are computed by following the Least Squares Method (i.e., minimizing the sum of squares error) as:

$$a = \frac{M\sum xy - \sum x \sum y}{M\sum x^2 - (\sum x)^2} \tag{9-13}$$

$$b = \frac{\sum y - a\sum x}{M} \tag{9-14}$$

where y is the *DelayPerFrame[m]* (in samples), x the frame number m, M is the number of points used for this estimate (only the frames with active speech), a is the slope of the curve and b is the Y-intercept.

If 90% of the delay variations during active speech can be explained by this linear model (the delay values differ less than 32ms from its expected value in the linear function) we estimate the sample rate ratio *SRRatio* as:

$$SRRatio = 1 - \frac{a}{N} \tag{9-15}$$

where N is the number of samples in a frame.

In case this condition is not satisfied during the entire sample duration, the ITU-T P.863 algorithm separates the file into sentences and repeats the procedure. If all sentences have a valid estimate $SRRatio_i$ (90% of frames in the sentence follow the partial linear model) and all $SRRatio_i$ go in the same sense (all >1.0 or all <1.0), the selected value for *SRRatio* is the one closer to 1.0 (conservative approach for resampling). Otherwise a ratio of 1.0 is reported.

The effective sample rate can be calculated by multiplying *SRRatio* with the nominal sample rate.

The corresponding parts of the algorithm can be found in:

CTempAlignment::GetSampleRateRatio_linear(…), in file: TempAlignment.pdf (see Annex B).

9.5 Resampling

If the difference between the nominal sample rate and the detected sample rate is larger than 0.5%, the signal with the higher sample rate will be down sampled and the entire processing starts from the beginning. This happens at most one time to avoid excessive looping in case of signals for which the sample rate ratio cannot be determined in a reliable manner.

Even if the sample rate determination cannot be made with perfect accuracy, e.g., in case of signals with additional variable delay, the detected sample rate ratio is still accurate enough to bring the signals back to the safe operating range of the temporal alignment.

9.6 Level, frequency response and time alignment pre-processing

9.6.1 Determination of the overall system gain

The ITU-T P.863 algorithm is designed to take into account the impact of play back level on the perceived quality and consequently needs a calibration factor that maps the digital signal representation level in dBov towards the play back level in dB(A). This calibration factor is chosen to be 2.8 for a signal that is played at 73 dB(A) in diotic presentation (same signal to both ears) when the digital representation is at a level of –26 dBov. For each other combination of digital signal level and play back level, the calibration factor can be calculated from:

$$C = 2.8 * 10^{(-26 - dBov)/20} * 10^{(73 - dB(A))/20} \tag{9-16}$$

Because the ITU-T P.863 algorithm was validated on signals played at levels between 53 and 79 dB(A), the ITU-T P.863 algorithm allows one to assess the impact of playing the signal at other than optimal levels, especially the impact of playing the signal at a low level. Play back of the signal

at a low level can result in both a quality improvement and a quality degradation, depending on the type of distortion found in the degraded file.

9.6.2 IRS receive pre-filtering in ITU-T P.863

The ITU-T P.863 algorithm can operate in two modes, narrowband mode, and super-wideband mode. In the narrowband mode, both the reference and degraded signals are pre-filtered with an IRS receive filter representing a listening situation in which subjects judge the quality of the speech signals over an IRS receive handset in monotic mode or over an IRS receive headset in monotic mode. In the super-wideband mode, both the reference and degraded signals are not filtered, representing a listening situation in which subjects judge the quality of the speech signals over a diffuse field equalized headset in diotic mode.

9.7 Perceptual model

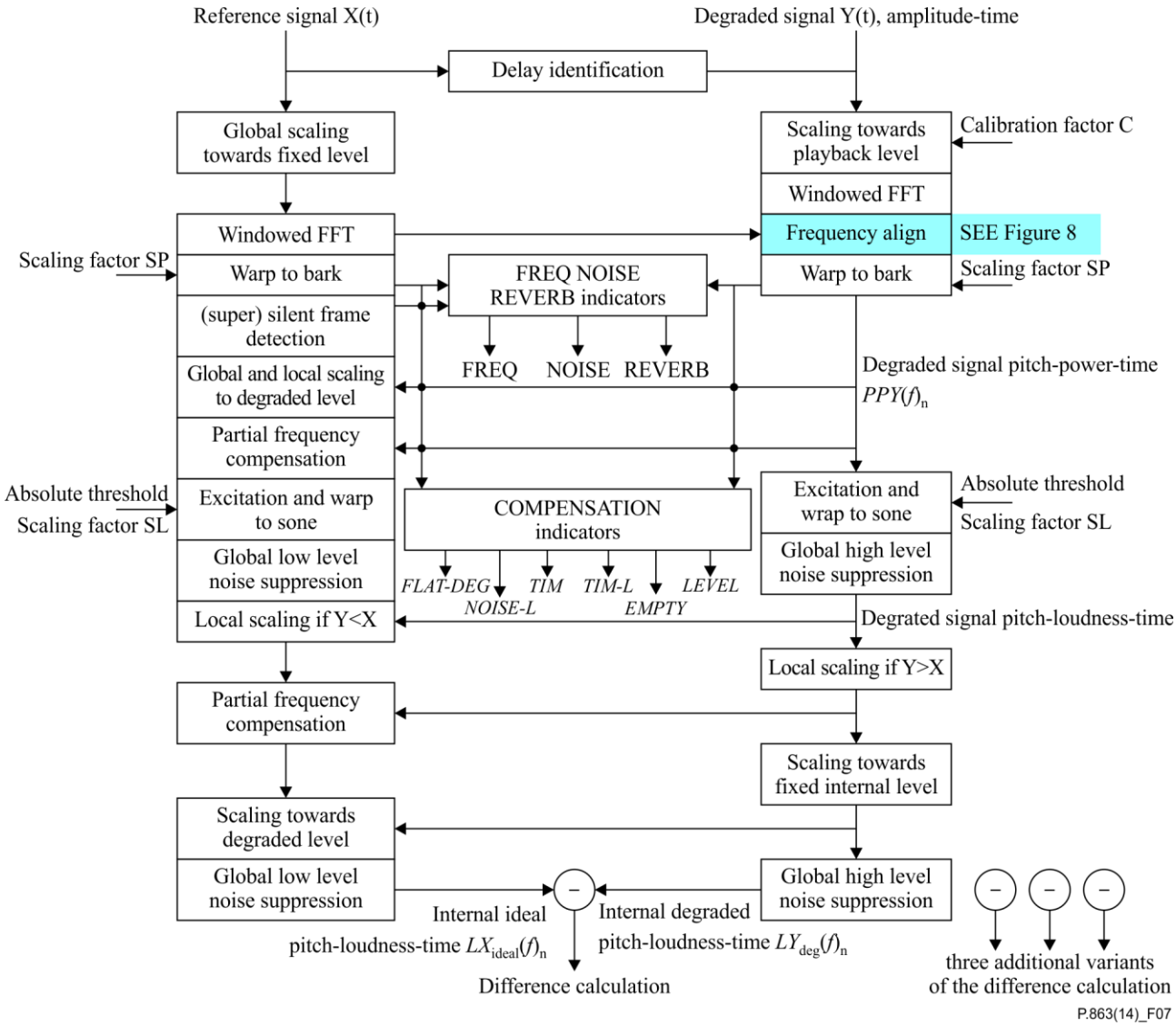
An overview of the perceptual model is given in Figures 7 through 10. The corresponding description can be found in: CPairParameters::DisturbanceProcess( ), in the file: Disturbance.pdf (see Annex B).

Figure 7 provides the basics of the perceptual model used in the calculation of the internal representation. For files where the frequency axis is warped, a special frequency domain align part is used as given in Figure 8. The pitch power densities (power as function of time and frequency) of reference and degraded are derived from the time and frequency aligned time signals. These densities are then used to derive the first three ITU-T P.863 quality indicators for frequency response distortions (FREQ), additive noise (NOISE) and room reverberations (REVERB). Furthermore six different compensation indicators are calculated that are used as a compensation for small errors in the final MOS prediction, a level indicator (LEVEL) a noise indicator for very loud noise levels (NOISE-L), three timbre indicators for quantifying unbalanced timbres corrections in the degraded signal independent from the timbre of the reference signal (TIM, TIM-L and FLAT-DEG) and an indicator for quantifying the effect of noise in frequency bands where his only little speech power (e.g., when wideband noise is added to narrowband speech, EMPTY). From the pitch power densities the internal representations of reference and degraded are derived in a number of steps as described in clauses 9.7.3 through 9.7.11. Four different variants of these densities are calculated, one representing the main branch (Disturbance.pdf p. 22), one representing the main branch for big distortions (Disturbance.pdf p. 44), one branch focused on added distortions (Disturbance.pdf p. 35) and one focused on added big distortions (Disturbance.pdf p. 51).

The four different variants of calculating the pitch power densities of reference and degraded signals are the inputs to the calculation of the final disturbance densities as given in Figure 9. The variants for the reference densities are referred to as ideal densities because low levels of noise in the reference are removed and timbre distortions are partially compensated for. Two final disturbance densities are calculated, one representing the final disturbance as a function of time and frequency, and one representing the final disturbance as a function of time and frequency but focused on the processing of added disturbances (clause 9.7.12).

Figure 10 gives an overview of the calculation of the MOS-LQO score from the two final disturbance densities and the FREQ, NOISE, REVERB indicators (clauses 9.7.13 and 9.7.14).

Details of the several steps in the processing can be found in the descriptive pdf files in the electronic attachment to this Recommendation (see Annex B).



NOTE – Four different variants of the internal representations are calculated each focused on a specific set of distortions (see Figure 9).

Figure 7 – Overview of the first part of the ITU-T P.863 perceptual model: Calculation of the internal representation of the reference and degraded

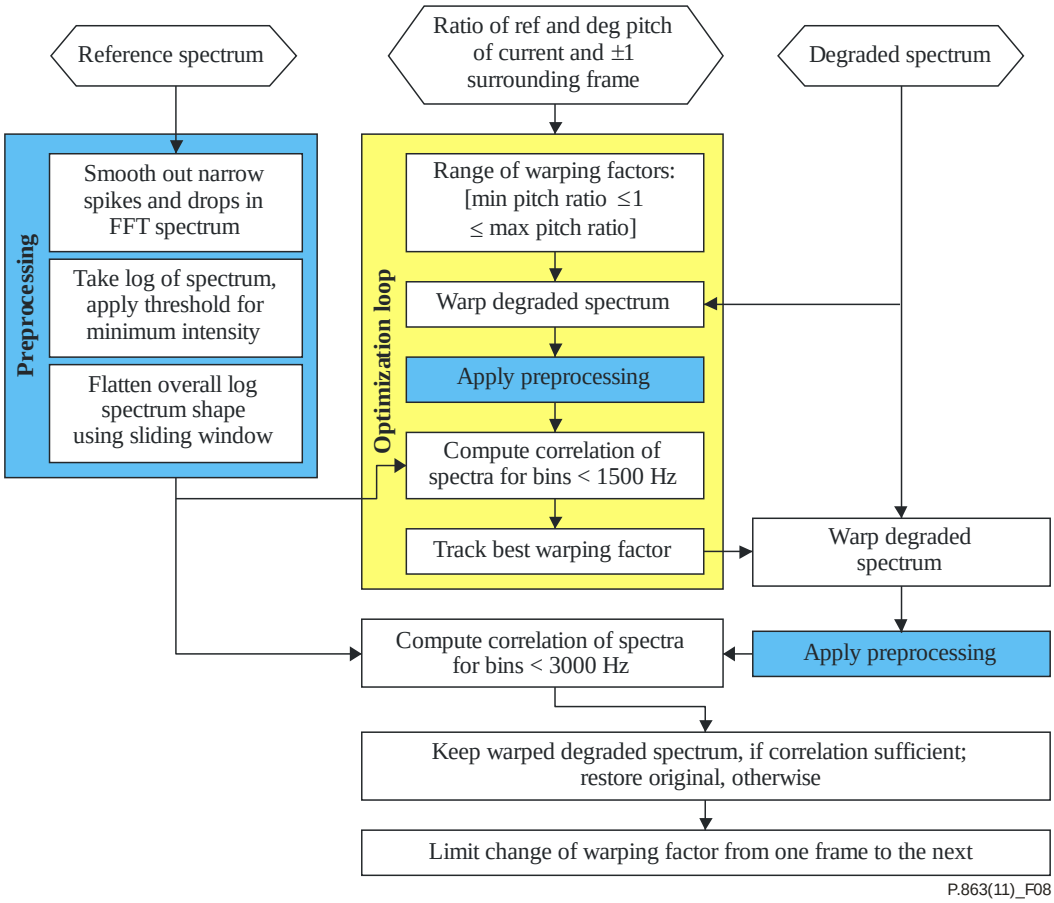
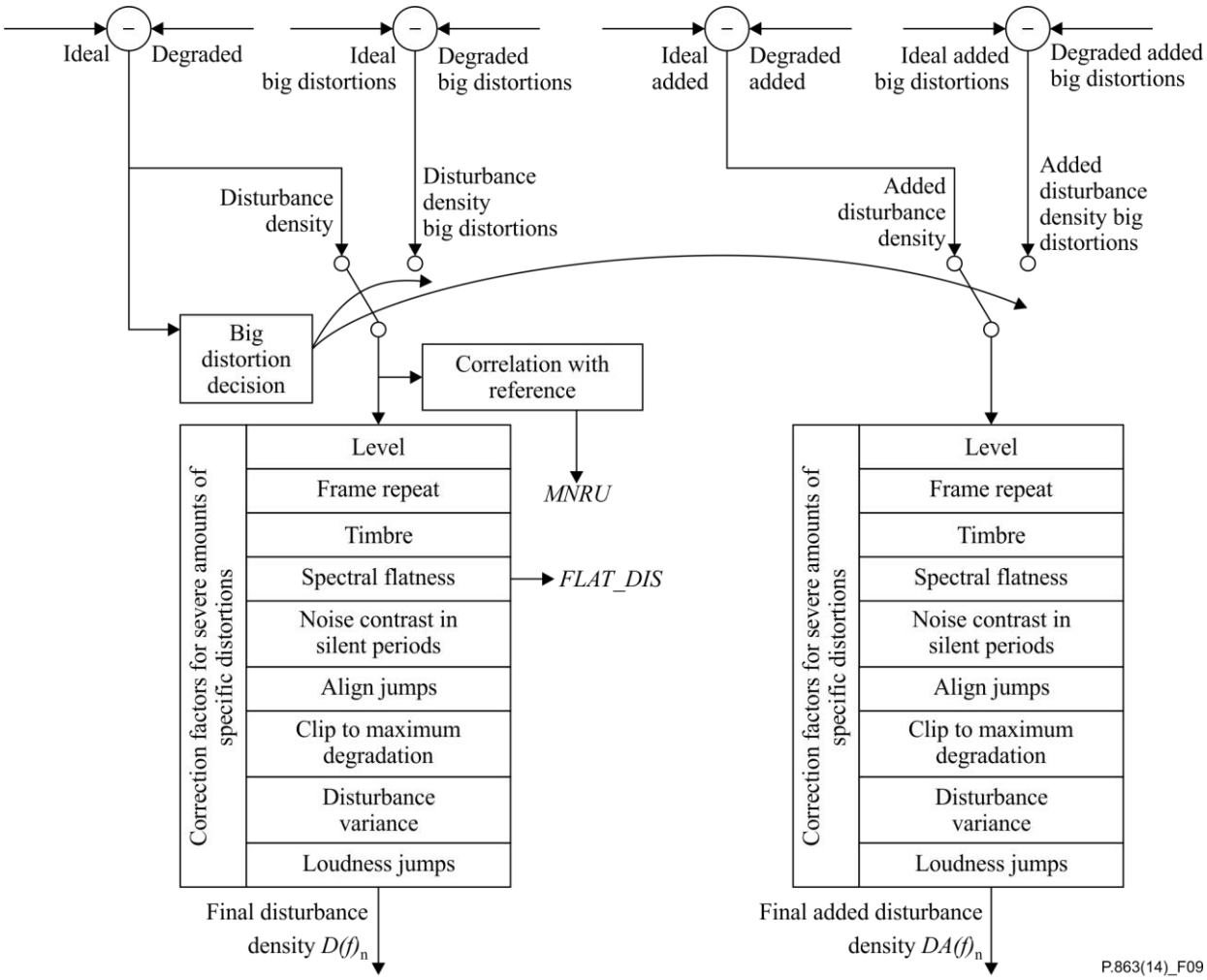


Figure 8 – Overview of the frequency domain alignment used in the ITU-T P.863 perceptual model



NOTE – See clause 9.7.12.

Figure 9 – Overview of the second part of the ITU-T P.863 perceptual model: Calculation of final disturbance densities from the four different variants of the internal representations

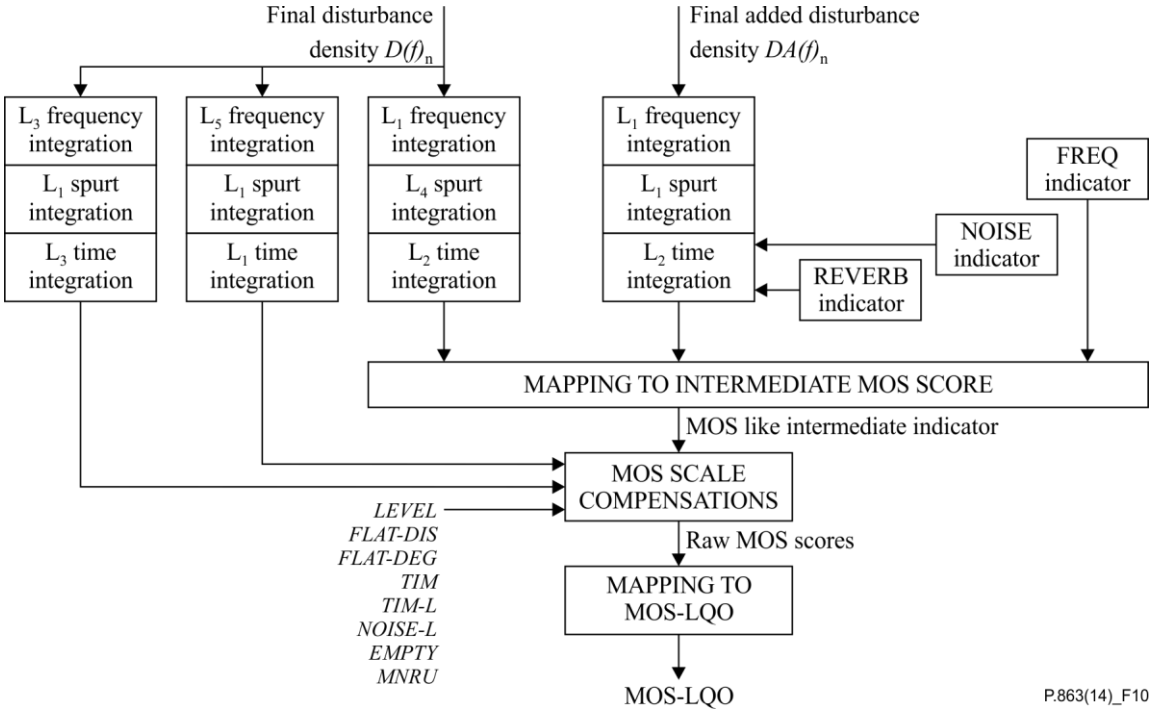


Figure 10 – Overview of the third part of the ITU-T P.863 perceptual model: Calculation of the final objective listening quality MOS score (MOS-LQO) from the final disturbance densities

9.7.1 Pre-computation of constant settings

9.7.1.1 FFT window size depending on the sample frequency

The window size W is set dependent on the sampling frequency f_s:

$$\begin{aligned} 0 < f_s \le 9\,\text{kHz} &\rightarrow W = 256 \\ 9 < f_s \le 18\,\text{kHz} &\rightarrow W = 512 \\ 18 < f_s \le 36\,\text{kHz} &\rightarrow W = 1024 \\ 36 < f_s \le 72\,\text{kHz} &\rightarrow W = 2048 \end{aligned} \tag{9-17}$$

The ITU-T P.863 algorithm was tested on 8, 16, and 48 kHz sampling. Resampling will not reproduce exactly the same MOS score as would occur in a subjective test, especially if the resampling deviates significantly from a factor of 2.

The overlap between successive frames is 50% using a Hanning window. The power spectra – the sum of the squared real and squared imaginary parts of the complex FFT components – are stored in separate real-valued arrays for both the reference and the degraded signal. Phase information within a single frame is discarded in the ITU-T P.863 algorithm and all calculations are based on the power representations only.

9.7.1.2 Start stop point calculation

In subjective tests, noise will usually start before the beginning of the speech activity in the reference signal. However, leading steady state noise in a subjective test decreases the impact of steady state noise, while in objective measurements that take into account leading noise it will increase the impact; therefore, it is expected that omission of leading and trailing noises is the correct perceptual approach. Therefore, the start and stop points used in the ITU-T P.863 processing are calculated from the beginning and end of the reference file. The sum of five successive absolute sample values must exceed 500 from the beginning and end of the original speech file in order for that position to be

designated as the start or end. The interval between this start and end is defined as the active processing interval. Distortions outside this interval are ignored in the ITU-T P.863 processing.

9.7.1.3 The power and loudness scaling factor

For calibration of the FFT time to frequency transformation, a sine wave with a frequency of 1 000 Hz and an amplitude of 40 dB SPL is generated. This sine wave is transformed to the frequency domain using a windowed FFT with a length determined by the sampling frequency. After converting the frequency axis to the Bark scale, the peak amplitude of the resulting pitch power density is then normalized to a power value of 10^4 by multiplication with a power scaling factor *SP*.

The same 40 dB SPL reference tone is used to calibrate the psychoacoustic (Sone) loudness scale. After warping the intensity axis to a loudness scale using Zwicker's law [b-Zwicker], the integral of the loudness density over the Bark frequency scale is normalized to 1 Sone using the loudness scaling factor *SL*.

9.7.2 Calculation of the pitch power densities

The degraded signal Y(t) is multiplied by the calibration factor C (see clause 9.6.1) and then transformed to the time-frequency domain with 50% overlapping FFT frames. The reference signal is scaled towards a fixed optimum level.

For files where the frequency axis is warped when compared to the reference, a dewarping in the frequency domain is carried out on the FFT frames. In the first step of this dewarping, both the reference and degraded FFT power spectra are preprocessed to reduce the influence of both very narrow frequency response distortions, as well as overall spectral shape differences on the following calculations. The preprocessing consists of performing a sliding window average of length equivalent to 100 Hz over both power spectra, taking the logarithm, and performing a sliding window normalization as described in clause 9.2.5.1, using a window length equivalent to 218.75 Hz. The ratio of the reference to degraded pitch of the current frame is then computed (see chapter 4 of [b-Beerends 1989]) and used to determine the search range for the warping factor, which lies between 1 and said pitch ratio. If possible, this search range is extended by the minimum and maximum pitch ratio found for one preceding and one following frame pair.

The function then iterates through the search range and warps the degraded power spectrum with the warping factor of the current iteration, and processes the warped power spectrum as described above. The correlation of the processed reference and processed warped degraded spectrum is then computed for bins between the common lower frequency limit and 1 500 Hz. After complete iteration through the search range, the "best" (i.e., that resulted in the highest correlation) warping factor is retrieved. The correlation of the processed reference and best warped degraded spectra is then compared against the correlation of the original processed reference and degraded spectra. The "best" warping factor is then kept if the correlation increases by a set threshold. If necessary, the warping factor is limited by a maximum relative change to the warping factor determined for the previous frame pair.

After the dewarping the frequency scale in Hz is warped towards the pitch scale in Bark reflecting that at low frequencies, the human hearing system has a finer frequency resolution than at high frequencies. This is implemented by binning FFT bands and summing the corresponding powers of the FFT bands with a normalization of the summed parts. The warping function that maps the frequency scale in Hertz to the pitch scale in Bark approximates the values given in the literature. The resulting reference and degraded signals are known as the pitch power densities $PPX(f)_n$ and $PPY(f)_n$ with *f* the frequency in Bark and the index n representing the frame index.

9.7.3 Computation of the speech active, silent and super silent frames

ITU-T P.863 operates on three classes of frames:

- speech active frames where the frame level of the reference signal is above a level that is about 20 dB below the average,

- silent frames where the frame level of the reference signal is below a level that is about 20 dB below the average, and
- super silent frames where the frame level of the reference signal is below a level that is about 35 dB below the average level.

9.7.4 Calculation of the frequency, noise and reverb indicators

The global impact of frequency response distortions, noise and room reverberations is separately quantified.

For the impact of overall global frequency response distortions, an indicator is calculated from the average spectra of reference and degraded signals. In order to make the estimate of the impact for frequency response distortions independent of additive noise, the average noise spectrum density of the degraded over the silent frames of the reference signal is subtracted from the pitch loudness density of the degraded signal. The resulting pitch loudness density of the degraded and the pitch loudness density of the reference are then averaged in each Bark band over all speech active frames for the reference and degraded file. The difference in pitch loudness density between these two densities is then integrated over the pitch to derive an average frequency response difference indicator. This indicator is then combined with the rate of change over consecutive Bark pitch bands to obtain the indicator for quantifying the impact of frequency response distortions (FREQ).

For the impact of additive noise, an indicator is calculated from the average spectrum of the degraded signal over the silent frames of the reference signal. The difference between the average pitch loudness density of the degraded over the silent frames and a zero reference pitch loudness density (representing perfect silence) determines a noise loudness density function that quantifies the impact of additive noise. This noise loudness density function is then integrated over the pitch to derive an average noise impact indicator (NOISE).

For the impact of room reverberations, the energy over time function (ETC) is calculated from the reference and degraded time series. The ETC represents the envelope of the impulse response. In a first step the loudest reflection is calculated by simply determining the maximum value of the ETC curve after the direct sound. In the ITU-T P.863 model, direct sound is defined as all sounds that arrive within 60 ms. Next, a second loudest reflection is determined over the interval without the direct sound and without taking into account reflections that arrive within 100 ms from the loudest reflection. Then, the third loudest reflection is determined over the interval without the direct sound and without taking into account reflections that arrive within 100 ms from the loudest and second loudest reflection. The energy of the three loudest reflections are then combined into a single reverb indicator (REVERB).

9.7.5 Scaling of the reference

The reference signal is now at the ideal level while the degraded signal is represented at a level that coincides with the play back level. Before a comparison is made between the reference and degraded, the overall level and small changes in local level are compensated to the extent that is necessary for the quality calculation. The global level equalization is carried out on the basis of the average power of reference and degraded in the speech band between 400 and 3 500 Hz. The reference signal is scaled towards the degraded signal and the impact of the level difference is thus compensated. For correctly modelling slowly varying gain distortions, a local scaling is carried out for level changes up to about 3 dB.

9.7.6 Partial compensation of the original pitch power density for linear frequency response distortions

To deal with filtering in the system under test, which introduces non-audible linear frequency response distortions, the reference signal is partially filtered with the transfer characteristics of the system under test. This is carried out by calculating the average power spectrum of the original and

degraded pitch power densities over all speech active frames. Per Bark bin, a partial compensation factor is calculated from the ratio of the degraded spectrum to the original spectrum.

9.7.7 Modelling of masking effects, calculation of the pitch loudness densities

Masking is modelled by calculating a smeared representation of the pitch power densities. Both time and frequency domain smearing are taken into account (see Figure 11). The time-frequency domain smearing uses a convolution approach as given in [b-Beerends 1994]. From this smeared representation, the representations of the reference and degraded pitch power density are re-calculated suppressing low amplitude time-frequency components, which are partially masked by loud components in the neighbourhood in the time-frequency plane. This suppression is implemented in two different manners, a subtraction of the smeared representation from the non-smeared representation and a division of the non-smeared representation by the smeared representation. The resulting representations of the pitch power density are then transformed to pitch loudness density representations using a modified version of Zwicker's power law [b-Zwicker]:

$$LX(f)_n = SL * \left(\frac{P_0(f)}{0.5}\right)^{0.22 * f_B * P_{fn}} * \left[\left(0.5 + 0.5\frac{PPX(f)_n}{P_0(f)}\right)^{0.22 * f_B * P_{fn}} - 1\right] \qquad (9\text{-}18)$$

with SL the loudness scaling factor, $P_0(f)$ the absolute hearing threshold, f_B and P_{fn} a frequency and level dependant correction defined by:

$$f_B = -0.03 * f + 1.06 \qquad \text{for } f < 2.0 \text{ Bark}$$

$$f_B = 1.0 \qquad \text{for } 2.0 \le f \le 22 \text{ Bark}$$

$$f_B = -0.2 * (f - 22.0) + 1.0 \qquad \text{for } f > 22 \text{ Bark}$$

$$P_{fn} = (PPX(f)_n + 600)^{0.008}$$

with f representing the frequency in Bark, $PPX(f)_n$ the pitch power density in frequency time cell f, n. The resulting two dimensional arrays $LX(f)_n$ and $LY(f)_n$ are called pitch loudness densities.

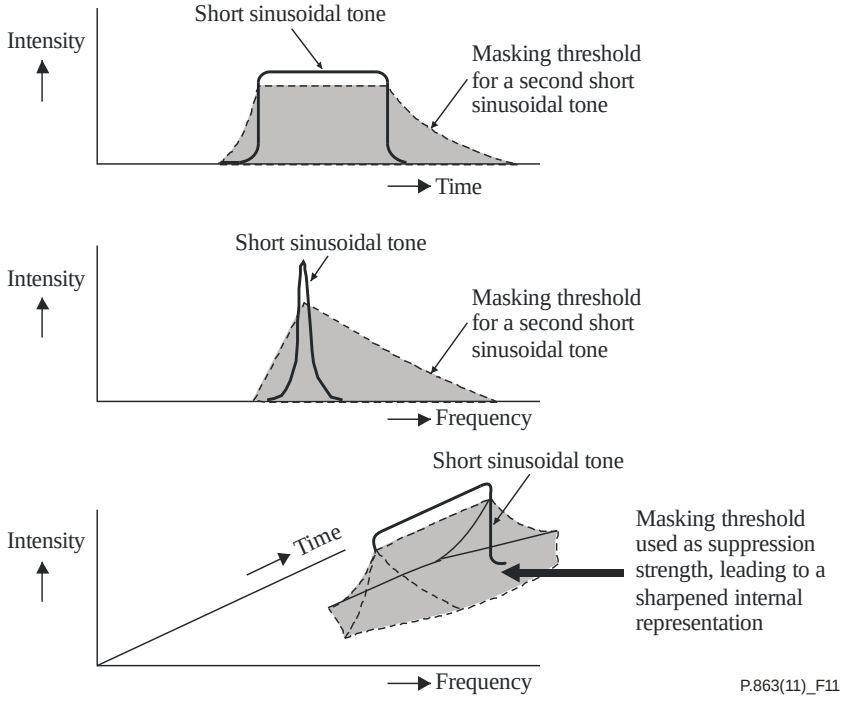


Figure 11 – Masking approach used in the ITU-T P.863 perceptual model

9.7.8 Compensation of the noise in reference and degraded signals

Low levels of noise in the reference signal, which are not affected by the system under test (e.g., a transparent system), will be attributed to the system under test by subjects and thus have to be suppressed in the calculation. This is carried out by calculating the average steady state noise loudness density of the reference signal $LX(f)_n$ over the super silent frames as a function of pitch. This average noise loudness density is then partially subtracted from all pitch loudness density frames of the reference signal. The result is an idealized internal representation of the reference signal.

Steady state noise that is audible in the degraded signal has a lower impact than non-steady state noise. This holds for all levels of noise and the impact of this effect can be modelled by partially removing steady state noise from the degraded signal. This is carried out by calculating the average steady state noise loudness density of the degraded signal $LY(f)_n$ frames for which the corresponding frame of the reference signal is classified as super silent, as a function of pitch. This average noise loudness density is then partially subtracted from all pitch loudness density frames of the degraded signal. The partial compensation uses a different strategy for low and high levels of noise. For low levels of noise the compensation is only marginal while the suppression that is used becomes more aggressive for loud additive noise. The result is an internal representation of the degraded signal with an additive noise that is adapted to the subjective impact as observed in listening test.

9.7.9 Partial compensation of the distorted pitch loudness density for time-varying gain between degraded and reference signal

Slow variations in gain are inaudible and small changes are already compensated for in the calculation of the reference signal representation (see clause 9.7.5). The remaining compensation necessary before the correct internal representation can be calculated is carried out in two steps; first the reference is compensated for signal levels where the degraded signal loudness is less than the reference signal loudness and second the degraded is compensated for signal levels where the reference signal loudness is less than the degraded signal loudness. The first compensation scales the reference signal towards a lower level for parts of the signal where the degraded shows a severe loss of signal such as in time clipping situations. The scaling is such that the resulting difference between reference and degraded represents the impact of time clips on the local perceived speech quality. Parts where the reference signal loudness is less than the degraded signal loudness are not compensated and thus additive noise and loud clicks are not compensated in this first step.

The second compensation scales the degraded signal towards a lower level for parts of the signal where the degraded signal shows clicks and for parts of the signal where there is noise in the silent intervals. The scaling is such that the resulting difference between reference and degraded represents the impact of clicks and slowly changing additive noise on the local perceived speech quality. The clicks are compensated in both the silent and speech active parts, the noise is compensated only in the silent parts.

9.7.10 Partial compensation of the original pitch loudness density for linear frequency response distortions

To deal with filtering in a system under test that introduces non-audible linear frequency response distortions, the reference signal was already partially filtered in the pitch power density domain. In order to further correct for the fact that linear distortions are less objectionable than non-linear distortions, the reference signal is now partially filtered in the pitch loudness domain. This is carried out by calculating the average loudness spectrum of the original and degraded pitch loudness densities over all speech active frames. Per Bark bin, a partial compensation factor is calculated from the ratio of the degraded loudness spectrum to the original loudness spectrum. This partial compensation factor is used to filter the reference signal with a smoothed, lower amplitude, version of the frequency response of the system under test. After this filtering, the difference between the reference and degraded pitch loudness densities that result from linear frequency response distortions is diminished

to a level that represents the impact of linear frequency response distortions on perceived speech quality.

9.7.11 Final scaling and noise suppression of the pitch loudness densities

Up to this point, all calculations on the signals are carried out on the play back level as used in the subjective experiment. For low play back levels, this will result in a low difference between reference and degraded pitch loudness densities and in general in a far too optimistic estimation of the listening speech quality. In order to compensate for this effect, the degraded signal is now first scaled in the low and upper frequency bands and then scaled towards a fixed global play back loudness. After this scaling, the reference signal is scaled towards the degraded signal and both the reference and degraded signal are now ready for a final noise suppression operation. This noise suppression takes care of the last parts of the steady state noise levels in the loudness domain that still have a too big impact on the speech quality calculation. The resulting signals are now in the perceptual relevant internal representation domain and from the ideal pitch-loudness-time $LX_{ideal}(f)_n$ and degraded pitch-loudness-time $LY_{deg}(f)_n$ functions the disturbance densities can be calculated. Four different variants of these densities are calculated, one representing the main branch, one representing the main branch for big distortions, one branch focused on added distortions and one focused on added big distortions.

9.7.12 Calculation of the final disturbance densities

Two final disturbance densities are calculated. The first one is derived from the difference between the ideal pitch-loudness-time and degraded pitch-loudness-time function. The second one is derived from the ideal pitch-loudness-time and a degraded pitch-loudness-time functions optimized with regard to parts where the degraded power density is larger than the reference power density (called ideal added pitch-loudness-time and degraded added pitch-loudness-time, see Figure 9) and from which a disturbance is calculated that is weighted with a factor dependant on the power ratio in each pitch-time cell, the asymmetry factor. The resulting disturbance density is referred to as the added density. In order to be able to deal with a large range of distortions two flavours of the densities are calculated, one derived from the difference between $LX_{ideal}(f)_n$ and $LY_{deg}(f)_n$ calculated with a perceptual model focused on small to medium distortions and one derived from the difference between $LX_{ideal}(f)_n$ and $LY_{deg}(f)_n$ calculated with a perceptual model focused on medium to big distortions. The switching between the two is carried out on the basis of a first estimation from the disturbance focused on small to medium level of distortions.

In the next steps, the final disturbance and added disturbance densities are compensated for severe amounts of specific distortions.

Severe deviations of the optimal listening level are quantified by an indicator directly derived from the signal level of the degraded signal. This global indicator (LEVEL) is also used in the calculation of the MOS-LQO.

Severe distortions introduced by frame repeats are quantified by an indicator derived from a comparison of the correlation of consecutive frames of the reference signal with the correlation of consecutive frames of the degraded signal.

Severe deviations from the optimal timbre are quantified by an indicator derived from the ratio of the upper frequency band loudness and the lower frequency band loudness of the disturbance densities. Compensations are carried out per frame and on a global level. The global level of timbre deviation is quantified with the FLAT-DIS indicator which is also used in the calculation of the MOS-LQO.

Severe noise level variations which focus the attention of subjects towards the noise are quantified by a noise contrast indicator derived from the silent parts of the reference signal.

Finally, the disturbance and added disturbance densities are clipped to a maximum level and the variance of the disturbance and the jumps in the loudness are used to compensate for specific time structures of the disturbances.

9.7.13 Aggregation of the disturbance over pitch, spurts and time, mapping to the intermediate MOS score

The final disturbance $D(f)_n$ and added disturbance $DA(f)_n$ densities are integrated per frame over the pitch axis resulting in two different disturbances per frame, one derived from the disturbance and one derived from the added disturbance, using an L_1 integration (see Figure 10):

$$D_n = \sum_{f=1,..Numberof\ Barkbands} |D(f)_n| W_f \tag{9-19}$$

$$DA_n = \sum_{f=1,..Numberof\ Barkbands} |DA(f)_n| W_f \tag{9-20}$$

with W_f a series of constants proportional to the width of the Bark bins.

Next, these two disturbances per frame are averaged over speech spurts of six consecutive frames with an L_4 and an L_1 weighting for the disturbance and for the added disturbance.

$$DS_n = \sqrt[4]{\frac{1}{6} \sum_{m=n,..n+6} D_m^4} \tag{9-21}$$

$$DAS_n = \frac{1}{6} \sum_{m=n,..n+6} D_m \tag{9-22}$$

Finally, a disturbance and an added disturbance are calculated per file from an L_2 averaging over time:

$$D = \sqrt[2]{\frac{1}{numberOfFrames} \sum_{n=1,..numberOfFrames} DS_n^2} \tag{9-23}$$

$$DA = \sqrt[2]{\frac{1}{numberOfFrames} \sum_{n=1,..numberOfFrames} DAS_n^2} \tag{9-24}$$

The added disturbance is compensated for loud reverberations and loud additive noise using the REVERB and NOISE indicators. The two disturbances are then combined with the frequency indicator (FREQ) to derive an internal indicator that is linearized with a third order regression polynomial to get a MOS-like intermediate indicator. In narrowband mode, this indicator is calculated in a simplified manner, the two side branches L_{313} and L_{511} operating on the final disturbance are replaced by a single L_{584} branch operating on the final added disturbance. The main branch indicators L_{142} and L_{223} are replaced by L_{223} and L_{132} operating on respectively the final disturbance and added disturbance (see Figure 10).

9.7.14 Computation of the final MOS-LQO ITU-T P.863 score

The raw ITU-T P.863 score is derived from the MOS-like intermediate indicator using ten different compensation factors, three as defined in the 2011 version of P.863:

- Two compensation factors for specific time-frequency characteristics of the disturbance, one calculated with an L_{511} aggregation over frequency, spurts and time, and one calculated with an L_{313} aggregation over frequency, spurts and time (see Figure 10),
- One compensation factor for very low presentation levels using the LEVEL indicator of the degraded signal. The corresponding variable in the algorithm is globalScaleDistortedToFixedlevelHulp in the files Time.pdf and is derived from globalScaleDistortedToFixedlevel as calculated in Disturbance.pdf, see Annex B).

In order to have a better behaviour for transparent conditions and for conditions that have sinusoidal degradations the FLATNESS indicator as used in the 2011 version of P.863 is replaced by a separate

flatness indicator compensation for silent intervals calculated over the disturbances and for the silent intervals calculated over the degraded files. The spectral flatness compensation factors are calculated by dividing the geometric mean of the disturbance and degraded pitch power densities by the arithmetic mean of the disturbance (FLAT_DIS, corresponding variable is frameFlatnessDisturbanceAvgCompensation000silent in the files Time.pdf and Disturbance.pdf, see Annex B) and degraded pitch power densities (FLAT_DEG, corresponding variable is frameFlatnessDistortedAvgCompensationSilent000 in the files Time.pdf and Disturbance.pdf).

Furthermore two new MOS compensation factors are defined for modelling the impact of non-optimal timbre of the degraded signal:

- A compensation factor for the impact of the timbre of the active speech parts calculated on the basis of the power in the upper (above 6 Bark) and lower (below 11 Bark) frequency bands of the degraded file (TIM, corresponding variable in the algorithm is distortedLoudnessTimbreHighPerFrameAvgActive000 in the files Time.pdf and Disturbance.pdf, see Annex B). For each frame the loudness contributions of the lower spectral parts are subtracted from the loudness contributions of the higher spectral parts and the difference is averaged over all active speech parts as derived from the aligned reference signal.
- A compensation factor for the impact of the timbre of the loud active speech parts calculated on the basis of the power in the upper (above 10 Bark) and lower (below 10 Bark) frequency bands of the degraded file for frames that have a level above the average active speech level (TIM-L, corresponding variable in the algorithm is distortedLoudnessTimbrePerFrameLoudAvg000 in the files Time.pdf and Disturbance.pdf, see Annex B). For each frame the loudness contributions of the lower spectral parts are subtracted from the loudness contributions of the higher spectral parts and the difference is averaged over all loud active speech parts as derived from the aligned reference signal.

Note that these two indicators quantify the timbre differences between an ideal timbre voice recording and the degraded voice output. If the reference signal is recorded with the ideal voice timbre the indicator quantifies the impact of timbre changes introduced by the system under test. If the reference signal is not recorded with the ideal voice timbre the indicator quantifies to a certain extent the timbre of the reference recording.

Finally three additional compensation factors are used:

- A compensation factor for the behaviour with very high loud levels of additive noise that dominate perception with soft speech sounds (NOISE-L, corresponding variable in the algorithm is CVCratioSNRlevelRangecompensation0001 in Time.pdf and Disturbance.pdf, see Annex B). This compensation uses the average power level of the reference speech frames calculated over all soft pitch power density frames, i.e., using all frames between about −16 and −10 dB below the average active speech level. For this softset of frames both the average power level of the original reference ($P_{soft,ref,average)}$ and degraded frames ($P_{soft,degraded,average)}$ are calculated. The average power level of the reference and degraded frames are calculated using all non-loud speech speech active reference frames, (between about −35 and −3 dB below active speech level, $P_{active,ref,average}$ and $P_{active,degraded,average}$). From the four resulting average powers a NOISE-L multiplication factor for the MOS is calculated as follows:

$$\text{NOISE-L} = (\Delta_2 + (P_{\text{soft, degraded, average}} + \Delta_1) / (P_{\text{active, degraded, average}} + \Delta_1)) \Big/ (\Delta_2 + (P_{\text{soft, ref, average}} + \Delta_1) / (P_{\text{active, ref, average}} + \Delta_1))$$

  The parameters Δ_1 and Δ_2 are constant values that are used to adapt the behavior of the model to the behavior of subjects.

- A compensation factor for quantifying the impact of noise in frequency bands that contain no or little speech energy (EMPTY). This compensation first calculates the average amount of noise in the upper frequency bands (above 3 000 Hz) of the disturbance density over the silent intervals (NsilHigh) as derived from the reference speech. For all speech active frames the amount of energy in the upper frequency bands of the degraded signal is determined (above 3 000 Hz) from which the NsilHigh is subtracted and limited to a minimum to obtain a compensation factor (cf). Next the NsilHigh is devide by cf to obtain the empty speech band compensation factor (EMPTY).
- A compensation factor for the behaviour with loud modulated noise distortions as found with reference MNRU conditions. This compensation (MNRU, corresponding variable in Annex B is correlationOriginalWithDisturbanceCompensation000ForMNRU in Time.pdf and Disturbance.pdf) is based on the correlation between the loudness of the original speech file per frame and the disturbance level per frame using an L_3 aggregation over the Bark bands calculated over all speech active frames as derived from the reference signal.

The training of this mapping is carried out on a large set of degradations, including degradations that were not part of the ITU-T P.863 benchmark. These raw MOS scores are for the major part already linearized by the third order polynomial mapping used in the calculation of the MOS-like intermediate indicator (clause 9.7.13).

Finally, the raw ITU-T P.863 MOS scores are mapped towards the MOS-LQO scores using a third order polynomial that is optimized for the ITU-T P.863 set of databases. In narrowband mode, the maximum ITU-T P.863 MOS-LQO score is 4.5 while in super-wideband mode, this point lies at 4.75. An important consequence of the idealization process is that under some circumstances, when the reference signal contains noise or when the voice timbre is severely distorted, a transparent chain will not provide the maximum MOS score of 4.5 in narrowband mode or 4.75 in super-wideband mode. If in the operational situation the system under test is not judged over diffuse field equalized headphones, the signals that are provided to the ITU-T P.863 algorithm, both reference and degraded, should be filtered with the transfer characteristic that represents the final evaluation situation.

10 Conclusions

The ITU-T P.863 algorithm was tested on a wide range of test conditions and distortions and provides a very good tool for predicting subjectively determined MOS scores. The obtained prediction error is low (see Appendix I), even for distortion types that were not assessed in the ITU-T P.863 benchmark.

It is important to understand and consider the two different operational modes supported by the ITU-T P.863 algorithm:

- super-wideband mode for listening over super-wideband headphones;
- narrowband mode for listening over loosely coupled IRS type handsets.

In the super-wideband mode the impact of play back level is modelled and the default calibration factor (C) of 2.8 has to be used in combination with the standard –26 dBov scaling for play back levels of 73 dB(A) SPL (diotic). Play back levels down to 53 dB(A) SPL and up to 78 dB(A) SPL may be used and MOS-LQO scores should be reported in the format MOS-LQOsw (dB level). In narrowband mode only the play back level of 79dB(A) SPL (monotic) is supported. Narrowband mode MOS scores are referred to as MOS-LQOn.

The maximum ITU-T P.863 MOS-LQO score is 4.5 in narrow-band while in super-wideband mode this point lies at 4.75. Under some circumstances, when the reference signal contains noise or when the voice timbre is distorted, a transparent chain will not provide the maximum MOS score of 4.5 in narrowband mode or 4.75 in super-wideband mode.

It is important to realize that in super-wideband mode a high quality narrowband speech file recorded over a transparent link will get a MOS score that is highly dependent on the voice characteristics of

the recorded speech. A child voice containing a lot of high frequencies will be scored much lower than a male voice containing less high frequencies which is thus less affected by the narrowband low pass filtering. Note that a high quality transparent recording in the narrowband mode will always get a MOS score of 4.5.

It is recommended that all future quality measurements be carried out directly in super-wideband mode with a minimum of two high quality male and two high quality female voices.

Annex A

Conformance data and conformance tests

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

A.1 List of files provided for conformance validation

The conformance validation process described below makes reference to the following files, which are provided in the conform subdirectory of the electronic attachment:

For conformance tests in *narrowband* operational mode:

- Test_1a_results_ref.txt — *File pairs and ITU-T P.863 scores for test 1(a)*
- Test_1b_results_ref.txt — *File pairs and ITU-T P.863 scores for test 1(b)*
- Test_2a_results_ref.txt — *File pairs and ITU-T P.863 scores for test 2(a)*
- Test_2b_results_ref.txt — *File pairs and ITU-T P.863 scores for test 2(b)*
- process.bat — *Sample batch script to assist with preparing material for tests 1(b) and 2(a)*

For conformance tests in *super-wideband* operational mode:

- Test_3b_results_ref.txt — *File pairs and ITU-T P.863 scores for test 3(b)*
- Test_4b_results_ref.txt — *File pairs and ITU-T P.863 scores for test 4(b)*
- Test_5c_results_ref.txt — *File pairs and ITU-T P.863 scores for test 5(c)*
- Test_6c_results_ref.txt — *File pairs and ITU-T P.863 scores for test 6(c)*

The 'voipref' speech files are in raw format (16-bit linear PCM, little-endian byte ordering, at 8 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 conformance tests 1a to 6c, there are BAT files prepared. The BAT assume that the reference executable is called 'POLQA.exe' and is called as follows:

POLQA <reference.raw> <degraded.raw> <Sampling Frequency / Hz> <Mode 0: NB, 1: SWB>

A.2 Conformance tests

A.2.1 Conformance data sets

The data sets for the conformance tests are as given in Table A.1 below:

Table A.1 – Data sets for the conformance tests

Test	Number of file pairs	(a) 8 kHz data set	(b) 16 kHz data set	(c) 48 kHz data set
1	1328	Downsampled from ITU-T P-series Supplement 23 using ITU-T Software Tool Library[2] and process.bat. To be used in narrowband-mode of the ITU-T P.863 algorithm only (Test 1(a)). *– Mandatory –*	ITU-T P-series Supplement 23 To be used in narrowband-mode of the ITU-T P.863 algorithm (Test 1(b)). *– Not Mandatory –*	*Not applicable*
2	40	Clause A.1 'voipref' data. To be used in narrowband-mode of the ITU-T P.863 algorithm only (Test 2(a)). *– Mandatory –*	Upsampled from clause A.1 'voipref' data using Software Tool Library[2] and process.bat. To be used in narrowband-mode of the ITU-T P.863 algorithm (Test 2(b)). *– Not Mandatory –*	*Not applicable*
3	1328	*Not applicable*	ITU-T P-series Supplement 23 To be used in super-wideband mode of the ITU-T P.863 algorithm (Test 3(b)). *– Mandatory –*	*Not applicable*
4	40	*Not applicable*	Upsampled from clause A.1 'voipref' data using Software Tool Library[2] and process.bat. To be used in super-wideband mode of the ITU-T P.863 algorithm (Test 4(b)). *– Mandatory –*	*Not applicable*
5	200	*Not applicable*	*Not applicable*	Clause A.1 'SWB_TNO_601_48k' as attached. To be used in super-wideband mode of the ITU-T P.863 algorithm (Test 5(c)). *– Mandatory –*

[2] See [ITU-T G.191]

Table A.1 – Data sets for the conformance tests

Test	Number of file pairs	(a) 8 kHz data set	(b) 16 kHz data set	(c) 48 kHz data set
6	50	*Not applicable*	*Not applicable*	Clause A.1 'SWB_SQ_48k' as attached. To be used in super-wideband mode of the ITU-T P.863 algorithm (Test 6(c)). *– Mandatory –*
7	No data set defined. This test is open-ended, based on general, unknown data. *– Mandatory –*			

A.2.2 Conformance requirements

The test requirements are the confirmation of very narrow distribution of differences to the reference values provided in clause A.1.

The allowed distribution of differences across all mandatory tests 1(a), 2(a), 3(b), 4(b), 5(c) and 6(c) are summarized in the following table below. The requirements are based on the absolute difference in the ITU-T P.863 score between the implementation under test and the reference values given in clause A.1.

Table A.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 other databases than the defined ones in this annex, the same error distribution must not be exceeded. For unknown data, a test set of at least 2000 file pairs – preferably from complete subjective experiments – has to be taken for that statistics.

A.3 Conversion of sampling rates

In Test 1(a), 8 kHz resampled versions of the files associated with [ITU-T P-Sup.23] are used, on a file-by-file basis. The original and degraded files must be *downsampled* using the ITU-T Software Tool Library ([ITU-T G.191]), program `filter`, using the following command:

```
filter -down HQ2 inputfile.raw outputfile.raw
```

This assumes that the 16 kHz input speech file is called inputfile.raw and the 8 kHz output file is called outputfile.raw.

A batch script to assist with this, and the original and degraded file names are provided in the files listed above.

In Test 2(b) and 4(b), a 16 kHz re-sampled versions of the Annex A VoIP test files are used on a file-by-file basis. The original and degraded files must be *upsampled* using the ITU-T Software Tool Library ([ITU-T G.191]), program filter, using the following command:

filter -up HQ2 inputfile.raw outputfile.raw

This assumes that the 8 kHz input speech file is called inputfile.raw and the 16 kHz output file is called outputfile.raw.

A batch script to assist with this, and the original and degraded file names are provided in the files listed above.

A.4 Digital attachments

Processing scripts and reference results are contained in an electronic attachment to this Recommendation in the following folder:

- "process_results" (with .bat files)

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

- "VOIP"
- "SWB_SQ_48k"
- "SWB_TNO_601"

Annex B

Detailed Descriptions of the ITU-T P.863 algorithm in pdf-format

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

Detailed descriptions of the components of the ITU-T P.863 algorithm are contained in the PDF files in the folder: "Detailed Descriptions", included as an electronic attachment to this Recommendation.

Appendix I

Reporting of the performance results for the ITU-T P.863 algorithm based on the rmse* metric

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

I.1 Purpose of this appendix

This appendix reports the performance results of the ITU-T P.863 algorithm for all 62 databases used in the ITU-T P.863 evaluation; this is done in order to give the readers of this Recommendation some insight into the procedures which, during the standards making process, finally led to the establishment of ITU-T P.863.

NOTE 1 – The statistical data reported here have been restricted to those performance metrics which formed the basis of the standards making process in ITU-T.

NOTE 2 – These statistical data are explicitly not intended for the users of the ITU-T P.863 algorithm assuming that the typical user of the ITU-T P.863 algorithm will not have access to and knowledge about the speech databases and the statistical evaluations having led to these data.

NOTE 3 – ITU-T P.863 performance criteria from a user's perspective will be determined in a so-called characterization phase, the results of which will be published separately.

I.2 Overview

The performance metric is the so-called rmse*, that is, an rmse that considers the confidence interval of the individual MOS scores.

The basis for the main objectives for the evaluation of the ITU-T P.863 algorithm is a so-called epsilon-insensitive rmse It is calculated as the traditional rmse but small differences to the target value will not be counted. This rmse considers only differences related to an epsilon-wideband around the target value. This 'epsilon' is defined as the 95% confidence interval of the subjective MOS value. By definition, the uncertainty of the MOS is taken into account in this evaluation. The rmse* is calculated on a prediction error as illustrated in Figure I.1.

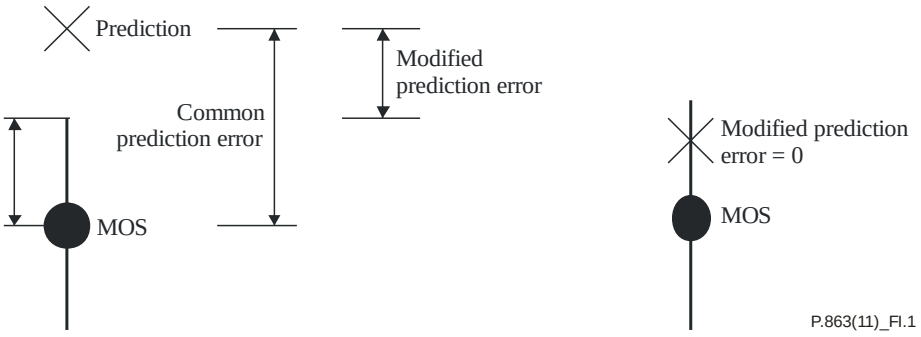


Figure I.1 – Calculation of rmse*

The complete statistical evaluation method is described in clause I.4.

I.3 Performance results for the ITU-T P.863 algorithm

The two rmse* values for the ITU-T P.863 algorithm per database shown in Tables I.1 to I.3 are obtained after a monotonous 3rd order mapping or after a 1st order linear mapping of the ITU-T P.863 results, respectively. The rmse* values reflect the 'per condition' performance of the ITU-T P.863 algorithm. The prediction becomes less accurate as the rmse* increases. Values of rmse* < 0.1 can be seen as proper and accurate predictions. The prediction becomes less accurate as the rmse* increases; however, high values can be caused by either single outliers or a general lower prediction accuracy. The ITU-T P.863 algorithm does not result for any tested database in a rmse* > 0.3.

Table I.1 – Two rmse* values for ITU-T P.863 per database (Set 1)

Database	rmse* 3rd	rmse* 1st
Set 1	ITU-T P.863	ITU-T P.863
NB_BT_P862_BGN_ENG	0.0945	0.1703
NB_BT_P862_PROP	0.1514	0.1517
NB_DT_P862_1st	0.1648	0.1616
NB_DT_P862_BGN_GER	0.0917	0.1638
NB_DT_P862_Share	0.0683	0.0707
NB_ERIC_AMR_4B	0.1310	0.1301
NB_ERIC_P862_NW_MEAS	0.1590	0.1668
NB_TNO_P862_KPN_KIT97	0.2317	0.2615
NB_TNO_P862_NW_EMU	0.1154	0.1300
NB_TNO_P862_NW_MEAS	0.2200	0.2460
NB_ITU_SUPPL23_EXP1a	0.1385	0.1658
NB_ITU_SUPPL23_EXP1d	0.0753	0.0772
NB_ITU_SUPPL23_EXP1o	0.1133	0.1192
NB_ITU_SUPPL23_EXP3A	0.1824	0.2278
NB_ITU_SUPPL23_EXP3C	0.0601	0.1164
NB_ITU_SUPPL23_EXP3d	0.0649	0.0776
NB_ITU_SUPPL23_EXP3o	0.0890	0.1164
NB_FT_P563_PROP	0.0915	0.1185
NB_LUC_P563_PROP	0.0702	0.0717
NB_OPT_P563_PROP	0.1081	0.1691
NB_PSY_P563_PROP	0.1381	0.2246
NB_SQ_P563_PROP	0.1393	0.1384
Average	0.1227	0.1489

Table I.2 – Two rmse* values for ITU-T P.863 per database (Set 2)

Database	rmse* 3rd	rmse* 1st
Set 2	ITU-T P.863	ITU-T P.863
NB_ATT_iLBC	0.1629	0.163
NB_ERIC_Field_GSM_EU	0.1865	0.2003
NB_ERIC_Field_GSM_US	0.1378	0.1395
NB_GIPS_EXP1	0.0798	0.1878
WB_GIPS_EXP3	0.0926	0.2026
SWB_GIPS_EXP4	0.0851	0.0885
NB_QUALCOMM_EXP1b	0.0952	0.1217
WB_QUALCOMM_EXP1w	0.0874	0.1082
NB_QUALCOMM_EXP2B	0.117	0.1223
NB_QUALCOMM_EXP3w	0.0466	0.0579
WB_QUALCOMM_EXP3w	0.0265	0.0826
SWB_48kHz101_ERICSSON	0.2775	0.2926
WB_48kHz102_ERICSSON	0.1732	0.1756
SWB_48kHz201_FT_DT	0.2504	0.25
SWB_48kHz202_FT_DT	0.2603	0.279
SWB_48kHz301_OPTICOM	0.2465	0.2644
SWB_48kHz302_OPTICOM	0.1468	0.1783
SWB_48kHz401_PSYTECHNICS	0.1348	0.1575
WB_48kHz402_PSYTECHNICS	0.1325	0.1326
SWB_48kHz501_SWISSQUAL	0.1974	0.2035
SWB_48kHz502_SWISSQUAL	0.252	0.2481
SWB_48kHz601_TNO	0.1942	0.1913
SWB_48kHz602_TNO	0.1171	0.1426
Average	0.1522	0.1735

Table I.3 – Two rmse* values for ITU-T P.863 per database (Set 3)

Database	rmse* 3rd	rmse* 1st
Set 3	ITU-T P.863	ITU-T P.863
SWB_48kHz103_ERICSSON	0.185	0.1978
NB_8kHz104_Ericsson	0.2618	0.272
SWB_48kHz203_FTDT	0.2522	0.2499
WB_16kHz204_FTDT	0.2266	0.2251

Table I.3 – Two rmse* values for ITU-T P.863 per database (Set 3)

Database	rmse* 3rd	rmse* 1st
SWB_48kHz303_OPTICOM	0.1897	0.1982
SWB_48kHz403_PSYTECHNICS	0.1372	0.1359
NB_8kHz404_PSYTECHNICS	0.1795	0.1783
SWB_48kHz503_SWISSQUAL	0.1857	0.1829
NB_8kHz504_SWISSQUAL	0.2194	0.2154
SWB_48kHz603_TNO	0.1144	0.1191
NB_8kHz_NTT_PTEST_1	0.0776	0.0773
NB_QUALCOMM_EXP4	0.1107	0.1294
WB_QUALCOMM_EXP5	0.0664	0.1237
NB_QUALCOMM_EXP6a	0.1802	0.204
NB_QUALCOMM_EXP6b	0.1262	0.2444
NB_16kHz_HUAWEI_1	0.1013	0.1173
NB_16kHz_HUAWEI_2	0.1457	0.1947
Average	0.1623	0.1803

An important parameter is the worst case performance that was also part of the evaluation procedure.

Table I.4 – Worst case performance of the ITU-T P.863 algorithm (Sets 1 to 3)

Database	rmse* 3rd	rmse* 1st
Absolute worst case over the three sets	0.2775	0.2926
Average of the three worst experiments	0.2665	0.2812

The performance of the ITU-T P.863 algorithm compared to the algorithm in [b-ITU-T P.862] is shown in Tables I.5 and I.6. Here the ITU-T P.862 values are mapped using [b-ITU-T P.862.1]. Tables I.5 and I.6 are restricted to narrow-band databases. The ITU-T P.863 algorithm shows a reduction of rmse* by 27% compared with that of [b-ITU-T P.862.1] after 3rd order mapping and one of 33% after first order mapping.

Table I.5 – Performance of ITU-T P.863 compared to ITU-T P.862 (Set A)

Database	rmse* 3rd		rme* 1st	
Set A (narrowband)	ITU-T P.862.1	ITU-T P.863	ITU-T P.862.1	ITU-T P.863
NB_BT_P862_BGN_ENG	0.1490	0.0945	0.2182	0.1703
NB_BT_P862_PROP	0.1603	0.1514	0.1860	0.1517
NB_DT_P862_1st	0.1916	0.1648	0.2070	0.1616
NB_DT_P862_BGN_GER	0.0973	0.0917	0.1465	0.1638

Table I.5 – Performance of ITU-T P.863 compared to ITU-T P.862 (Set A)

Database	rmse* 3rd		rme* 1st	
NB_DT_P862_Share	0.1263	0.0683	0.1276	0.0707
NB_ERIC_AMR_4B	0.0918	0.1310	0.0999	0.1301
NB_ERIC_P862_NW_MEAS	0.2214	0.1590	0.2406	0.1668
NB_TNO_P862_KPN_KIT97	0.2967	0.2317	0.3370	0.2615
NB_TNO_P862_NW_EMU	0.3017	0.1154	0.2983	0.1300
NB_TNO_P862_ NW_MEAS	0.2493	0.2200	0.2654	0.2460
NB_ITU_SUPPL23_EXP1a	0.1342	0.1385	0.1644	0.1658
NB_ITU_SUPPL23_EXP1d	0.0780	0.0753	0.0957	0.0772
NB_ITU_SUPPL23_EXP1o	0.1091	0.1133	0.1386	0.1192
NB_ITU_SUPPL23_EXP3a	0.1939	0.1824	0.2357	0.2278
NB_ITU_SUPPL23_EXP3c	0.1370	0.0601	0.1891	0.1164
NB_ITU_SUPPL23_EXP3d	0.1258	0.0649	0.1290	0.0776
NB_ITU_SUPPL23_EXP3o	0.1537	0.0890	0.1725	0.1164
NB_FT_P563_PROP	0.1139	0.0915	0.1188	0.1185
NB_LUC_P563_PROP	0.0632	0.0702	0.0821	0.0717
NB_OPT_P563_PROP	0.1150	0.1081	0.1315	0.1691
NB_PSY_P563_PROP	0.1623	0.1381	0.1696	0.2246
NB_SQ_P563_PROP	0.1915	0.1393	0.1941	0.1384
Average	0.1574	0.1227	0.1795	0.1489

Table I.6 – Performance of ITU-T P.863 compared to ITU-T P.862 (Set B)

Database	rmse* 3rd		rme* 1st	
Set B (narrowband)	ITU-T P.862.1	ITU-T P.863	ITU-T P.862.1	ITU-T P.863
NB_ATT_iLBC	0.2268	0.1629	0.2243	0.1630
NB_ERIC_Field_GSM_EU	0.2401	0.1865	0.2458	0.2003
NB_ERIC_Field_GSM_US	0.1986	0.1378	0.2245	0.1395
NB_GIPS_EXP1	0.2943	0.0798	0.3906	0.1878
NB_QUALCOMM_EXP1b	0.1588	0.0952	0.2593	0.1217
NB_QUALCOMM_EXP2b	0.1826	0.1170	0.2591	0.1223
NB_QUALCOMM_EXP3w	0.1546	0.0466	0.2219	0.0579
NB_8kHz104_ERICSSON	0.3570	0.2618	0.3826	0.2720
NB_48kHz404_PSYTECHNICS	0.3260	0.1795	0.3412	0.1783

Table I.6 – Performance of ITU-T P.863 compared to ITU-T P.862 (Set B)

Database	rmse* 3rd		rme* 1st	
NB_8kHz504_SWISSQUAL	0.4203	0.2194	0.4166	0.2154
NB_8kHz_NTT_PTEST_1	0.1073	0.0776	0.1150	0.0773
NB_QUALCOMM_EXP4	0.1730	0.1107	0.2533	0.1294
NB_QUALCOMM_EXP6a	0.2480	0.1802	0.3074	0.2040
NB_QUALCOMM_EXP6b	0.1491	0.1262	0.2865	0.2444
NB_16kHz_HUAWEI_1	0.1719	0.1013	0.2026	0.1173
Average	0.2272	0.1388	0.2754	0.1620

Table I.7 gives the performance of the ITU-T P.863 algorithm compared with that of [b-ITU-T P.862.2]. In the ITU-T P.863 evaluation set six 'common' wideband databases (up to 7 000 Hz audio bandwidth) were used. The ITU-T P.863 algorithm results in a reduced rmse* of 56% compared with that of [b-ITU-T P.862.2].

Table I.7 – Performance of ITU-T P.863 compared to ITU-T P.862.2 (Set C)

Database	rmse* 3rd		rme* 1st	
Set C (wideband)	ITU-T P.862.2	ITU-T P.863	ITU-T P.862.2	ITU-T P.863
WB_48kHz102_ERICSSON	0.4521	0.1732	0.4482	0.1756
WB_16kHz402_PSYTECHNICS	0.3245	0.1325	0.3646	0.1326
WB_GIPS_EXP3	0.3467	0.0926	0.4255	0.2026
WB_QUALCOMM_EXP1w	0.2606	0.0874	0.3664	0.1082
WB_QUALCOMM_EXP3w	0.2852	0.0466	0.3727	0.0579
WB_16kHz204_FT_DT	0.4221	0.2266	0.4158	0.2251
WB_QUALCOMM_EXP5	0.3236	0.0664	0.3773	0.1237
Average	0.3450	0.1179	0.3958	0.1465

I.4 Calculation of rmse*

I.4.1 Uncertainty of subjective results and calculation of confidence intervals

In the statistical performance evaluation of the ITU-T P.863 algorithm, the uncertainty of the subjective votes was taken into account. This is important since an objective model will not be able to predict an average subjective opinion more accurately than the average subject itself.

Usually, the standard deviation and its corresponding confidence interval (ci95) are calculated, in order to determine the degree of uncertainty of the subjects' votes. These statistical parameters are either based on a so-called file-based or condition-based analysis to either determine the uncertainty of the subjects per file or per test condition. A test condition consists of a set of files with, for example, two male and two female talkers where the files are processed under the same technical circumstances (i.e., applying the same distortions).

The reasons for the average subject's uncertainty of opinion are various.

Examples: A specific talker or sentence the talker speaks might offend one part of the listener group while others might be pleased by the voice and the content of the file. But not only the listeners might be biased by the different talkers and the content but also a system under test might have its 'preferences'. For example, when low pitched voices are better encoded than high pitched voices. The subjects might vote accordingly but, within a test condition, we would observe a large variability of the average opinion on that condition.

Consequently, in a file-based analysis, the two effects above can be easily identified and separated. In a condition-based analysis, however, the two effects will be mixed-up, thus making it difficult for an objective model to predict a solid opinion. Since the ITU-T P.863 algorithm is primarily evaluated on a condition-basis, these effects had been taken into account by introducing the rmse* metric. The next clauses explain how the statistical parameters were calculated for use in file- and condition-based analysis.

I.4.2 The confidence interval

Usually, the confidence interval (ci95) is calculated under consideration of all individual scores for a single file or test condition. The standard deviation σ and the number of individual scores *M* determines the confidence interval. It is recommended to use the accurate T-value for the given *M*.

$$ci_{95} = t(0.05, M)\frac{\sigma}{\sqrt{M}} \tag{I-1}$$

Depending on whether a file-based analysis or a condition-based analysis is performed the calculation of the standard deviation σ is computed as described in the following clauses. The number of scores *M* is then be replaced by the number of votes per file *K* in the file-based analysis and in the condition-based analysis *M* is replaced by the number of votes per condition *N*.

I.4.3 The standard deviation for file-based analysis

The standard deviation σ_j of the individual votes $v_{j,l,k}$ of listener *k* for an audio file spoken by talker *l* and processed by condition *j* is defined as follows:

$$\sigma_{j,l} = \sqrt{\frac{\sum_{k=1}^{K}\left(v_{j,l,k} - MOSLQS_{j,l}\right)^2}{K-1}} \tag{I-2}$$

Condition j, $j \in \{0,1,...\}$,

Listener k, $k \in \{1,2,...,K\}$,

Talker l, $l \in \{1,2,...,L\}$ *(L is equivalent with the number of files per test condition),*

with $MOSLQS_{j,l}$ as "Mean Opinion Score Listening Quality Subjective" for talker *l* of condition *j* as defined in:

$$MOSLQS_{j,l} = \frac{1}{K}\sum_{k=1}^{K} v_{j,l,k} \tag{I-3}$$

I.4.4 The standard deviation for condition-based analysis

The standard deviation σ_j for condition-based analysis is defined as follows making use also of the definitions in the previous clause:

$$\sigma_j = \sqrt{\frac{\sum_{l=1}^{L}\sum_{k=1}^{K}\left(v_{j,l,k} - MOSLQS_{j,l}\right)^2}{N-1}} = \sqrt{\frac{K-1}{N-1}\sum_{l=1}^{L}\sigma_{j,l}{}^2} \quad \text{(I-4)}$$

N denotes the number of votes per condition.

I.4.5 Exceptional cases

The calculation of σ_j and $\sigma_{j,l}$ as well the corresponding ci95 as described above is the regular way of calculation. However, it requires the access to the individual votes or at least to a per-file standard deviation and/or confidence interval along with the number of votes behind.

For some of the existing databases, this information is not available. In those cases, the following simplification will be applied:

1) If only the *MOS* per-condition is provided, this value is used for the per-file evaluation as well. This per-condition MOS will be used as MOS for each file. The required ci95 per-file is obtained according to the following rules in that case.

2) If only *standard deviation* per-condition is provided, this value is used for the per-condition evaluation instead of σ_j as described above. It is accepted that the systematic over-/under-prediction of a speaker or a file may influence (increase) that value.

   The ci95 per-file (required for secondary analysis) can be derived by the simplification that the standard deviation is constant and equal for all files in that condition. Under this simplification the ci95 per-file can be derived by:

$$ci_{95}(per\ file) = t(0.05, M)\frac{\sigma(per\_condition)}{\sqrt{M}} \quad \text{(I-5)}$$

   with *M* as the number of votes per-file[3].

3) If only the *confidence interval ci95* per-condition is provided, this value is used for the per-condition evaluation directly. It is accepted that the systematic over-/under-prediction of a speaker or a file may influence (increase) that value.

   The ci95 per-file (required for secondary analysis) can be derived again by the simplification that the standard deviation is constant and equal for all files in that condition. Under this simplification the ci95 per-file can be derived by:

$$ci_{95}(per\_file) = ci_{95}(per\_condition)\frac{\sqrt{N}}{\sqrt{M}} \quad \text{(I-6)}$$

   with *N* as number of vote for the entire condition and *M* as the number of votes per-file[3].

4) If there is neither *standard deviation* nor *confidence interval* available the ci95 (per-condition) is set fix to 0.2. The ci95 (per-file) is calculated as in point 2.

[3] In case that *M* is unknown, it might be assumed by *N* divided by the number of files scored in one condition (i.e., if *N* = 96 and for files were scored in one condition, *M* = 24).

I.4.6 Epsilon-insensitive root mean square error (rmse*)

The basis for the main objectives for the ITU-T P.863 evaluation was a so-called epsilon-insensitive rmse. It is calculated as the traditional rmse but small differences to the target value will not be counted. This rnse considers only differences related to epsilon-wideband around the target value. This 'epsilon' is defined as the 95% confidence interval of the subjective MOS value. By definition, the uncertainty of the MOS is taken into account in this evaluation. This modified rmse can be described as follows:

$$Perror(i) = \max(0, |MOSLQS(i) - MOSLQO(i)| - ci_{95}(i)) \quad \text{(I-7)}$$

where the index *i* denotes the condition or the speech sample.

The final modified *rmse** is calculated as usual but based on *Perror* with the formula:

$$rmse^* = \sqrt{\left(\frac{1}{N-d}\sum_N Perror(i)^2\right)} \quad \text{(I-8)}$$

where the index *i* denotes the condition or the speech sample, *N* denotes the number of conditions or speech samples and *d* the number of freedoms. The degree of freedom *d* is set to 4 in case a 3rd order mapping is applied in prior to the values, *d* is set to 2 in case of a 1st order mapping.

The rmse* is calculated per database and gives an impression how the prediction error exceeds the ci95.

I.5 Scatter plots

In order to illustrate the distribution of MOS values that are related to a certain rmse* value, this appendix provides some scatter plots taken from the evaluation of the ITU-T P.863 algorithm. For each bandwidth (narrowband, wideband and super-wideband) two charts are presented. The first chart shows the performance for the best case experiment and the second one the performance of the worst case performance of the model. As selection criterion, the first order mapped per condition rmse* has been used. Please note that in the title of each chart the per condition rmse* after 3rd order mapping is mentioned.

I.5.1 Scatter plots of the best and worst case narrowband experiment

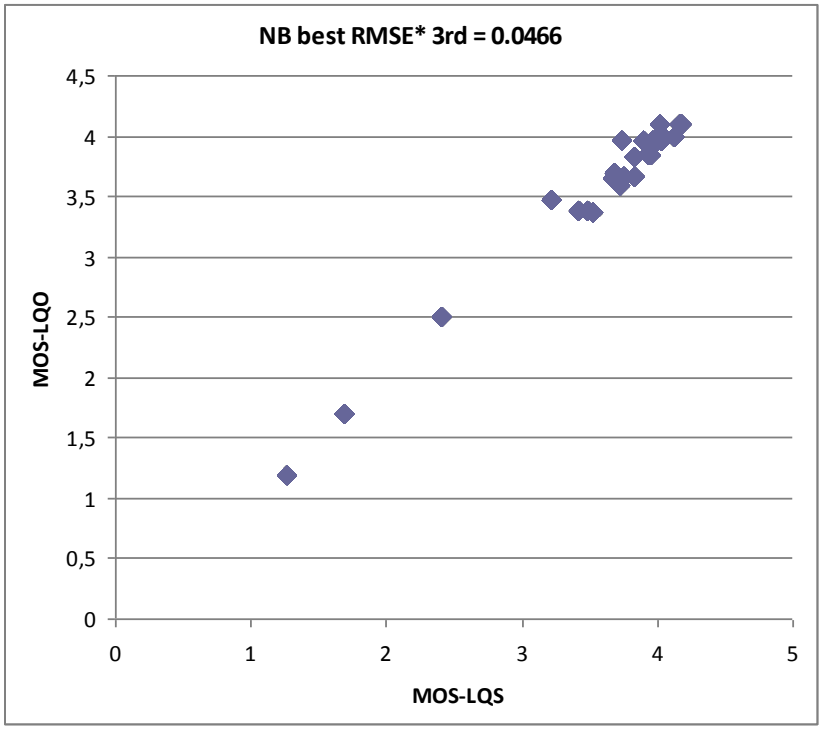


Figure I.2 – Performance for the best case narrowband experiment after 1st order mapping

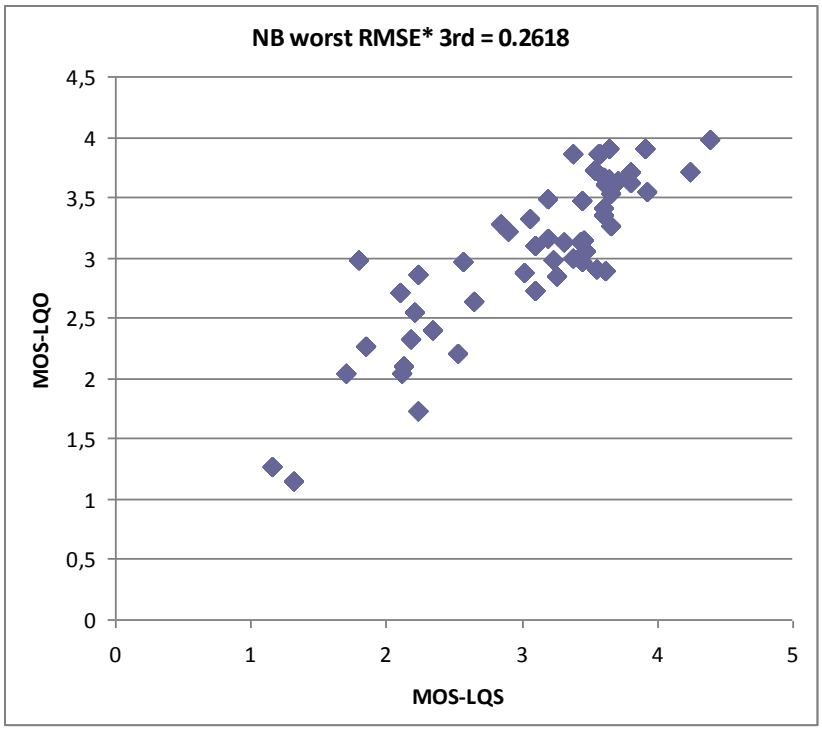


Figure I.3 – Performance for the worst case narrowband experiment after 1st order mapping

I.5.2 Scatter plots of the best and worst case wideband experiment

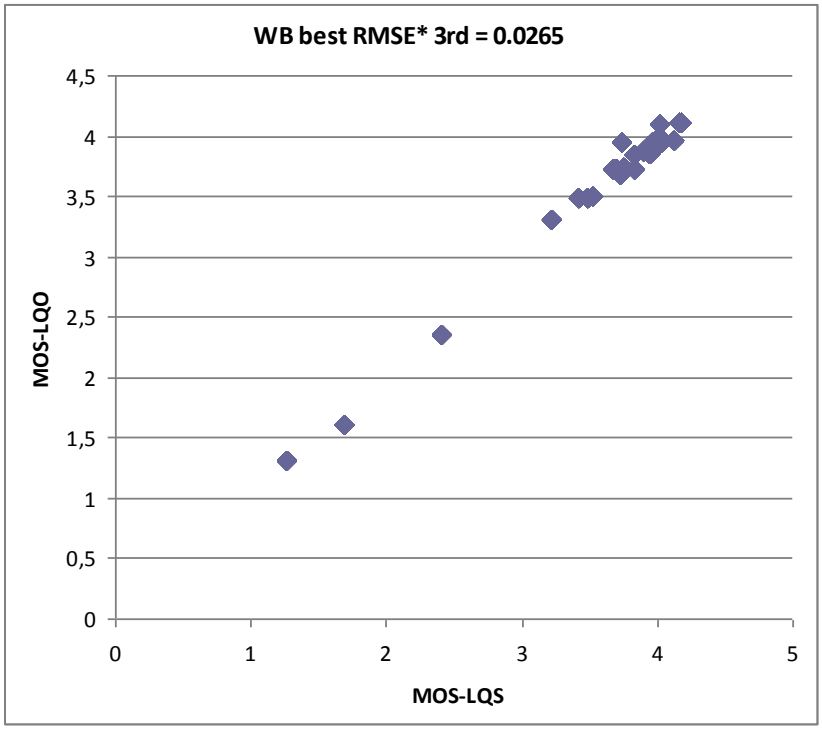


Figure I.4 – Performance for the best case wideband experiment after 1st order mapping

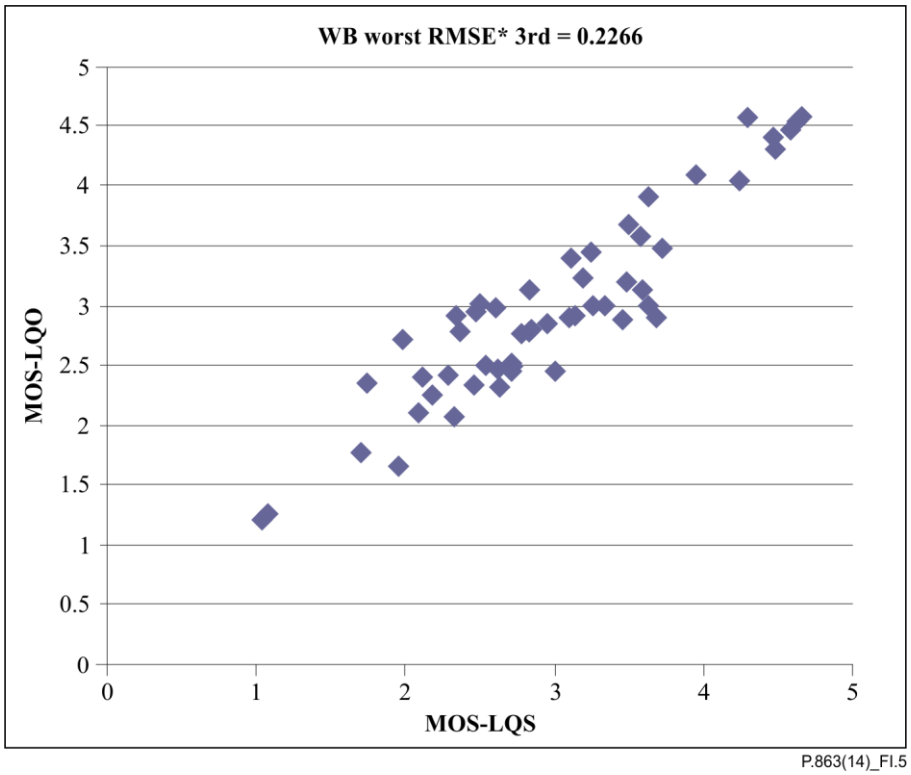


Figure I.5 – Performance for the worst case wideband experiment after 1st order mapping

I.5.3 Scatter plots of the best and worst case super-wideband experiment

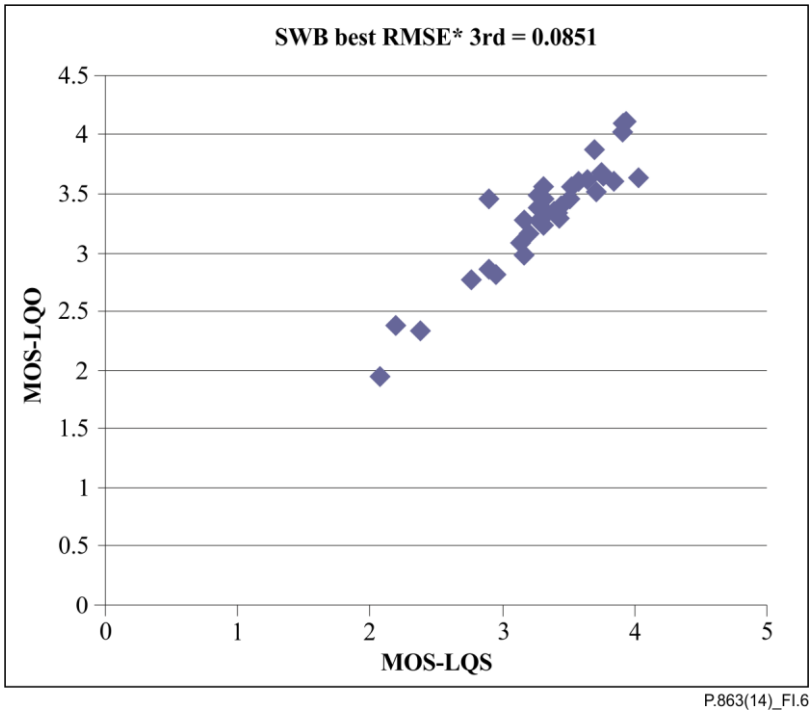


Figure I.6 – Performance for the best case super-wideband experiment after 1st order mapping

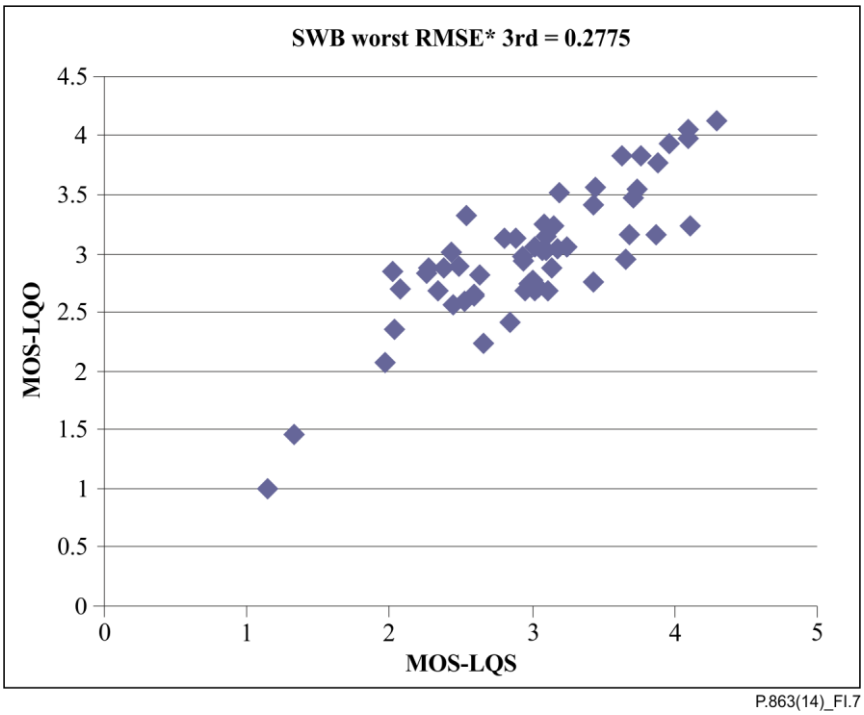


Figure I.7 – Performance for the worst case super-wideband experiment after 1st order mapping

Appendix II

Description of the "full-scale" subjective tests in a super-wideband context conducted for the ITU-T P.863 algorithm training and validation

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

The "full-scale" super-wideband experiment design was introduced during the development of ITU-T P.863 to accommodate for the lack of super-wideband speech subjective assessment methodology. The subjective test methodology used for ITU-T P.863 is based on the ACR methodology of [ITU-T P.800] and [ITU-T P.830] but features additional constraints. All super-wideband databases provided by the proponents had to follow the defined design rules detailed in this appendix.

This experiment design was denoted as "full-scale" as each subjective test had to cover the entire range of degradation dimensions which includes background noise, linear filtering, presentation level and temporal clipping.

II.1 Database structure and subjects requirement

The full-scale super-wideband experiments had to comply with the following rules:

- Minimum 44 conditions including the anchor conditions.
- Each file must be assessed by at least 8 subjects.
- Each condition must include at least four talkers.
- Each condition must be assessed by at least 96 votes.

II.2 Anchor conditions

The full-scale super-wideband experiments had to contain the following 12 anchor conditions:

Reference:	clean, 0 dB attenuation, super-wideband (50 to 14 kHz)
MNRU:	10 dB and 25 dB (modified MNRU using ITU-T P.50 shaped noise for modulation)
Background noise:	12 dB Hoth and 20 dB babble SNR (relation between ITU-T P.56 measured at the clean reference and the r.m.s noise level)
Level according to ITU-T P.56:	–10 dB and –20 dB from nominal level (–26 dBov)
Linear filtering:	narrowband IRS send and receive filtered[4], 500 to 2 500 Hz and 100 to 5 000 Hz
Temporal clipping:	2% and 20% packet loss, packet size 20 ms without packet loss concealment. A tool for packet loss insertion was provided by SwissQual.

[4] The linear filtering IRS (send+receive) will be constructed as follows: 48 kHz reference signal down-sampled to 16 kHz, modified IRS send filtered, modified IRS receive filtered, up-sampled to 48 kHz, presentation level 0 dB relative to nominal level.

II.3 Design rules of test conditions for full-scale mandatory tests

The distribution of the degradations had to meet the following requirements:

- Background Noise (>30% noisy), see definition of Background Noise in clause II.5
- Audio band-limitations (>10% super-wideband, >30%wideband, >30% narrowband)
- Presentation level
    - >10% –20 to –12 dB
    - >10% –12 to –3 dB
    - >60% –3 to +3 dB
    - >10% +3 to +6 dB

The limitations given above applied to the test conditions only (except anchor conditions).

Further degradation types were required for the mandatory full-scale experiments. Here the objectives were to reach an average over all full-scale databases submitted. This allowed individual databases to focus on certain degradation types or even not include some types at all. The objectives had to be met without consideration of the anchor conditions.

- At least 40% live and at least 40% simulated network situations
- At least 15% variable delay (VoIP, video-telephony)/time warping
- About 5% amplitude clipping (overload, saturation)
- At least 60% speech codecs as used in telecommunication scenarios today including tandeming
- At least 15% bit-errors (wireless connections), five different patterns
- Five different PLC strategies; at least 15%
- Five different packet loss pattern, including front-end clipping (temporal clipping); at least 15%
- Speech enhancement systems in networks and terminals, including gain variations, send and receive side; about 10%
- Reverberations caused by acoustical coupling on the *receiving* side; three reverberations/rooms/locations; no more than 20% acoustical recordings
    - Non-linear distortions produced by the microphone/transducer at acoustical interfaces (combined with reverberation)
    - Influence of linear distortions (spectral shaping), also time variant
- Reverberations caused by acoustical coupling on the *sending* side; three reverberations/rooms/locations; no more than 20% of acoustical recordings
    - Non-linear distortions produced by the microphone/transducer at acoustical interfaces (combined with reverberation)
    - Influence of linear distortions (spectral shaping), also time variant.

II.4 Reference and degraded speech material

Each reference sample consisted of a sentence pair spoken by a speaker. It was required to use a minimum of 16 different reference samples, spoken by at least four different speakers and with no repetition of text. Repetitive use of reference samples was allowed, but it was required that subjects were not presented with a reference sample repetition within six consecutive samples.

Each reference speech file consists of two sentences separated by a gap of at least 1 s but not more than 2 s. The minimum amount of active speech in each file is 3 s. The first speech activity starts between 0.5 s and 2 s. The last speech activity ends between 0.5 s and 2.5 s before the end of the

sample (file). The speech activity according to [ITU-T P.56] calculated over the entire file has to be between 35% and 65%. The entire file length has to be between 8 s and 12 s.

The noise floor of the reference files should not exceed –84 dBov (A) in the leading and trailing parts as well as in the gaps between the sentences.

Usually, the temporal structure of the speech signals is widely kept largely preserved during transmission. The degraded and captured signals have to follow the same rules as defined above[5].

Special rules for background noise conditions and time warping are defined separately later in this appendix.

II.5 Transmission and capturing capture of speech material superimposed interlaced with background noises

The term Background Noise is used to refer to an additive noise to superimposed on the speech. Multiplicative or modulated noises such as MNRU are not considered as background noise.

The noise should be added at least over the entire file-length. In each case noisy sequences using additive background noise have to start with a preamble of 1 s to 2 s of noise and the duration of the trailing noise must be at least 1s but not longer than 2.5 s. In the case of shorter leading/trailing sequences in the reference files, the added noise file should extend these sequences. The maximum sample length of 12 s and the minimum speech activity of 35% must be adhered to.

In each case, noisy sequences using additive background noise will start with a preamble between 1 s and 2 s noise only and the duration of the trailing noise must be 1 s but not longer than 2.5 s.

Due to the potential time-warping effect of the actual noisy gap between the two sentences, the active speech and the leading/trailing sequences may differ from the reference signal.

It can be helpful to transmit longer initial and trailing sequences over the test set up (e.g., for reaching convergence of noise suppression). The leading and trailing sequences have to be cut to the required length before presenting those samples in the auditory test and to the ITU-T P.863 algorithm.

For simplification, the SNR is described as the ratio between the power in the active speech parts[6] and the A-weighted power of the noise in speech pauses (e.g., between the individual sentences). A condition is considered as a background noise condition if the SNR is less than 35 dB by applying the described measuring procedure.

The desired amount of noise should be present for the listeners. That means the noise has to be present in the files used in the listening test. This defines the measuring point of the SNR the point at which the SNR is measured for each sample.

Basically, the S/N ratio for rating a condition as a 'Background Noise' or as a 'Clean' condition or as 'Clean' has to be measured at the file to be presented in the listening test. If there are no doubts about the use of noise reduction systems, the measurement can be done at the input files to the test condition. In case of the use of IF noise suppression systems are used anywhere in the test channel or the terminal or if there is any doubts, the measurement has to be done at the received and recorded files. Files with a SNR greater than 35 dB are not considered as 'Background Noise'.

---

5 The structure of the degraded signal does not have to exactly match the structure of the corresponding reference/input signal. It only has to meet the general rules. For example, the reference signal may have a leading silence period of 1.2 s but the degraded signal has one of 2.3 s due to the capture process.

6 NOTE – The example program of [ITU-T P.56] can deliver wrong results of active speech level in case of high background noise. Here speech pauses are not recognized anymore and are counted as speech. Consequently, the measured ASL is closer to the main r.m.s. level.

II.6 Transmission and capturing capture of speech material under time warping conditions

In the case of time warping and/or temporal stretching/compression, the temporal structure may change compared to the reference signal used as input signal. The same limits as defined in clause II.4 are applied to the captured degraded signals under time warping conditions. This means, that in case of intended or expected temporal stretching or compression is present, reference files have to be chosen. They have to meet the requirements of clause II.4 after transmission.[7]

II.7 Subjective test set up for assessing super-wideband speech quality

The basic design of the subjective test is derived from [ITU-T P.800] and [ITU-T P.830] using the ACR scale. However as there was no standardized test for assessing super-wideband speech quality available during the ITU-T P.863 benchmark new supplementary design rules were developed.

The room used for recording the reference super-wideband speech files must have a reverberation time below 300 ms above 200 Hz (e.g., an anechoic chamber). Recordings must be made using omni-directional microphones. The distance to the microphone must be approximately 10 cm. Background noise must be below 30 dBSPL(A). Speech signals will be band pass filtered to 20 Hz to 14 kHz. Directional microphones are allowed under the condition that the spectral balance is the same as with the before mentioned omni-directional microphones.

All super-wideband tests are based on 48 kHz sampled mono speech signals that are filtered with a 50 to 14 000 Hz band-pass filter (according to [ITU-T G.191], 14KBP) before using them in the subjective test. Speech files are played to the subject using a diotic headphone presentation (both ears receive the same mono signal) with a diffuse field equalization.

The play back presentation level is calibrated with a pink noise reference signal scaled to the default digital level of –26 dBov (obtained with the ITU-T P.56 algorithm) at a nominal level of 73 dB (A) SPL at the entrance of an artificial ear (ear reference point). The play back level of the speech signal is restricted to –21 dBov and –46 dBov.

Subjects must have a normal hearing in audio bandwidth up to 8 kHz (max. hearing loss of 20 dB). The subjects are split into three groups according to their age. These groups consist of subjects which are between 15 and 30 years, between 30 and 50 years and above 50 years old. At least 20% of all the subjects in the experiment must fall into each group. Within each group at least 40% of the subjects must be female and at least 40% must be male.

II.8 Limitations in subjective test results

Ten of the full-scale super-wideband subjective tests used in the ITU-T P.863 benchmark were examined for consistency of the subjective scores. These ten tests were conducted by seven different laboratories. The analysis showed that the perceived quality of signals of different degradation types can vary between experiments due to dependencies on test design, source material characteristics and listener's judgement of the test scenario. In particular it was noted that the rank order of conditions from different distortion classes can change from experiment to experiment.

---

[7] NOTE – In the case of intended stretching, reference files with separating pauses below 2 s should be used to give space for stretching. In the case of intended compression, reference files with separating pauses close to 2 s and with sufficient active speech should be used. Potentially, longer leading/trailing sequences have to be transmitted. The given file durations can be derived by post cut-off prior to the listening tests.

Apart from the anchor conditions, most of the other conditions are unique across the tests, thus it is not possible to perform other comparisons. However, the described effects are very likely to influence all conditions within the test. Therefore, the results of individual experiments may not provide consistent comparison of all degradation types.

As these experiments involve a mixture of narrowband, wideband and super-wideband, these conclusions cannot be generalized for pure narrowband or wideband experiments, although some similar effects, on a smaller scale, can be expected.

It was observed that the subjective scores vary expectedly and consistently when degradation types are considered separately. For example, the increase of packet loss or background noise would lower the subjective scores. This indicates that the ACR method is likely to provide repeatable assessments of super-wideband signals when the number of degradation types is limited.

In case of a wider range of degradation types (as it is the case for super-wideband full-case experiments), it was demonstrated that the remaining conditions contained in the test influence the results. Both the frequency and the intensity of a given degradation dimension in a test can have a strong impact on subjective scoring. One possible consequence for future tests could be the use of stronger requirements with respect to the frequency and intensity distribution of a given degradation dimension, especially regarding the distribution of background noise conditions. These enhanced constraints should help reduce the variability across future full-scale experiments.

Part of the variability in the subjective scores is most likely also due to the lack of guidance given to the subjects. The visibility of technical equipment in labs and the scoring of samples via headphones may decrease the impression of being in a "telephony situation". In this context, low levels and noises might not be seen as being caused by the channel (telephony system), and therefore not be considered as distortions. More detailed instructions on how to assess the telephony situation may help subjects deal with the new, unknown application that is super-wideband speech.

The lack of subject guidance could also be addressed with a different subjective test methodology. An approach like that of [b-ITU-T P.835], querying opinions for specific attributes, may help subjects to consider the different degradation types in their overall speech quality rating. A comparison approach, where the source signal is presented to the subjects such as the [ITUT P.800] DCR or [b-ITU-R BS.1116] methodologies, may also increase subject awareness of the presence of distortions.

The rank order variation of different degradation types cannot be predicted by objective models. This issue is therefore challenging as models must be trained with this variability in mind. Nevertheless, some full-scale databases are necessary to examine the (average) dependency between individual degradation dimensions. Consequently, the evaluation of objective models on full-scale super-wideband databases should be treated with caution and the models' performance is expected to be lower than on more traditional subjective tests.

Appendix III

Prediction of acoustically recorded narrowband speech

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

III.1 Background

Recommendation ITU-T P.863 is specified for the prediction of listening quality of acoustically recorded speech data in a super-wideband context only. That means the reference signal in this context is always a super-wideband speech signal and ITU-T P.863 is used in super-wideband operational mode.

This appendix advises how ITU-T P.863 can be used for the prediction of listening quality of acoustically recorded speech data in a narrowband context. Narrowband context means that the reference signal is narrowband. The prediction is using a narrowband scale and ITU-T P.863 predicts as a listener in a pure narrowband listening test.

Therefore no modifications to ITU-T P.863 are required.

III.2 Requirements for acoustically recorded speech data to be assessed by ITU-T P.863

Besides the common rules for speech signals and acoustical recordings as described in ITU-T P.863, the prediction of listening quality of acoustically recorded speech in a narrowband context is restricted to the following items.

- Recordings that are close to the ear, e.g., using handsets and headphones.
- Low variation compared to a nominal level in recording and presentation level. The nominal level stands for 79 dB(A) SPL in monotic recording/presentation and 73 dB(A) SPL in diotic recording/presentation.
- The reference signal to ITU-T P.863 must be flat filtered. No IRS send characteristic must be applied to the reference signal.
- It is recommended to apply the DC-removal filter as described in Annex C of [ITU-T P.501] to any speech signal used before applying the speech signal to ITU-T P.863.

ITU-T P.863 is not recommended for loudspeaker recordings or other recordings with considerable lower levels than the nominal level. ITU-T P.863 is applied to one ear signal only; binaural effects are not taken into account.

III.3 Pre-processing of speech and use of ITU-T P.863

It is recommended to reduce the sampling frequency of the flat reference signal and the test signal to 8 kHz to ensure audio narrowband and resample the signal afterwards to 48 kHz as required for super-wideband mode. In addition, the digital level of both signals should be to –26 dB OVL SPL according to [ITU-T P.56] independent of whether the signal was recorded monotically or diotically. These steps have to be done in a pre-processing step and are not an integral part of ITU-T P.863. At best, the reference signal is directly gained from a flat super-wideband signal by down-sampling and level readjustment.

Even though in this application reference signals are used, which have audio narrowband, ITU-T P.863 itself has to be used in super-wideband operational mode. In this mode the ITU-T P.863 internal IRS receive filter characteristic is not used; instead a flat input filter is applied as required for acoustically recorded data.

III.4 Interpretation of results

The outcome of ITU-T P.863 are mean opinion score (MOS) predictions without further mapping. The predicted MOS values are directly given on a one to five point scale. Experiments on test data have not shown a systematic bias or different interpretation of the scale. On average across the experiments, a good approximation of results was reached.

The MOS-LQO can be interpreted as a prediction of listening quality as it would be perceived in a narrowband Listening-Only-Test with monotic or diotic presentation on the nominal level.

III.5 Example results

Three narrowband experiments with acoustically recorded speech material were evaluated. The three experiments were provided by Deutsche Telekom in German.

The experiments PAAM_1 and PAAM_2 have been conducted for the PAAM project in 2002/2003. They consist of recordings using an ITU-T P.79 type 3.4 ear coupler. In both tests two types of randomly chosen off-the-shelf handset-phones and a headphone have been used as devices in receiving direction, the test conditions apply a set of typical codecs (PAAM_1) and high-/low-pass conditions (PAAM_2) in front of the device. The purpose of the experiments was not the evaluation and comparison of different handsets rather the evaluation of a reproducible rank-order of the test conditions independent from the used handset when acoustically recorded. Both experiments are based on simulated transmission, only the receiving and playing handset was a physical device.

The third experiment (NB_1, 2001) is based mostly on recordings using a good standard headset (headset 1) that was used as listening device over a variety of real VoIP connections on a laptop-PC. In contrast a few conditions have been produced by using a low-cost headset (headset 2) and a PC loudspeaker/desktop microphone. At sending side, an artificial mouth was used along with a headset microphone, wireless handset devices and ordinarily shaped handsets.

All three experiments were conducted according to [ITU-T P.800] with a naïve listening panel.

	DTAG PAAM_1	**DTAG PAAM_2**	**DTAG NB_1**
Pearson correlation	0.98	0.97	0.95
rmse* (raw)	0.17	0.27	0.16
rmse* (1^{st} order mapping)	0.02	0.04	0.12

The experiments are predicted with good accuracy. However, the experiments reflecting transmission technologies as used in 2002/2003 and include only a few different acoustical devices.

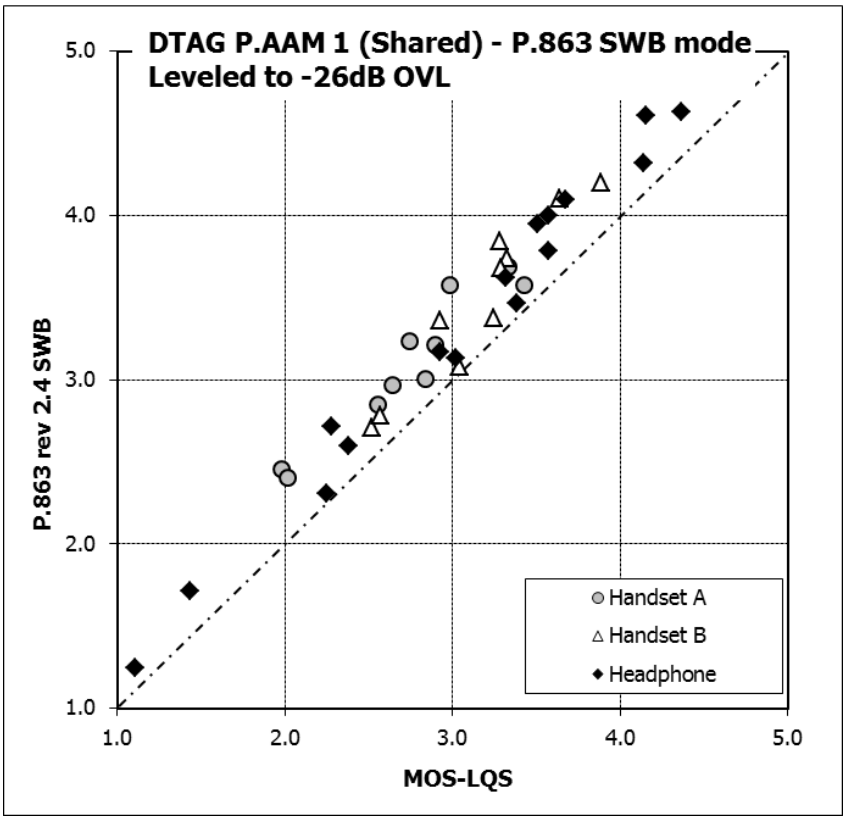


Figure III.1 –Results for experiment 1

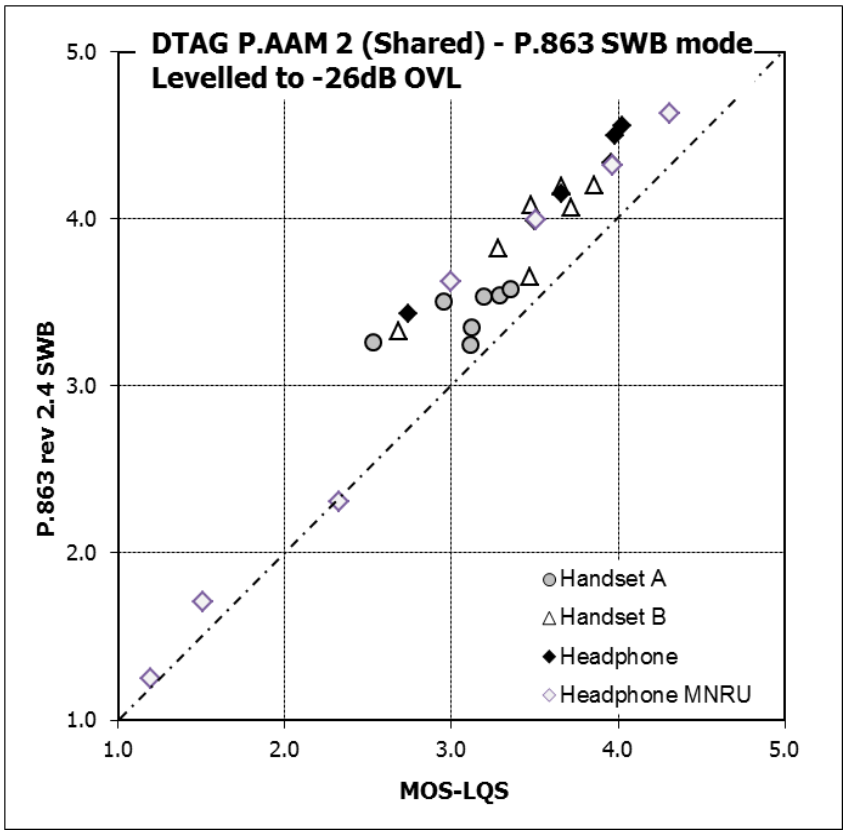


Figure III.2 –Results for experiment 2

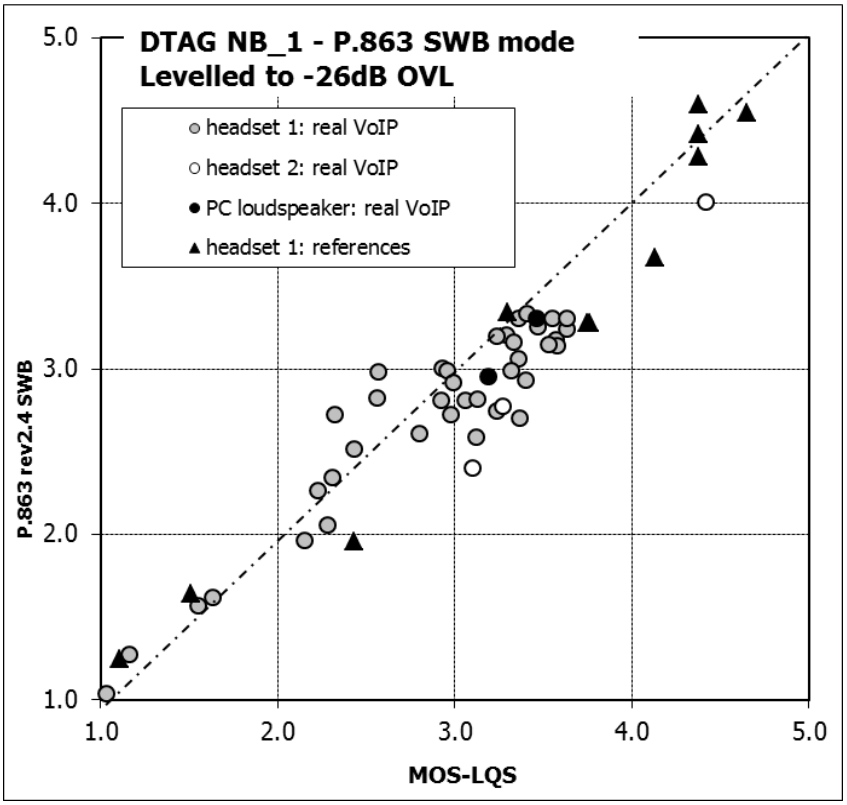


Figure III.3 – Results for experiment 3

Bibliography

[b-ITU-T P.79] Recommendation ITU-T P.79 (2007), *Calculation of loudness ratings for telephone sets*

[b-ITU-T P.835] Recommendation ITU-T P.835 (2003), *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*.

[b-ITU-T P.861] Recommendation ITU-T P.861 (1998), *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*.

[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-ITU-T P.862A2] Recommendation ITU-T P.862 Amendment 2 (2005), *Revised Annex A – Reference implementations and conformance testing for ITU-T Recs P.862, P.862.1 and P.862.2*.

[b-ITU-T P.862.1] Recommendation ITU-T P.862.1 (2003), *Mapping function for transforming P.862 raw result scores to MOS-LQO*.

[b-ITU-T P.862.2] Recommendation ITU-T P.862.2 (2007), *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*.

[b-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*.

[b-ITU-T P.863.1] Recommendation ITU-T P.863.1 (2014), *Application guide for Recommendation ITU-T P.863.*

[b-ITU-R BS.1116] Recommendation BS.1116-2 (2014), *Methods for the subjective assessment of small impairments in audio systems*.

[b-Barkowsky] Barkowsky, M., Bialkowski, J., Bitto, R., and Kaup, A. (2007), *Temporal registration using 3D phase correlation and a maximum likelihood approach in the perceptual evaluation of video quality*, IEEE 9th Workshop on Multimedia Signal Processing, pp. 195-198.

[b-Beerends 1989] Beerends, J.G. (1989), *Pitches of simultaneous complex tones, Chapter 5: A stochastic subharmonic pitch model*, Ph.D. dissertation, Technical University of Eindhoven, April 1989. < http://alexandria.tue.nl/extra3/proefschrift/PRF6B/8903333.pdf>

[b-Beerends 1992] Beerends, J.G. and Stemerdink, J.A. (1992), *A Perceptual Audio Quality Measure based on a psychoacoustic sound representation*, Journal of the Audio Engineering Society, vol. 40, December, pp. 963-978.

[b-Beerends 1994] Beerends, J.G. and Stemerdink, J.A. (1994), *A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation*, Journal of the Audio Engineering Society, vol. 42, No. 3, March, pp. 115-123.

[b-Beerends 2002] Beerends, J.G., Hekstra, A.P., Rix, A.W. and Hollier, M.P. (2002), *Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment Part II: Psychoacoustic Model*, Journal of the Audio Engineering Society, vol. 50, October, pp. 765-778.

[b-Beerends 2007] Beerends, J.G., Busz, B.P., Oudshoorn, P., Van Vugt, J.M., Ahmed, O.K. and Niamut, O.A. (2007), *Degradation Decomposition of the Perceived Quality of Speech Signals on the Basis of a Perceptual Modelling Approach*, Journal of the Audio Engineering Society, vol. 55, December, pp. 1059-1076.

[b-Coalition] The POLQA Coalition: OPTICOM GmbH, Erlangen, Germany; SwissQual AG, Solothurn, Switzerland; TNO Telecom, Delft, The Netherlands, POLQA – Perceptual Objective Listening Quality Analysis, Technical White Paper, July 2010.
< http://www.polqa.info/download/registration_form.php>

[b-Goh] Goh, C., Hamadicharef, B., Henderson, G.T., and Ifeachor, E.C. (2005), *Comparison of fractal dimension algorithms for the computation of EEG biomarkers for dementia*, Proceedings of the 2nd International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2005), Lisbon, Portugal.

[b-Rix] Rix, A.W., Hollier, M.P., Hekstra, A.P., and Beerends, J.G. (2002), *Perceptual evaluation of speech quality (PESQ), The New ITU Standard for objective measurement of perceived speech quality, Part I – Time alignment*, Journal of the Audio Engineering Society, vol. 50, October, pp. 755-764.

[b-SwissQual] SwissQual AG, *Transition to SQuad08 and Wideband Voice Tests, Technical White Paper*, April 2009. <email: info@swissqual.com>

[b-Zwicker] Zwicker, E. and Feldtkeller, R. (1967), *Das Ohr als Nachrichtenempfänger*, S. Hirzel Verlag, Stuttgart.

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
Series P	**Terminals and subjective and objective assessment methods**
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

Printed in Switzerland
Geneva, 2015