International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.501

Amendment 1

(06/2022)

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

Objective measuring apparatus

Test signals for use in telephony and other speech-based applications

Amendment 1:
New Appendix II: ITU-T P.863 results on ITU-T P.501 and ITU-T P.565 speech samples

Recommendation ITU-T P.501 (2020) – Amendment 1

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
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 P.700
Methods for objective and subjective assessment of speech quality	Series	P.80
Methods for objective and subjective assessment of speech and video quality	Series	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 speech and video	Series	P.1500

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

Recommendation ITU-T P.501

Test signals for use in telephony and other speech-based applications

Amendment 1

New Appendix II: ITU-T P.863 results on ITU-T P.501 and ITU-T P.565 speech samples

Summary

Amendment 1 provides the Recommendation ITU-T P.863 Ed. 3 (04/2020) reference scores on typical codec conditions for the reference samples in Recommendation ITU-T P.501 Annexes C and D. The ITU-T P.863 scores on the same codec conditions are also reported for the speech reference sample provided as annexes to Recommendations ITU-T P.565 and ITU-T P.565.1.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.501	1996-08-30	12	11.1002/1000/3635
2.0	ITU-T P.501	2000-05-18	12	11.1002/1000/5080
2.1	ITU-T P.501 (2000) Amd. 1	2004-05-14	12	11.1002/1000/7411
3.0	ITU-T P.501	2007-06-29	12	11.1002/1000/9065
4.0	ITU-T P.501	2009-12-14	12	11.1002/1000/10657
5.0	ITU-T P.501	2012-01-13	12	11.1002/1000/11459
5.1	ITU-T P.501 (2012) Amd. 1	2012-07-14	12	11.1002/1000/11686
5.2	ITU-T P.501 (2012) Amd. 2	2014-10-29	12	11.1002/1000/12330
5.3	ITU-T P.501 (2012) Amd. 3	2015-06-29	12	11.1002/1000/12515
6.0	ITU-T P.501	2017-03-01	12	11.1002/1000/13173
6.1	ITU-T P.501 (2017) Amd. 1	2018-06-13	12	11.1002/1000/13623
7.0	ITU-T P.501	2020-05-29	12	11.1002/1000/14271
7.1	ITU-T P.501 (2020) Amd. 1	2022-06-17	12	11.1002/1000/14998

Keywords

Speech, speech-based applications, telephonometry, telephony, test signals.

* 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/software copyrights, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Page

Recommendation ITU-T P.501

Test signals for use in telephony and other speech-based applications

1) Appendix II

Add the following appendix after appendix I:

Appendix II

ITU-T P.863 results on ITU-T P.501 and ITU-T P.565 speech samples

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

II.1 Reference speech samples in ITU-T P.501 Annex C and Annex D

[ITU-T P.863] and [ITU-T P.863.1] recommend evaluating the quality of a system under test by use of multiple samples with different content, and covering male and female voices. The idea is to reduce sample and talker dependencies of speech quality assessments by averaging the individual speech quality measurements (similar to [ITU-T P.800] listening tests). Usually, four double sentences are used. However, the use of multiple files requires steady test conditions as can be maintained mainly under lab conditions. For fluctuating quality conditions as e.g., in real-field drive tests, shorter samples can be used and are recommended if they fulfil constraints regarding similar behaviour as an average of multiple samples. This is the case for the especially designed short speech samples in Annex D.

Annex C and Annex D provide speech samples prepared and proven for use with [ITU-T P.863].

Annex C contains sets of samples, where always four samples (double sentences spoken by one speaker) are spoken in one language. This structure of samples is in line with the ITU-T P.800 speech sample structure, and adapted thus to static testing as in a lab environment.

Annex D provides samples where always a male and female voice sentence is combined in one shorter 6-second long sample. These short samples in Annex D are especially composed to result in similar speech quality values for the main coding and band-limiting conditions as the four samples as in Annex C. Because of the short length, they are better adapted to non-static quality conditions, where speech quality prediction should be based on shorter evaluation periods. In addition, Annex D samples are balanced and prepared for [ITU-T P.863] use regarding levelling, noise floor, length of pauses and filtering.

II.2 Processing and scoring of speech samples

The applied processing chain is illustrated in Figure II.1. The AMR-WB codec corresponds to the bug-fixed version as in ITU-T G.722.2 (2003) Annex C Corrigendum 1 (2018).

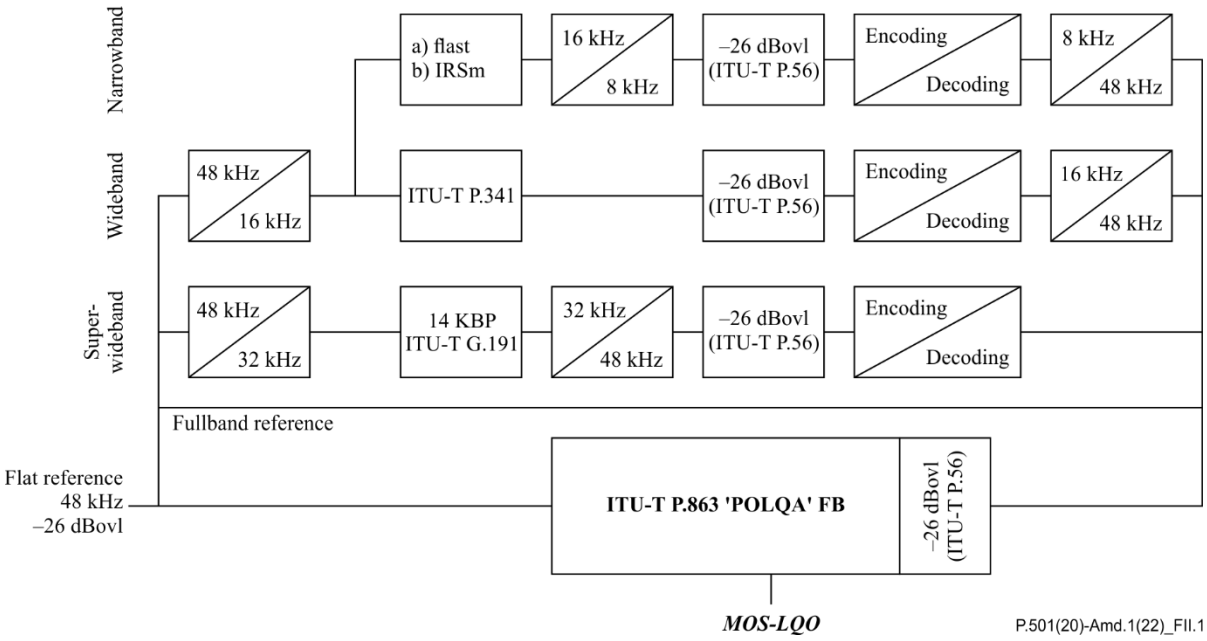

Figure II.1 – Processing chain for test conditions and codecs

II.3 ITU-T P.863 scores on Annex C and Annex D samples

Even if both sets of reference speech samples in Annex C and Annex D have been selected carefully, there is a certain variation in measured speech quality across the individual test conditions.

Figures II.2 and II.3 show the spread of ITU-T P.863 results for both sets of samples, where for Annex C the results for four samples of the same language are averaged.

As expected, the span of the results is smaller if averaging four samples compared to just having one single sample as in Annex D. However, the span for the short, single samples is still rather small due to the specific selection to meet results as typical across multiple samples.

It is also observed that the influence of the language itself is smaller than by the sample with its talker and text dependency. Even if Annex C covers eight different languages, the results for the same test conditions are very similar. It is assumed that the remaining differences are more driven from the talkers chosen and recording conditions than by the language itself.

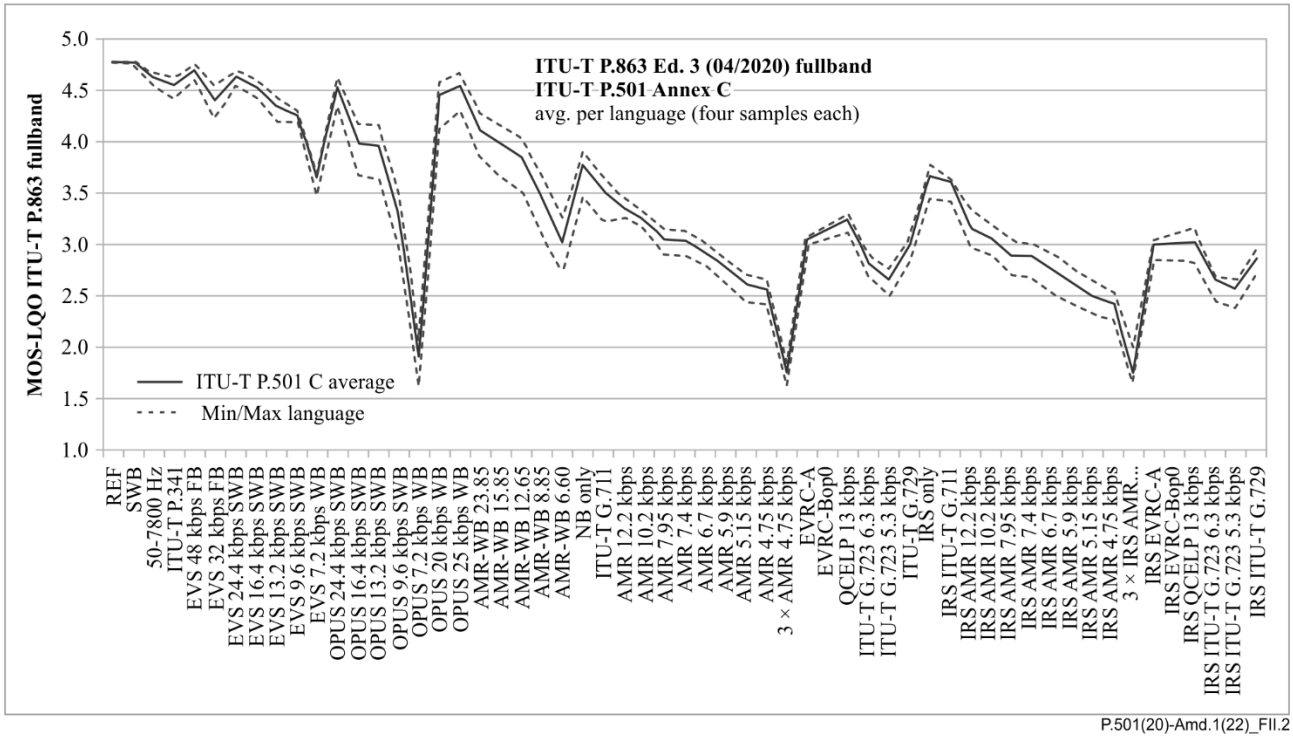

Figure II.2 – Averaged ITU-T P.863 scores on typical codec conditions for Annex C samples

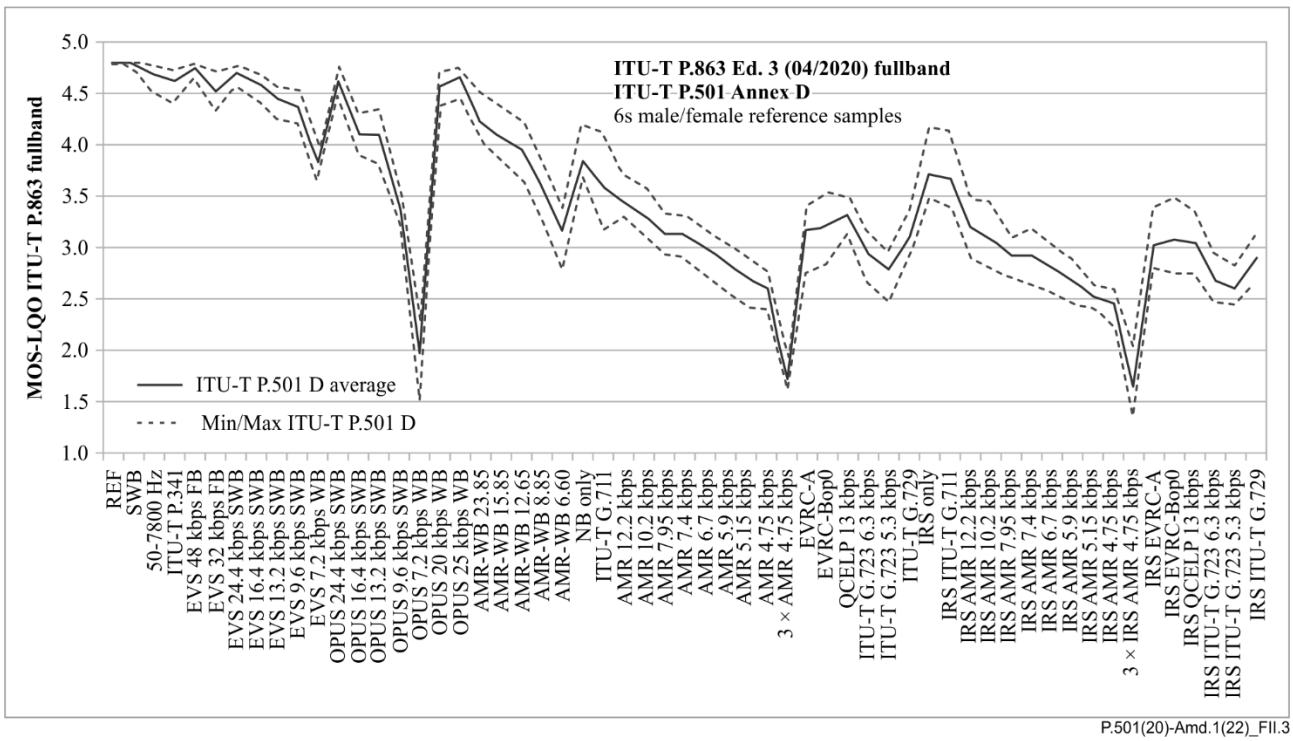

Figure II.3 – Averaged ITU-T P.863 scores on typical codec conditions for Annex D samples

Figure II.4 shows the individual results of all eight samples in Annex D. From the figure it is clear that the individual files are not constantly biased by each other, but rather show some other systematic differences. For example, the Italian (IT) sample produces results systematically above average for narrowband conditions, but is close to the average for wideband and above. In contrast, the German (GE) sample systematically results in below-average scores for narrowband conditions.

In general, the spread between the samples is smaller for conditions above wideband. The larger spread for band-limited conditions, especially narrowband, is most likely caused by the larger influence of the individual spectral distribution of the samples than by coding artefacts.

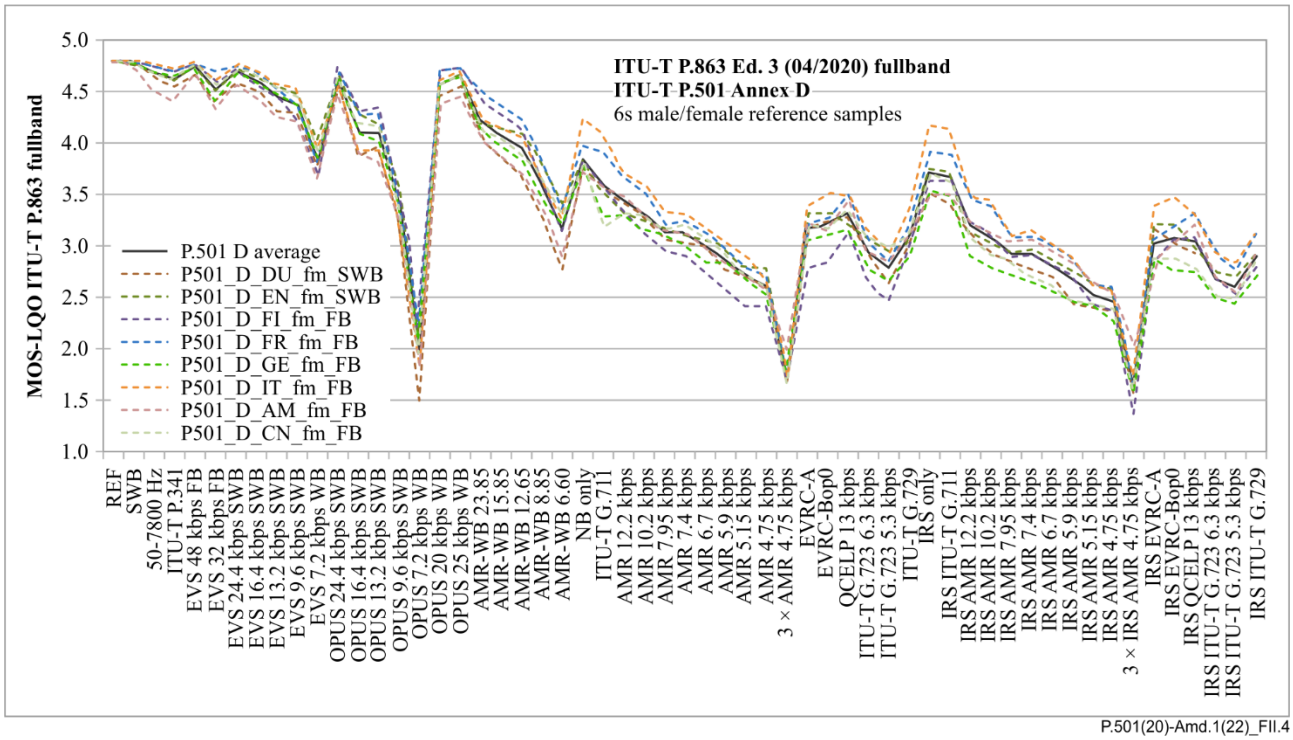

Figure II.4 – ITU-T P.863 scores on typical codec conditions for each of the Annex D samples

The EN sample (British English) is the sample that is closest to the average of all samples in Annexes D and C (see Figure II.5). It is closely followed by the AM sample (American English) that produces just slightly more pessimistic results for wideband and super-wideband conditions, and by the CN sample (Chinese) being a bit pessimistic in narrowband.

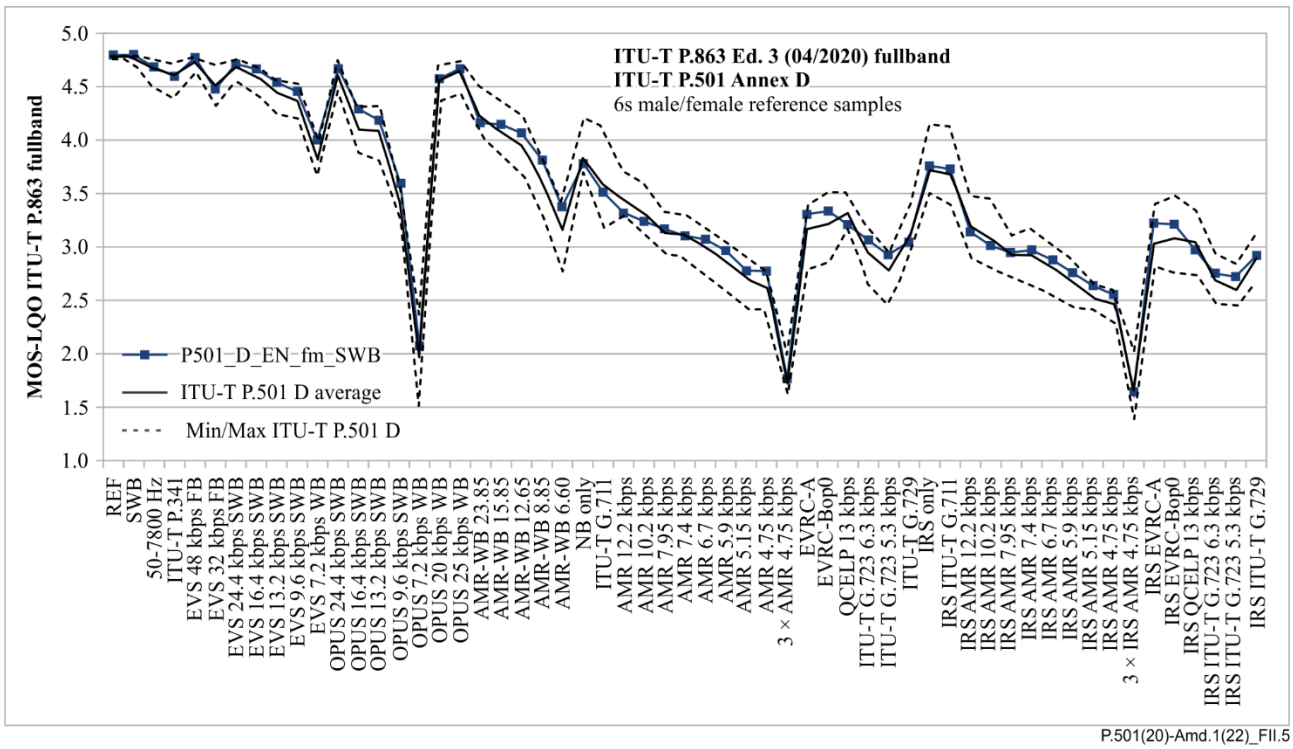

Figure II.5 – [ITU-T P.863] scores for the EN sample in Annex D

II.4 ITU-T P.863.1 scores on Annex C and Annex D samples

[ITU-T P.863.1] provides reference scores (see Table II.1) for typical codec conditions as average across all Annex C speech samples.

Table II.1 – Reference scores on typical codec conditions for ITU-T P.501 Annex C samples

Codec	[b-ITU-T P.862.1]	[ITU-T P.863] NB	[ITU-T P.863] FB Ed. 3 (04/2020)
Transparent 20-20'000 Hz	–	–	4.79
50-14'000 Hz			4.77
EVS 48 kbit/s FB			4.70
EVS 24.4 kbit/s SWB			4.64
EVS 16.4 kbit/s SWB			4.55
EVS 13.2 kbit/s SWB			4.35
EVS 9.6 kbit/s SWB			4.27
EVS 7.2 kbit/s WB	–	–	3.66
50-7'600 Hz	–	–	4.63
P.341	–	–	4.56
AMR-WB 23.85	–	–	4.10
AMR-WB 15.85	–	–	3.99
AMR-WB 12.65	–	–	3.85
AMR-WB 8.85	–	–	3.46
AMR-WB 6.60	–	–	3.02
NB only (50-3'800 Hz)	4.5	4.50	3.79

Table II.1 – Reference scores on typical codec conditions for ITU-T P.501 Annex C samples

Codec	[b-ITU-T P.862.1]	[ITU-T P.863] NB	[ITU-T P.863] FB Ed. 3 (04/2020)
IRS only	4.5	4.48	3.68
IRS G.711	4.4	4.4	3.61
IRS G.729 A	3.6	3.9	2.86
IRS G.723.1 6.3 kbit/s	3.6	3.8	2.66
IRS G.723.1 5.3 kbit/s	3.4	3.7	2.56
IRS AMR 12.2 kbit/s	4.0	4.2	3.17
IRS AMR 10.2 kbit/s	3.9	4.1	3.07
IRS AMR 7.95 kbit/s	3.7	3.9	2.89
IRS AMR 7.4 kbit/s	3.7	3.9	2.88
IRS AMR 6.7 kbit/s	3.6	3.8	2.76
IRS AMR 5.9 kbit/s	3.4	3.7	2.63
IRS AMR 5.15 kbit/s	3.3	3.6	2.50
IRS AMR 4.75 kbit/s	3.2	3.5	2.43
IRS EVRC-A	3.7	3.9	2,99
IRS EVRC-Bop0	3.7	4.0	3.02
IRS QCELP 13 kbit/s	3.9	4.0	3.03
NOTE 1 – IRS stands for IRS mod (send); 32 P.501 Annex C samples (Fullband, −26 dB SPL acc. P.56, cut to 8 s, initial pause 500 ms, noise floor −85 dB). NOTE 2 – [ITU-T P.863] in FB mode scores transparent FB speech samples at ≥4.75 MOS-LQO. Since, bandwidth limitations relative to FB signals are treated as degradations, any limitation leads to lower scores. On average, a plain wideband signal (50-7'600 Hz) is scored at 4.63 MOS-LQO and a plain narrowband signal (50-3'800 Hz) at 3.79. This means that in narrowband networks scores of 3.79 MOS-LQO are not exceeded on average, and real connections are usually below that value due to additional degradations. NOTE 3 – The AMR-WB codec conditions have been processed without DTX enabled.			

II.5 Results obtained with the speech sample from [ITU-T P.565.1]

Along with {ITU-T P.565], a dedicated speech sample for use with the model described in [ITU-T P.565.1] was provided as Annex D. This speech sample follows the construction principle of Annex D using American English, but is not based on samples in ITU-T P.501.

This sample has been evaluated using the same processing chain as the analysis of the ITU-T P.501 samples above.

The predicted scores (see Figure II.6) follow very well the average results across the Annex D samples for super-wideband codecs such as enhanced voice service (EVS) and OPUS, the wideband codec AMR-WB and plain filtered speech signals.

For all types of narrowband coding, this sample shows a clear overprediction compared to averages across the samples specified in this Recommendation. This overprediction clearly exceeds the maximum boundary of the Annex D samples for AMR, and is about 0.5 mean opinion score (MOS) above their average score.

This means that when applying this sample in practice or along with [ITU-T P.565.1], in typical narrowband conditions using AMR or EVRC, significant higher MOS values will be predicted compared to typical samples from Annex D, such as EN or AM.

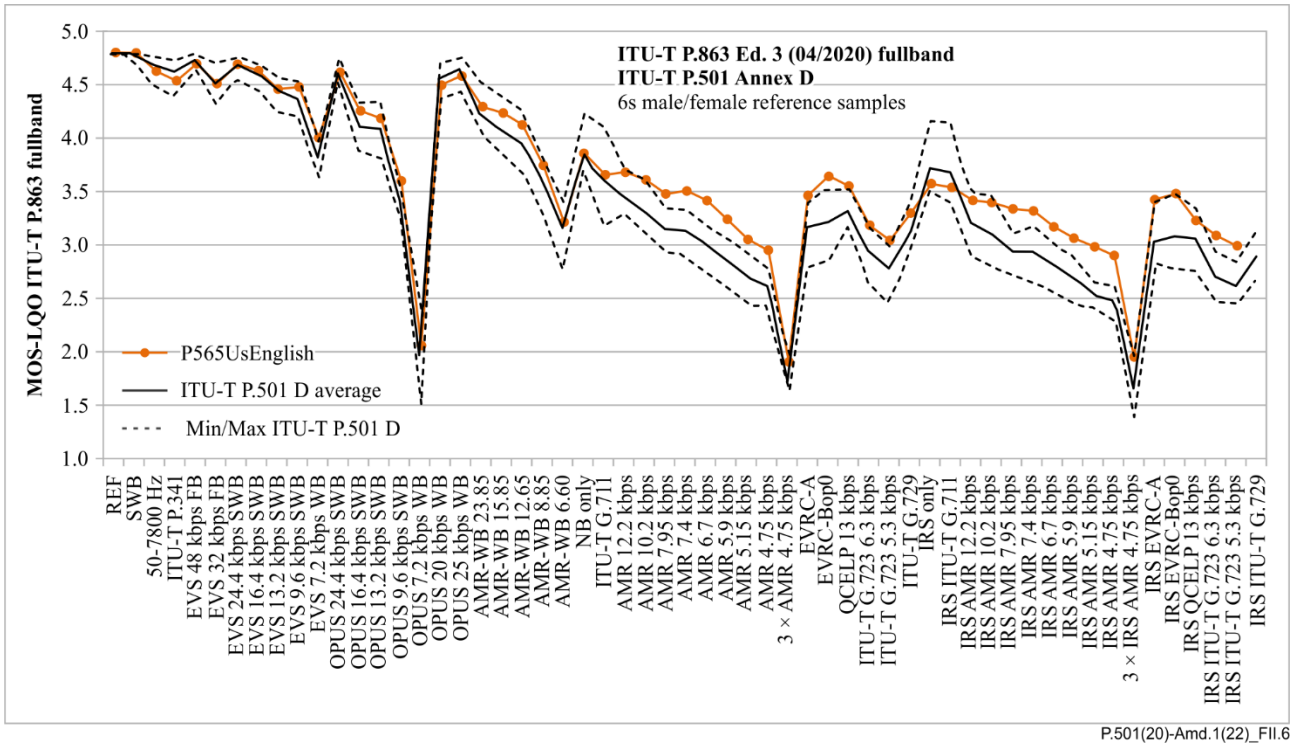

Figure II.6 – [ITU-T P.863] scores on typical codec conditions for the speech sample in [ITU-T P.565/P.565.1] vs. the Annex D reference samples

II.6 Summary

The speech samples in Annex D result in very similar average results as the ones in Annex C. If looking at results 'per language', the four samples as in Annex C result in a narrower spread than the single short samples in Annex D. Nevertheless, the spread of the individual short samples in Annex D is still smaller than for arbitrarily taken samples or the individual samples from Annex C. This narrow spread of samples in Annex D is due to the careful selection and composition of the speech samples.

The speech sample as published in Annex D shows a clear overprediction for narrowband codecs compared to Annex C and Annex D samples.

2) Bibliography

Add the following entries to the Bibliography:

Bibliography

[b-ITU-T P.565] Recommendation ITU-T P.565 (2021), *Framework for creation and performance testing of machine learning based models for the assessment of transmission network impact on speech quality for mobile packet-switched voice services.*

[b-ITU-T P.565.1] Recommendation ITU-T P.565.1 (2021), *Machine learning model for the assessment of transmission network impact on speech quality for mobile packet-switched voice services.*

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2022