

INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.563

(05/2004)

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

Objective measuring apparatus

**Single-ended method for objective speech quality
assessment in narrow-band telephony applications**

Notice to Subscribers of ITU-T Recommendation P.563:

**To obtain one complimentary CD-ROM containing ITU-T
Recommendation P.563 and the Reference Implementation and
Conformance Data, please contact IHS/Retail at 1-800-854-7179.**



INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.563

(05/2004)

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

Objective measuring apparatus

Single-ended method for objective speech quality assessment in narrow-band telephony applications

ITU-T Recommendation P.563

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Subscribers' lines and sets	Series	P.30 P.300
Transmission standards	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 quality	Series	P.80 P.800
Audiovisual quality in multimedia services	Series	P.900

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

ITU-T Recommendation P.563

Single-ended method for objective speech quality assessment in narrow-band telephony applications

Summary

This Recommendation describes an objective single-ended method for predicting the subjective quality of 3.1 kHz (narrow-band) telephony applications. This Recommendation presents a high-level description of the method and advice on how to use it. An ANSI-C reference implementation, described in Annex A, is provided in separate files and forms an integral part of this Recommendation. 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 ANSI-C reference implementation shall take precedence in case of conflicts between the high-level description as given in this Recommendation and the ANSI-C reference implementation.

This Recommendation includes an electronic attachment containing an ANSI-C reference implementation and conformance testing data.

Source

ITU-T Recommendation P.563 was approved on 14 May 2004 by ITU-T Study Group 12 (2001-2004) under the ITU-T Recommendation A.8 procedure.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications. 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, implementors are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database.

© ITU 2005

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

CONTENTS

Page

1	Introduction	1
2	Normative references.....	1
3	Abbreviations.....	2
4	Scope	2
5	Convention.....	5
6	Requirements on speech signals to be assessed.....	5
7	Overview of P.563	6
7.1	Vocal tract analysis and unnaturalness of speech.....	7
7.2	Analysis of strong additional noise	8
7.3	Interruptions, mutes and time clipping.....	8
7.4	Distortion classification.....	8
8	Comparison between objective and subjective scores.....	9
8.1	Correlation coefficient.....	10
9	High level description of the functional blocks used in P.563	10
9.1	Description of basic speech descriptors and the signal pre-processing.....	10
9.2	Description of the functional block 'Vocal tract analysis and Unnatural Voice'.....	18
9.3	Description of the functional block 'Additive Noise'	32
9.4	Description of the 'Mutes/Interruptions' functional block components.....	42
9.5	Description of the Speech Quality Model	46
Annex A	– Source code for reference implementation and conformance tests.....	49
A.1	List of files provided for the ANSI-C reference implementation.....	49
A.2	List of files provided for conformance validation	50
A.3	Speech files provided for validation with variable delay	57
A.4	Conformance data sets.....	57
A.5	Conformance requirements	57
A.6	Conformance test on unknown data	57

Electronic attachment: ANSI-C reference implementation and conformance testing data.



ITU-T Recommendation P.563

Single-ended method for objective speech quality assessment in narrow-band telephony applications¹

1 Introduction

The P.563 algorithm is applicable for speech quality predictions without a separate reference signal. For this reason, this method is recommended for non-intrusive speech quality assessment, live network monitoring and assessment by using unknown speech sources at the far-end side of a telephone connection.

Real systems may include background noise, filtering and variable delay, as well as distortions due to channel errors and speech codecs. Up to now, methods for speech quality assessment of such systems, such as ITU-T Rec. P.862, require either a reference signal or they calculate only quality indexes based on a restricted set of parameters like level, noise in speech pauses and echoes.

The P.563 approach is the first recommended method for single-ended non-intrusive measurement applications that takes into account the full range of distortions occurring in public switched telephone networks and that is able to predict the speech quality on a perception-based scale MOS-LQO according to ITU-T Rec. P.800.1. This Recommendation is not restricted to end-to-end measurements; it can be used at any arbitrary location in the transmission chain. The calculated score is then comparable to the quality perceived by a human listener, who is listening with a conventional shaped handset at this point.

The validation of P.563 included all available experiments from the former P.862 validation process, as well as a number of experiments that specifically tested its performance by using an acoustical interface in a real terminal at the sending side. Furthermore, the P.563 algorithm was tested independently with unknown speech material by third party laboratories under strictly defined requirements.

It is recommended that P.563 be used for speech quality assessment in 3.1 kHz (narrow-band) telephony applications only.

2 Normative 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 Recommendation P.48 (1988), *Specification for an intermediate reference system*.
- ITU-T Recommendation P.800 (1996), *Methods for subjective determination of transmission quality*.
- ITU-T Recommendation P.810 (1996), *Modulated noise reference unit (MNRU)*.

¹ This Recommendation includes an electronic attachment containing an ANSI-C reference implementation and conformance testing data.

- ITU-T Recommendation P.830 (1996), *Subjective performance assessment of telephone-band and wideband digital codecs*.
- ITU-T Recommendation 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*.
- ITU-T P-series Recommendations – Supplement 23 (1998), *ITU-T coded-speech database*.

3 Abbreviations

This Recommendation uses the following abbreviations:

ACR	Absolute Category Rating
CELP	Code-Excited Linear Prediction
dBov	dB to overload point
DCME	Digital Circuit Multiplication Equipment
ERP	Ear Reference Point
HATS	Head and Torso Simulator
IRS	Intermediate Reference System
LPC	Linear Prediction Coefficient
MOS	Mean Opinion Score
MOS-LQO	Mean Opinion Score – Listening Quality Objective
MOS-LQS	Mean Opinion Score – Listening Quality Subjective
PCM	Pulse Code Modulation
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level

4 Scope

Based on the benchmark results presented within Study Group 12 in 2003, an overview of the test factors, coding technologies and applications to which this Recommendation applies is given in Tables 1 to 3. Table 1 presents the relationships of test factors, coding technologies and applications for which this Recommendation has been found to show acceptable accuracy. Table 2 presents a list of conditions for which the Recommendation is known to provide inaccurate predictions or is otherwise not intended to be used. Finally, Table 3 lists factors, technologies and applications for which P.563 has not currently been validated. Although correlations between objective and subjective scores in the benchmark were around 0.89 for both known and unknown data, the P.563 algorithm cannot be used to replace subjective testing but it can be applied for measurements where auditory tests would be too expensive or not applicable at all.

It should also be noted that the P.563 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 in the same way as it can be investigated by an auditory test assessing listening quality on an ACR scale. The P.563 algorithm scores the speech signal in that way, as it is presented to a human listener by using a conventional shaped handset and listening with a SPL of 79 dB at the ERP.

Because P.563 models the human quality perception in combination with a common receiving terminal, the degradation produced by a receiving terminal and other equipment in a real monitored connection, which are connected behind the measurement point, cannot be taken into account.

Because P.563 predicts listening quality scores, all effects degrading talking quality or conversational quality only cannot be taken into account. That means, the effects of loudness loss, delay, sidetone, talker-echo, and other impairments related to the talking quality or two-way interaction only are not reflected in the P.563 scores. Therefore, it is possible to have high P.563 scores, yet non-optimal quality of the connection overall.

It should be highlighted that P.563 is designed for the prediction of speech quality in public switched narrow-band telephone networks. The types and the amount of the distortions, technologies and applications in the validation procedure cover the range of common occurrences in such networks. Extreme situations, even if they fulfil the terms of Table 1, may be predicted inaccurately.

Table 1/P.563 – Factors for which P.563 had demonstrated acceptable accuracy and recommended application scenarios

Test factors
Characteristics of the acoustical environment (reflections, different reverberation times) as used in the validation phase. Mobile and conventional shaped handsets as well as handsfree terminals according to P.340 test-setup in office environments were used (See Note).
Environmental noise at the sending side
Characteristics of the acoustical interface of the sending terminal
Remaining electrical and encoding characteristics of the sending terminal
Speech input levels to a codec
Transmission channel errors
Packet loss and packet loss concealment with CELP codecs
Bit rates if a codec has more than one bit-rate mode
Transcodings
Effect of varying delay on listening quality in ACR tests
Short-term time warping of speech signal
Long-term time warping of speech signal
Transmission systems including echo cancellers and noise reduction systems under single talk conditions and as they will be scored on an ACR scale
Coding technologies
Waveform codecs, e.g., G.711; G.726; G.727
CELP and hybrid codecs ≥ 4 kbit/s, e.g., G.728, G.729, G.723.1
Other codecs: GSM-FR, GSM-HR, GSM-EFR, GSM-AMR, CDMA-EVRC, TDMA-ACELP, TDMA-VSELP, TETRA

Table 1/P.563 – Factors for which P.563 had demonstrated acceptable accuracy and recommended application scenarios

Recommended application scenarios for P.563
Live network monitoring using digital or analogue connection to the network
Live network end-to-end testing using digital or analogue connection to the network
Live network end-to-end testing with unknown speech sources at the far end side
NOTE – For more detailed information, please refer to the published testplans ("Joint Test Plan for Single-Ended Assessment Models", COM 12-D 121, January 2003).

Table 2/P.563 — P.563 is known to provide inaccurate predictions when used in conjunction with these variables, or is otherwise not intended to be used with these variables

Test factors
Listening levels, Loudness loss (See Note.)
Sidetone
Effect of delay in conversational tests
Talker echo
Music or network tones as input signal
Coding technologies
LPC vocoder technologies at bit rates < 4.0 kbit/s, e.g., IMBE, AMBE, LPC10e
Applications
Predicting talking quality
Two-way communication performance
NOTE – P.563 assumes a standard listening level of 79 dB SPL and compensates for non-optimum signal levels in the input files. The subjective effect of deviation from optimum listening level is therefore not taken into account.

Table 3/P.563 – Factors, technologies and applications for which P.563 has not or not fully been validated at the time of the standardization

Test factors
Amplitude clipping of speech (was not included in the evaluation data)
Talker dependencies and multiple simultaneous talkers
Singing voice and child's voice as input to a codec
Bit-rate mismatching between an encoder and a decoder if a codec has more than one bit-rate mode
Artificial speech signals as input to a codec
Listener echo
Effects/artifacts from isolated echo cancellers
Effects/artifacts from isolated noise reduction algorithms
Evaluation of synthetic speech and/or using it as input to a speech codec

Table 3/P.563 – Factors, technologies and applications for which P.563 has not or not fully been validated at the time of the standardization

Coding technologies
CELP and hybrid codecs < 4 kbit/s
MPEG-4 HVXC
Applications
Measurements at the acoustic interface of the receiving terminal/handset, e.g., using HATS

5 Convention

Subjective evaluation of telephone networks and speech codecs may be conducted using listening-only or conversational methods of subjective testing. For practical reasons, listening-only tests are the only feasible method of subjective testing during the development of speech codecs, when a real-time implementation of the codec is not available. Also listening-only tests are often not practicable for live network monitoring. This Recommendation discusses an objective measurement technique for estimating subjective quality obtained in listening-only tests, using listening equipment conforming to the IRS or modified IRS receive characteristics.

The P.563 approach predicts the results of ACR listening quality (LQS) subjective experiments by calculating a listening quality value (LQO) using the common MOS scale from 1 to 5. This Recommendation should, therefore, be considered to relate primarily to the ACR LQ opinion scale.

6 Requirements on speech signals to be assessed

The described algorithm is designed for evaluating human speech only. It cannot be used for the evaluation of music, noise or other non-speech audio signals. The applicability if singing voice is transmitted over telephone connections has not yet been validated.

The speech signal to be assessed has to be recorded at an 'electrical' interface. That means, recordings made by an artificial ear in the acoustical domain cannot be used. Furthermore, outcomes simulation of speech transmissions or other speech processing can be used if they are covered by the scope given in Table 1 and do not include a terminal simulation.

The digitized speech signal has to fulfil the following requirements:

- Sampling frequency: 8000 Hz
If higher frequencies are used for recording, a separate down-sampling by using a high quality flat low pass filter has to be applied. Lower sampling frequencies are not allowed.
- Amplitude resolution: 16 bit linear PCM
- Minimum active speech in file: 3.0 s
- Maximum signal length: 20.0 s
- Minimum speech activity ratio: 25%
- Maximum speech activity ratio: 75%
- Range of active speech level: –36.0 to –16.0 dBov
A level adjustment to –26 dBov is part of P.563. The recommended level limitation should avoid additional artefacts by low SNR or amplitude clipping respectively.

7 Overview of P.563

In comparison to P.862 (a so-called 'double-ended' or 'intrusive' method) that compares a high quality reference signal to the degraded signal on a basis of a perceptual model, P.563 predicts the speech quality of a degraded signal without a given reference speech signal. Figure 1 illustrates the differences in these approaches.

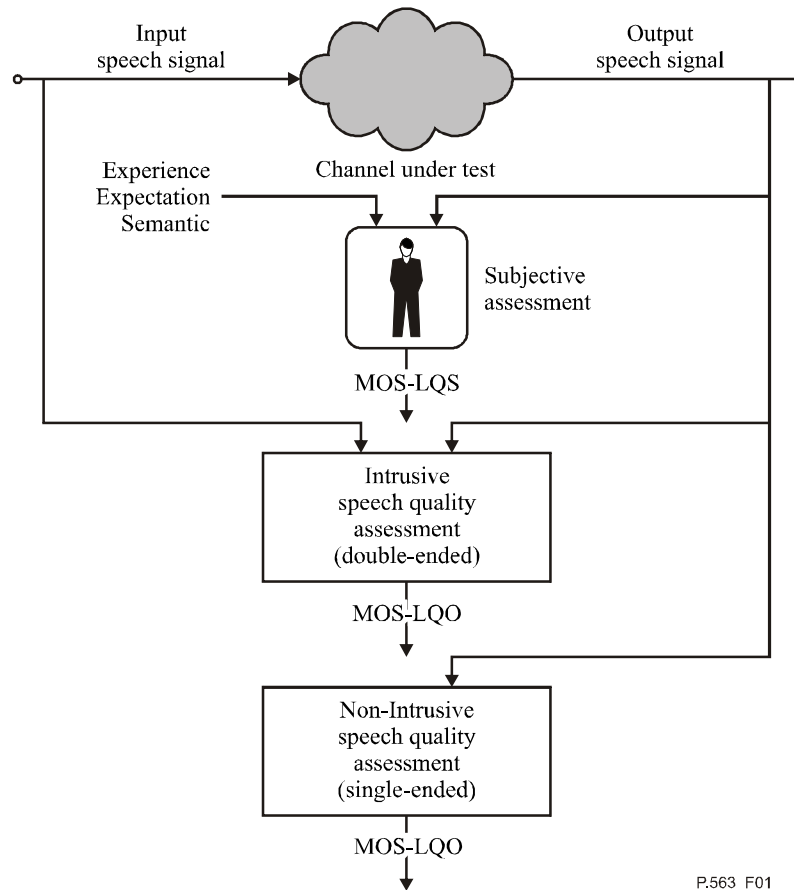


Figure 1/P.563 – Non-Intrusive versus Intrusive models

The P.563 approach could be visualized as an expert who is listening to a real call with a test device like a conventional handset into the line in parallel. This visualization explains also the main application and allows the user to rate the scores gained by P.563. The quality score predicted by P.563 is related to the perceived quality by linking a conventional handset at the measuring point.

Consequently, the listening device has to be part of the P.563 approach. Therefore, each signal will first be pre-processed. This pre-processing begins with the model of the receiving handset. Following this, a voice activity detector (VAD) is used to identify portions of the signal that contain speech and the speech level is calculated. Finally, a speech level adjustment to –26 dBov is applied. The pre-processed speech signal to be assessed will be investigated by several separate analyses, which detect, like a sensor layer, a set of characterizing **signal parameters**. This analysis will be applied at first to all signals. Based on a restricted set of **key parameters**, an assignment to a **main distortion** class will be made.

The key parameters and the assigned distortion class are used for the adjustment of the speech quality model. This provides a perceptual based weighting where several distortions are occurring

in one signal but one distortion class is more prominent than the others. The basic block-scheme of P.563 is shown in Figure 2.

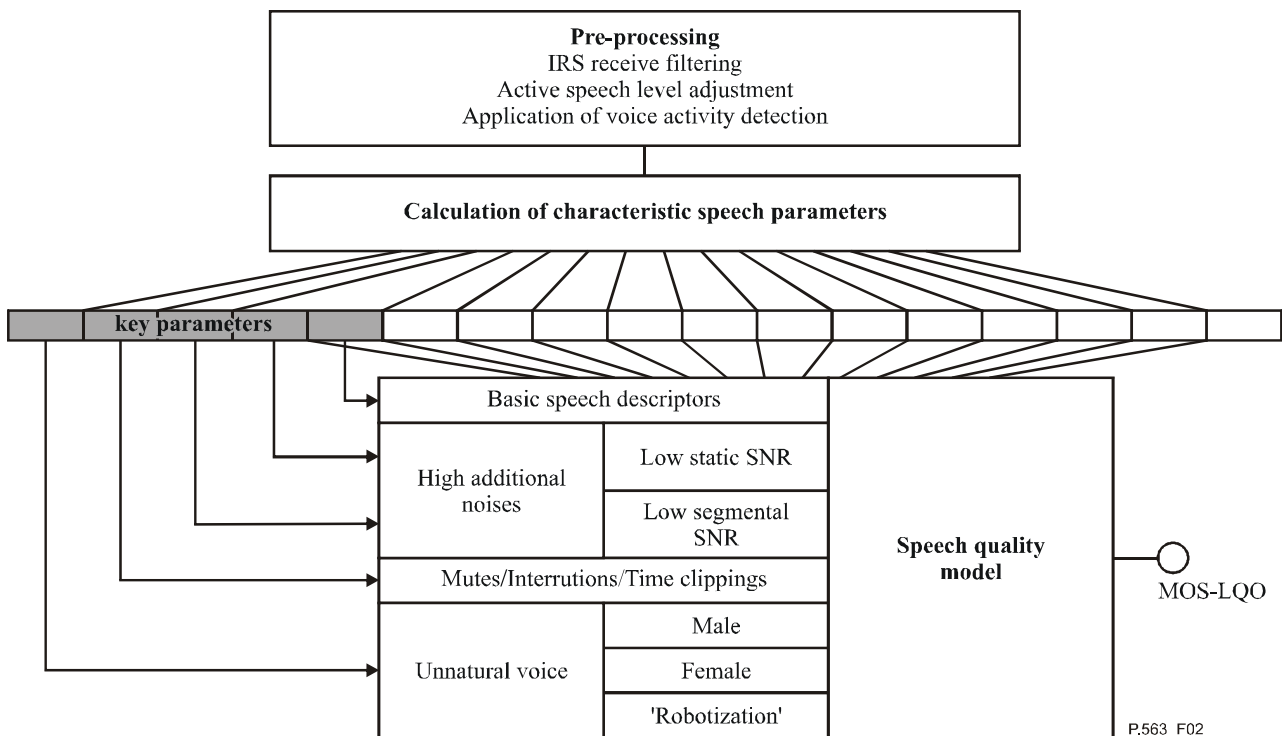


Figure 2/P.563 – Block scheme of P.563

Basically, the P.563 algorithm's signal parameterization can be divided into three independent functional blocks that correspond to the main classes of distortion:

- **Vocal tract analysis and unnaturalness of speech**
 - 1) Male voices;
 - 2) Female voices;
 - 3) Strong 'Robotization'.
- **Analysis of strong additional noise**
 - 1) Low static SNR (Background noise floor);
 - 2) Low segmental SNR (Noise that is related to the signal envelope).
- **Interruptions, mutes and time clipping**

In addition, a set of basic speech descriptors like active speech level, speech activity and level variations will be used, mainly for adjusting the pre-processing and the VAD. Some of the signal parameters calculated within the pre-processing stage will be used in these three functional blocks.

7.1 Vocal tract analysis and unnaturalness of speech

The main block looks for unnaturalness in the speech signal. This functional block contains a speech production model for extracting signal parts that could be interpreted as voice and separates them from the non-speech parts. Furthermore, high order statistical analysis gives additional information about how human-like the speech is.

The unnaturalness of speech will be rated separately for male and female voices. Furthermore, in the case of strong robotization², another separate rating is made, which is gender-independent.

In this clause the signal is investigated for the occurrence of tones like DTMF-tones or similar highly periodic signals that are not speech.

Other very annoying disturbances are repeated speech frames. In packet-based transmission systems, a typical error that can occur is the loss of packets. Some speech codecs employ error concealment methods in order to increase the received speech quality. In fact, some error concealment methods use packet (frame) repetitions that simply replace a lost packet by, for example, a previously successfully transmitted packet, and tend to decrease the quality of the signal rather than to increase it.

A more general description of the received speech quality is given by comparing the input signal with a pseudo reference signal generated by a speech enhancer.

7.2 Analysis of strong additional noise

The noise analysis calculates different characteristics of noise. Based on two key parameters, the decision will be made if additional noise is the main degradation. If additional noise is detected as the main degradation class, a decision is made for the type of noise. Either it is static and present over all the signal (at least during speech activity) such that the noise power is not correlated with the speech signal, or the noise power shows dependencies on the signal power envelope.

If there was noise found that is likely to be static, several detectors try to quantify the amount of noise 'locally' and 'globally'. The expression 'local' noise as it is used here, describes the signal parts found especially between phonemes, whereas 'global' noise was defined as the signal between utterances such as sentences. Distinguishing between those noise types is important as, for example, in mobile communications often different settings for speech active parts and non-active parts are applied, e.g., introduction of comfort noise.

7.3 Interruptions, mutes and time clipping

Mutes and interruptions also form a separate distortion class. Such distortions can only partly be described by outcomes of the vocal tract investigation. Hence, a separate analysis is made to detect and to rate time clippings and unnatural mutes in the signal.

Signal interruption can occur in two variants i.e., as temporal speech clipping or speech interruption. Both lead to a loss of signal information.

Temporal clipping may occur when voice activity detection is used, when DCME is used or the signal becomes interrupted. This clipping is an annoying phenomenon that cuts off a bit of speech in the instant it takes for the transmitter to detect presence of speech. It is possible to detect the interruptions of the speech signal, which occur during the active speech intervals. The algorithms used in P.563 are able to distinguish between normal word ends and abnormal signal interruptions as well as unnatural silence intervals in a speech utterance.

7.4 Distortion classification

Table 4 gives an overview for all calculated signal parameters. The key parameters that are used for classification of the main distortions are highlighted in grey.

The table also reflects the structure of the rest of this Recommendation. The main columns are in line with the sections or the corresponding main distortion classes respectively.

² Robotization is caused by a voice signal that contains too much periodicity.

All the names used for the parameters can be found and are explained in later sections of this Recommendation as well as in the example source code.

Table 4/P.563 – Overview for all used signal parameters in P.563

Basic speech descriptors	Unnatural speech		Noise analysis		Interruptions/ mutes
	Vocal tract analysis	Speech statistics	Static SNR	Segmental SNR	
<i>PitchAverage</i>	<i>Robotization</i>	<i>LPCcurt</i>	<i>SNR</i>	<i>EstSegSNR</i>	<i>SpeechInterruptions</i>
<i>SpeechSectionLevelVar</i>	<i>ConsistentArtTracker</i>	<i>LPCskew</i>	<i>EstBGNoise</i>	<i>SpecLevelDev</i>	<i>SharpDeclines</i>
<i>SpeechLevel</i>	<i>VTPMaxTubeSection</i>	<i>LPCskewAbs</i>	<i>NoiseLevel</i>	<i>SpecLevelRange</i>	<i>MuteLength</i>
<i>LocalLevelVar</i>	<i>FinalVtpAverage</i>	<i>CepCurt</i>	<i>HiFreqVar</i>	<i>RelNoiseFloor</i>	<i>UnnaturalSilence</i>
	<i>VTPPeakTracker</i>	<i>CepSkew</i>	<i>SpectralClarity</i>		<i>UnnaturalSilenceMean</i>
	<i>ArtAverage</i>	<i>CepADev</i>	<i>GlobalBGNoise</i>		<i>UnnaturalSilenceTotEnergy</i>
	<i>VtpVadOverlap</i>		<i>GlobalBGNoiseTotEnergy</i>		
	<i>PitchCrossCorrLOffset</i>		<i>GlobalBGNoiseRelEnergy</i>		
	<i>PitchCrossPower</i>		<i>GlobalBGNoiseAffectedSamples</i>		
	<i>BasicVoiceQuality</i>		<i>LocalBGNoiseLog</i>		
	<i>BasicVoiceQualityAsym</i>		<i>LocalBGNoiseMean</i>		
	<i>BasicVoiceQualitySym</i>		<i>LocalBGNoiseStddev</i>		
	<i>FrameRepeats</i>		<i>LocalBGNoise</i>		
	<i>FrameRepeatsTotEnergy</i>		<i>LocalBGNoiseAffectedSamples</i>		
	<i>UnnaturalBeeps</i>				
	<i>UnnaturalBeepsMean</i>				
	<i>UnnaturalBeepsAffectedSamples</i>				

8 Comparison between objective and subjective scores

Subjective votes are influenced by many factors such as the preferences of individual subjects and the context (the other conditions) of the experiment. Thus, a regression process is necessary before a direct comparison can be made. The regression must be monotonic so that information is preserved, and it is normally used to map the objective P.563 score onto the subjective score. A

good objective quality measure should have a high correlation with many different subjective experiments if this regression is performed separately for each one and, in practice, with P.563, the regression mapping is often almost linear, using a MOS-like scale.

A preferred regression method for calculating the correlation between the P.563 score and subjective MOS, which was used in the validation of P.563, uses a 3rd-order polynomial constrained to be monotonic. This calculation is performed on a per study basis. In most cases, condition MOS is the chosen performance metric, so the regression should be performed between condition MOS and condition-averaged P.563 scores. A condition should use at least four different speech samples. The result of the regression is a set of objective MOS scores for that test. In order to be able to compare objective and subjective scores the subjective MOS scores should be derived from a listening test that is carried out according to ITU-T Rec. P.830.

8.1 Correlation coefficient

The closeness of the fit between P.563 and the subjective scores may be measured by calculating the correlation coefficient. Normally, this is performed on condition-averaged scores, after mapping the objective to the subjective scores. The correlation coefficient is calculated with Pearson's formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

In this formula, x_i is the subjective condition MOS for condition i , and $\bar{x}$ is the average over the subjective condition MOS values, x_i . y_i is the mapped condition-averaged P.563 score for condition i , and $\bar{y}$ is the average over the condition-averaged P.563 MOS values y_i .

For 24 known ITU benchmark experiments, the average correlation was 0.88. For an agreed set of six experiments used in the independent validation, experiments that were unknown during the development of P.563, the average correlation was 0.90.

9 High level description of the functional blocks used in P.563

This clause explains the functional blocks used in P.563 and shown in Figure 2.

9.1 Description of basic speech descriptors and the signal pre-processing

9.1.1 Voice activity detection

The Voice Activity Detection (VAD) algorithm is based on an adaptive power threshold, using an iterative approach. Envelope frames above this threshold are classified as speech, and below as noise.

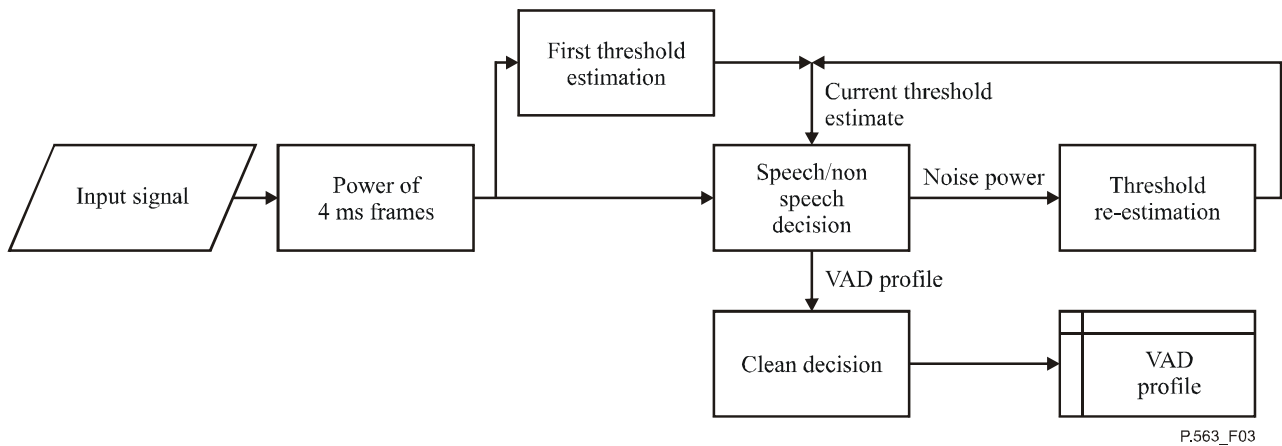


Figure 3/P.563 – Voice activity detection overview

As shown in Figure 3, the VAD threshold is initially set to the mean power of the signal. This initialization allows the first speech/noise decision to be made. The threshold is then iteratively re-estimated to be the mean plus two standard deviations of the power of the noise sections over the whole signal. Twelve iterations are completed, after which the final threshold is determined and used to make the final speech/noise decision.

$$\begin{cases} m_x = \sum_{i=0}^{n-1} \frac{x_{noise}(i)}{n} \\ th_{est} = m_x + 2 \times \sqrt{\sum_{i=0}^{n-1} \frac{(x_{noise}(i) - m_x)^2}{n}} \end{cases}$$

Additional processing is performed to label short events:

- Events that are above threshold for 12 ms or less are labelled as non-speech;
- Speech utterances that are separated by less than 200 ms are joined together.

9.1.2 IRS filtering

It is assumed that the listening tests were carried out using an IRS or a modified IRS receive characteristic in the handset. For some of the parameters, this has to be taken into account.

In this Recommendation, this is done in 2 steps:

Firstly, the level normalization to –26 dB is carried out by estimating the signal level in the range of 250 Hz to about 3000 Hz.

In the second step, a filter is applied in the frequency domain using an FFT filter over the length of the file. The filter characteristic is similar to the IRS receive characteristic given in ITU-T Rec. P.830.

The IRS filtered signal is used in the sections Sharp Declines, Robotization, Frame Repeats, Unnatural Beeps, Basic Voice Quality, Unnatural Silence, Local Background Noise and Global Background Noise.

9.1.3 Signal normalization and level calculation

Before being analysed, the signal is normalized to –26 dBov and filtered.

To normalize to –26 dBov, an estimate of the speech level is calculated based on the output from the VAD by calculating the RMS value of each active frame.

Once the signal level has been normalized, it is filtered using a 4th order Butterworth high-pass filter at 100 Hz cut-off frequency.

All the parameters that do not use the IRS filtered signal use this normalized signal, except mutes parameters which require the raw signal.

9.1.4 Variation of the speech section level

SpeechSectionsLevelVar is a descriptor of the level variation between sentences. The level of each speech section in the signal is computed. The difference between the maximum and the minimum level is calculated.

9.1.5 Local level variation

LocalLevelVar describes the energy variation of the signal. The RMS is calculated frame by frame. The first derivative of the RMS array is then calculated and power averages are extracted.

9.1.6 Pitch synchronous extraction

Pitch period length estimation and synchronous pitch mark labelling is required by multiple parts of the P.563 algorithm for extraction of parameters. The pitch synchronous extraction uses a hybrid temporal/spectral approach. Figure 4 shows an overall description of the algorithm.

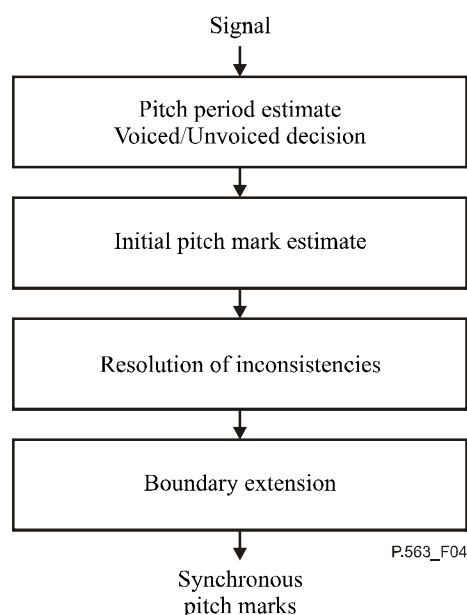


Figure 4/P.563 – Pitch synchronous extraction overview

9.1.6.1 Pitch period estimate and voice/unvoiced decision

Pitch period estimates are achieved through an autocorrelation method. The autocorrelation is calculated over 64 ms frames (512 elements), 50% overlapped.

The signal is first passed through a Hann window defined by:

$$y_i = 0.5x_i[1 - \cos(w)] \text{ for } i = 0 \dots (n-1)$$

where:

$$w = \frac{2\pi i}{n}$$

x is the input signal and

n is size of the frame

The autocorrelation is calculated using the inverse FFT of the power of the FFT:

$$R_{xx}(t) = y(t) \otimes y(t) = FFT^{-1}(FFT(y) \times FFT(y)^*)$$

R_{xx} is normalized by $R_{xx}(0)$, the maximum value of the calculated autocorrelation. The first 32 elements of the output array are filtered using half of a Hann window. The filtered autocorrelation is the n given by:

$$N_i = \frac{R_i}{R_0} \text{ for } i = 0, 1, 2, \dots, 31 \text{ and } N_i = \frac{R_i}{2R_0}[1 - \cos(w)] \text{ for } i = 32, 33, \dots, n-1$$

where:

$$w = \frac{2\pi i}{64}$$

The maximum value is then searched between the 20th and the 100th elements, which covers a pitch range from 150 to 780 Hz. If this maximum is greater than 0.5, then the frame is classified as voiced and the candidate pitch length (T_{max}) is saved.

To avoid pitch doubling, a comparison is made with the pitch length of the last frame marked as voiced (T_{old}). T_{old} is firstly readjusted to make sure its position is a local peak, then if the following equation is validated,

$$N_{T_{old}} > \frac{1 + N_{T_{max}}}{3}$$

T_{max} is considered to be doubled, and T_{old} is set as the pitch period of the current frame.

9.1.6.2 Pitch mark placement

Once the pitch period is estimated, pitch mark labelling is performed. The method used is the cross-correlation of windowed speech segments with an impulse train. It is performed frame-by-frame to find the optimum pitch mark locations. The algorithm seeks the offset that gives the maximum absolute cross-correlation.

The cross-correlation is defined by: $Rxy_i = x_i \otimes y_i$, with

$$x_i = 0.5z_i[1 - \cos(w)], \text{ for } i = 0 \dots n$$

where:

$$w = \frac{2\pi i}{n}$$

z is the input signal

n is the frame size

$y_i = 1$ when i is a multiple of the pitch period (l)

$y_i = 0$ everywhere else

As a result, the equation can be simplified:

$$Rxy_i = \sum_{j=0}^{n/l} x_{i+j \cdot l}$$

where:

i is the offset

l the pitch length

Once the calculation is completed, i_{max} is set to be the index of the maximum of Rxy . Pitch marks are placed within the frame at regular intervals separated by the pitch length l , offset by i_{max} .