INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.861

(08/96)

SERIES P: TELEPHONE TRANSMISSION QUALITY

Methods for objective and subjective assessment of quality

Objective quality measurement of telephone-band (300 - 3400 Hz) speech codecs

ITU-T Recommendation P.861

(Previously CCITT Recommendation)

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY

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 ITU-T List of Recommendations.

ITU-T RECOMMENDATION P.861

OBJECTIVE QUALITY MEASUREMENT OF TELEPHONE-BAND (300 - 3400 Hz) SPEECH CODECS

Summary

This Recommendation describes an objective method for estimating the subjective quality of telephone-band (300 - 3400 Hz) speech codecs.

This Recommendation specifies the production of source speech for objective quality measurement, codec and reference conditions for which the objective quality measurement method has been shown to provide valid results, the calculation of objective quality based on the objective quality measure called the Perceptual Speech Quality Measure (PSQM), the estimation of the subjective quality from the objective measurement results and an analysis of the results.

This Recommendation can be applied when evaluating the effects on subjective quality of speech codecs of speech input levels, talkers, bit rates and transcodings.

Source

ITU-T Recommendation P.861 was prepared by ITU-T Study Group 12 (1993-1996) and was approved under the WTSC Resolution No. 1 procedure on the 30th of August 1996.

Keywords

Objective speech quality measurement, Subjective speech quality evaluation.

FOREWORD

ITU (International Telecommunication Union) is the United Nations Specialized Agency in the field of telecommunications. The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of the ITU. The 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 Conference (WTSC), which meets every four years, establishes the topics for study by the ITU-T Study Groups which, in their turn, produce Recommendations on these topics.

The approval of Recommendations by the Members of the ITU-T is covered by the procedure laid down in WTSC Resolution No. 1 (Helsinki, March 1-12, 1993).

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.

© ITU 1996

All rights reserved. No part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from the ITU.

CONTENTS

Introduction

Subjective quality assessment of speech codecs is one of the key technologies in designing digital telecommunication networks. Recommendation P.830 defines subjective testing methodologies for speech codecs. Since subjective quality assessment is time-consuming and expensive, it is therefore desirable to develop an objective quality assessment methodology to estimate the subjective quality of speech codecs with less subjective testing.

The most widely-used objective speech quality measure demonstrating the performance of speech codecs is the Signal-to-Noise Ratio (SNR = S/N). However, it is pointed out that the SNR does not adequately predict subjective quality for modern network components. This is especially true for recent low bit-rate codecs. Therefore, a variety of more sophisticated objective quality measures, such as the LPC Cepstrum Distance Measure (CD) [1], Information Index (II) [2], Coherence Function (CHF) [3], Expert Pattern Recognition (EPR) [4], and Perceptual Speech Quality Measure (PSQM) [5] were developed. The performance of these systems, in terms of ability to give accurate estimates of subjective quality, has been investigated in ITU-T since the 1980s.

After careful comparisons among these objective quality measures, it was concluded that the PSQM best correlated with the subjective quality of coded speech. Therefore, this Recommendation describes objective quality assessment with the PSQM as the objective quality measure [12].

In order to assist the readers of this Recommendation in the development of their own implementation of the PSQM, a floppy diskette has been included with this Recommendation. A description of the contents of this diskette can be found in the `README` file on the diskette and in Appendix I.

Recommendation P.861

OBJECTIVE QUALITY MEASUREMENT OF TELEPHONE-BAND (300 - 3400 Hz) SPEECH CODECS

(Geneva, 1996)

1 Scope

Subjective quality assessment of speech codecs can be made in listening-only (one-way) tests or in conversational (two-way) tests. The objective quality measurement described in this Recommendation estimates the subjective quality in listening-only tests.

To demonstrate the subjective performance of a codec, the effects of a variety of quality factors should be investigated (see Recommendation P.830). The accuracy of the objective quality measurement described in this Recommendation has not been verified for examining all of the factors specified in Recommendation P.830. Table 1 is intended to be a guide to facilitate the readers' determination of the test factors, coding technologies and applications to which this Recommendation applies.

TABLE 1/P.861

Relationship of coding technologies, experimental factors and applications to this Recommendation

Test factors	Note
Speech input levels to a codec	1
Listening levels in subjective experiments	2
Talker dependencies	1
Multiple simultaneous talkers	2
Transmission channel errors	2
Bit rates if a codec has more than one bit-rate mode	1
Transcodings	1
Bit-rate mismatching between an encoder and a decoder if a codec has more than one bit-rate mode	2
Environmental noise in the sending side	2
Network information signals as input to a codec	2
Music as input to a codec	2
Delay	3
Short-term time warping of audio signal	2
Long-term time warping of audio signal	4
Temporal clipping of speech	2
Amplitude clipping of speech	2
Coding technologies	
Waveform	1
CELP and hybrids ≥4 kbit/s	1
CELP and hybrids <4 kbit/s	2

TABLE 1/P.861 (concluded)

Relationship of coding technologies, experimental factors and applications to this Recommendation

VOCODERs	2
Other coders	2
Applications	
Coder optimization	1
Coder evaluation	1
Coder selection	2
Network planning	5
Live network testing	6
In-service non-intrusive measurement devices	3

NOTES

1 The objective measure has demonstrated acceptable accuracy in the presence of this variable.

2 Insufficient information is available about the accuracy of the objective measure with regard to this variable.

3 The objective measure is known to provide inaccurate predictions when used in conjunction with this variable, or is otherwise not intended to be used with this variable.

4 The objective measure is known to provide inaccurate predictions when there is a significant amount of wander (more than 10% of the frame length). The applicability of the measure when there is a small amount of wander is for further study.

5 With caution, the objective measure might be used for some network planning purposes. The reader should note that there are important factors in network planning to which this Recommendation is not applicable (see the "Test factors" section of this table).

6 With caution, the objective measure might be used for some live network testing. The reader should note that there may be factors or technologies in a live network connection to which this Recommendation is not applicable (see the "Test factors" and "Coding technologies" sections of this table).

When comparing a codec with another codec or with a reference condition based on subjective experimental results, statistical tests that take the distributions of subjective votes into account are often used. Since the objective measurement in this Recommendation estimates only the mean of subjective votes (e.g. MOS, DMOS), such statistical tests cannot be applied to the results of objective measurement. Prediction of per cent poor or worse (%PoW) and per cent good or better (%GoB) are currently under study.

2 References

The following 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 are valid. All Recommendations and other references are subject to revision; all users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations listed below. A list of the currently valid ITU-T Recommendations is published regularly.

– ITU-T Recommendation P.50 (1993), *Artificial voices*.

– ITU-T Recommendation P.800 (1996), *Methods for subjective determination of transmission quality*.

– ITU-T Recommendation P.810 (1996), *Modulated Noise Reference Unit (MNRU)*.

- ITU-T Recommendation P.830 (1996), *Subjective performance assessment of telephone-band and wideband digital codecs*.
- CCITT Recommendation G.711 (1988), *Pulse Code Modulation (PCM) of voice frequencies*.
- CCITT Recommendation G.726 (1990), *40, 32, 24, and 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)*.
- CCITT Recommendation G.728 (1992), *Coding of speech at 16 kbit/s using low-delay code excited linear prediction*.
- ITU-T Recommendation G.729 (1996), *Coding of speech at 8 kbit/s using Conjugate Structure Algebraic Code-Excited Linear-Prediction (CS-ACELP).*
- CCITT Supplement No. 13 (1994) to the P-Series Recommendations.

3 Abbreviations

For the purposes of this Recommendation, the following abbreviations are used:

ACR	Absolute Category Rating
CELP	Code Excited Linear Prediction
DCR	Degradation Category Rating
DMOS	Degradation Mean Opinion Score
MOS	Mean Opinion Score
PSQM	Perceptual Speech Quality Measure

4 Definitions

For the purposes of this Recommendation, the following definition applies:

4.1 **dBov**: dB relative to the overload point of a digital system

5 Conventions

Subjective evaluation of 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. This Recommendation discusses an objective measurement technique for estimating subjective quality obtained in listening-only tests.

6 Summary of objective measurement procedure

Figure 1 illustrates the objective measurement procedure.

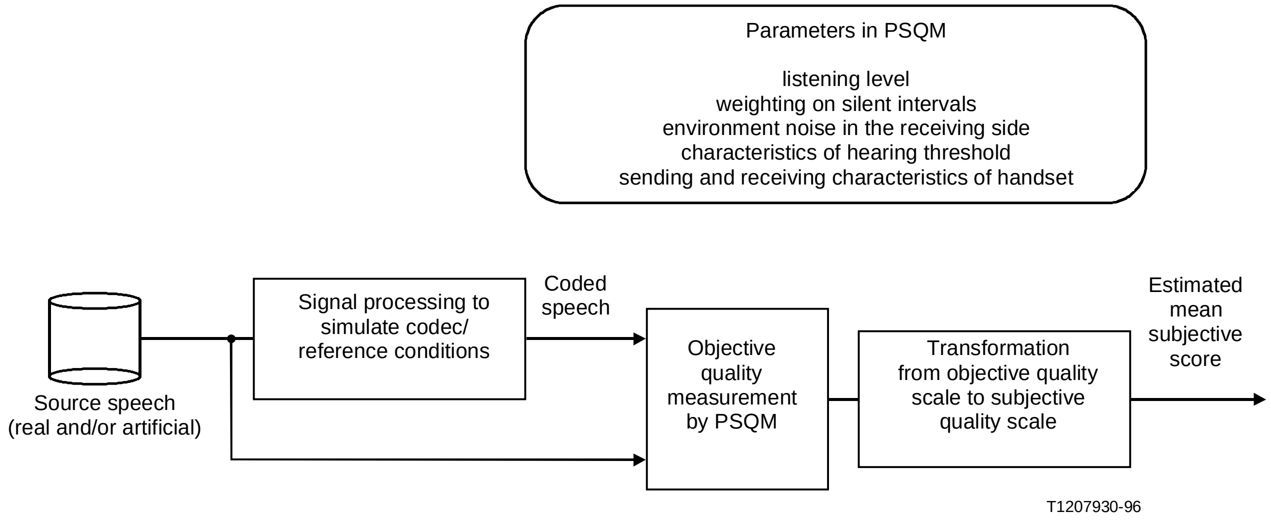
Parameters in PSQM
listening level
weighting on silent intervals
environment noise in the receiving side
characteristics of hearing threshold
sending and receiving characteristics of handset
Source speech (real and/or artificial)
Signal processing to simulate codec/ reference conditions
Coded speech
Objective quality measurement by PSQM
Transformation from objective quality scale to subjective quality scale
Estimated mean subjective score
T1207930-96

FIGURE 1/P.861

Objective quality measurement procedure

Objective quality measurement of speech codecs requires a number of steps:

1) preparation of source materials, i.e. recording of talkers and/or generation of the artificial voices conforming to Recommendation P.50;
2) selection of experimental parameters that will exercise the salient features of the codec and are able to be tested by objective measurement;
3) production of coded/reference speech;
4) calculation of the objective speech quality based on the Perceptual Speech Quality Measure (PSQM), using source and coded speech;
5) transformation from the objective quality scale to the subjective quality scale, if necessary;
6) analysis of results.

Each of these steps is described below.

7 Source speech material preparation

Source signals for objective measurement may be real voices or the artificial voices specified in Recommendation P.50, depending on the goals of the experiment.

Since the artificial voices defined in Recommendation P.50 reproduce the mean characteristics of human speech over various languages, they are useful in objectively estimating the mean subjective quality of a codec over these languages. When the talker-dependency of a codec or the performance of a codec for particular languages is concerned, it is recommended that real voices be used. In either case, no environmental noise should be added.

7.1 Real voices

When real voices are used in objective measurement, they should be produced, recorded and level-equalized in accordance with clause 7/P.830.

It is recommended that a minimum of two male talkers and two female talkers should be used for each testing condition. If talker dependency is to be tested as a factor in its own right, it is recommended that more talkers be used as follows:

– 8 male;

– 8 female;

– 8 children.

7.2 Artificial voices

When the artificial voices conforming to Recommendation P.50 are used in objective measurement, it is recommended that both male and female artificial voices be used. These signals should be passed through a filter with appropriate frequency characteristics to simulate sending frequency characteristics of a telephone handset, and level-equalized in the same manner as real voices (see Recommendation P.830).

ITU-T recommends the use of the Modified Intermediate Reference System (IRS) sending frequency characteristic as defined in Annex D/P.830.

8 Selection of experimental parameters

To demonstrate the performance of a codec, the effects of various quality factors on the performance of the codec should be examined. Recommendation P.830 provides guidance on subjectively assessing the following quality factors:

1) speech input levels to a codec;

2) listening levels in subjective experiments;

3) talkers (including multiple simultaneous talkers);

4) errors in the transmission channel between an encoder and a decoder;

5) bit rates if a codec has more than one bit-rate mode;

6) transcodings;

7) bit-rate mismatching between an encoder and a decoder if a codec has more than one bit-rate mode;

8) environmental noise in the sending side;

9) network information signals as input to a codec;

10) music as input to a codec.

Since the objective quality measure described in this Recommendation assumes:

1) source speech is "clean" (i.e. without added environmental noise in the sending side); and

2) there are no channel degradations such as transmission bit errors, frame erasures (e.g. as in mobile radio applications), or cell loss (e.g. as in ATM networks),

the quality factors to which this Recommendation applies are speech input levels, talkers (excluding multiple simultaneous talkers), bit rates and transcodings.

NOTE 1 – Objective measurement for quality factors other than those specifically noted as applicable in this Recommendation is still under study. Therefore, these factors should be measured only after the accuracy of an objective measure is verified in conjunction with subjective tests conforming to Recommendation P.830.

NOTE 2 - Although there are some indications that the objective measure can accurately predict quality under channel degradation conditions [10][11], the applicability of the measure to those conditions is still under study.

In addition to the codec conditions, Recommendation P.830 recommends the use of reference conditions in subjective tests. These conditions are necessary to facilitate the comparison of subjective test results from different laboratories or from the same laboratory at different times. Also, when expressing the objective test results in terms of equivalent-Q values, reference conditions using the narrow-band Modulated Noise Reference Unit (MNRU) as specified in Recommendation P.810 should be tested.

NOTE 3 – Including other standard codecs such as G.711 64-kbit/s PCM, G.726 32-kbit/s ADPCM, G.728 16-kbit/s LD-CELP, and G.729 8-kbit/s CS-ACELP as well as MNRU in objective quality measurement may help demonstrate the relative performance of the codec under test and standardized codecs.

Detailed explanations of these experimental parameters are found in Recommendation P.830.

9 Calculation of objective quality

This clause describes a method for measuring the quality of telephone-band (300 - 3400 Hz) coded speech using the Perceptual Speech Quality Measure (PSQM). The objective of PSQM is to mimic the sound perception of subjects in real-life situations [6]. The PSQM simulates experiments in which subjects judge the quality of speech codecs. It does this by comparing a coded signal to a source signal (Figure 2). Although this basic principle of comparison makes it especially suited for Degradation Category Rating (DCR) testing, Absolute Category Rating (ACR) experiments can be simulated as shown in the validation tests [12]. To the extent that PSQM is a faithful representation of human perception and judgement processes, inaudible differences between input and output will receive the same PSQM score. In particular, if the input and the output are identical, PSQM will predict perfect quality irrespective of the quality of the input signal.

Within PSQM, the physical signals constituting the source and coded speech are mapped onto psychophysical representations that match the internal representations of the speech signals (the representations inside our heads) as closely as possible. These internal representations make use of the psychophysical equivalents of frequency (critical band rates) and intensity (Compressed Sone). Masking is modelled in a simple way: only when two time-frequency components coincide in both the time and frequency domains, masking is taken into account.

Within the PSQM approach, the quality of the coded speech is judged on the basis of differences in the internal representation. This difference is used for the calculation of the noise disturbance as a function of time and frequency. In PSQM, the average noise disturbance is directly related to the quality of coded speech.

The transformation from the physical (external) domain to the psychophysical (internal) domain is performed by three operations:

– time-frequency mapping;

– frequency warping;

– intensity warping (compression).

Besides perceptual modelling, the PSQM method also uses cognitive modelling [7] in order to get high correlations between subjective and objective measurements.

Figure 3 shows a block diagram of the PSQM algorithm.

All the parameters and variables in this clause are summarized in Tables 2, 3 and 4.

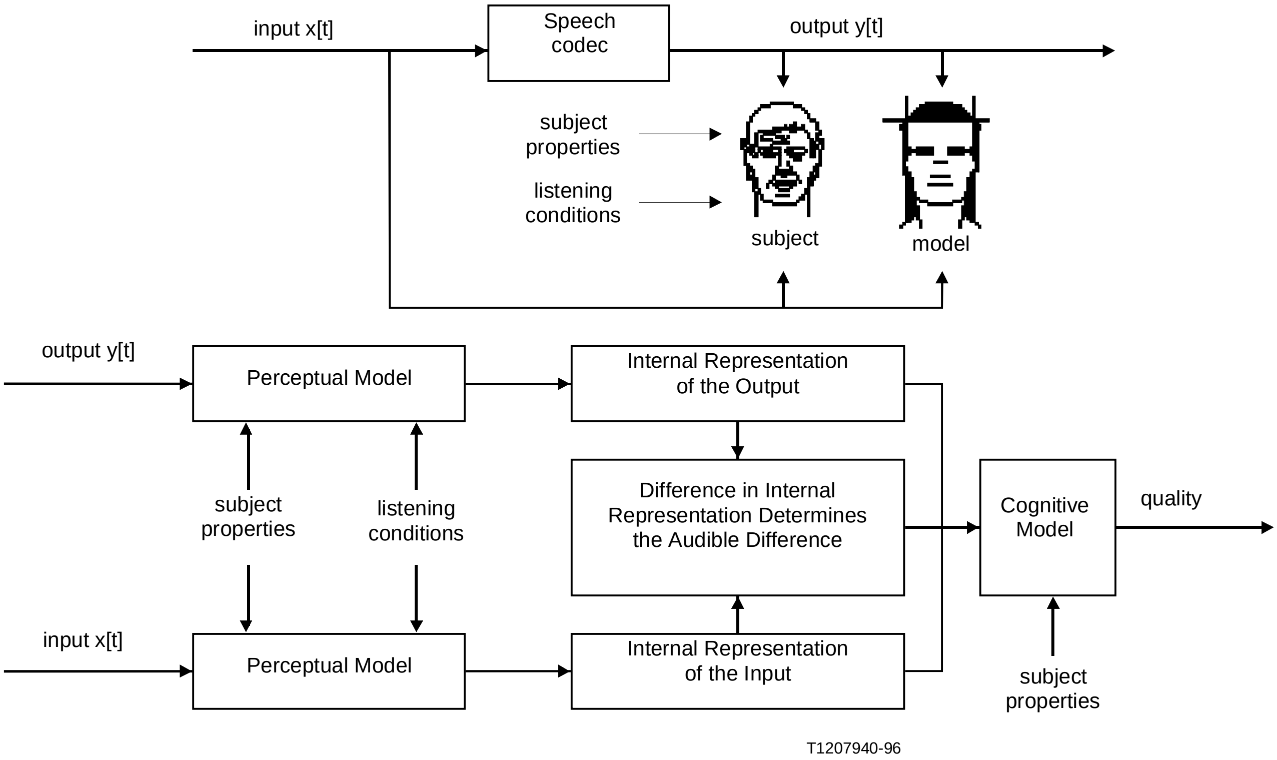

FIGURE 2/P.861

Overview of the basic philosophy used in the development of the PSQM – A computer model of the subject, consisting of a perceptual and a cognitive model, is used to compare the output of the speech codec with the input

TABLE 2/P.861

List of parameters in PSQM

Name	Description	Value
Nb	number of bands in critical band (Bark) domain	(see Table 4)
Nf	number of samples in time frame	512 for 16-kHz sampling frequency 256 for 8-kHz sampling frequency
F[j]	handset receiving frequency characteristics	IRS from Recommendation P.830 (Table 4 contains the IRS power transfer function)
H[j]	Hoth characteristics	(Table 4 contains the additive power of the Hoth characteristic)
P_0[j]	absolute threshold of hearing	(Table 4 contains the equivalent power representation of P_0[j])
Δf[j]	bandwidth of band j in Hertz	(see Table 4)
Δz	bandwidth of each subband in critical band domain	0.312
γ	exponent of compression function	0.001
W_{sil}	weighting factor on silent frames	0.2 (provisional)
W_{sp}	weighting factor on active speech frames	W_{sp} = (1- W_{sil})/W_{sil} = 4.0 (provisional)

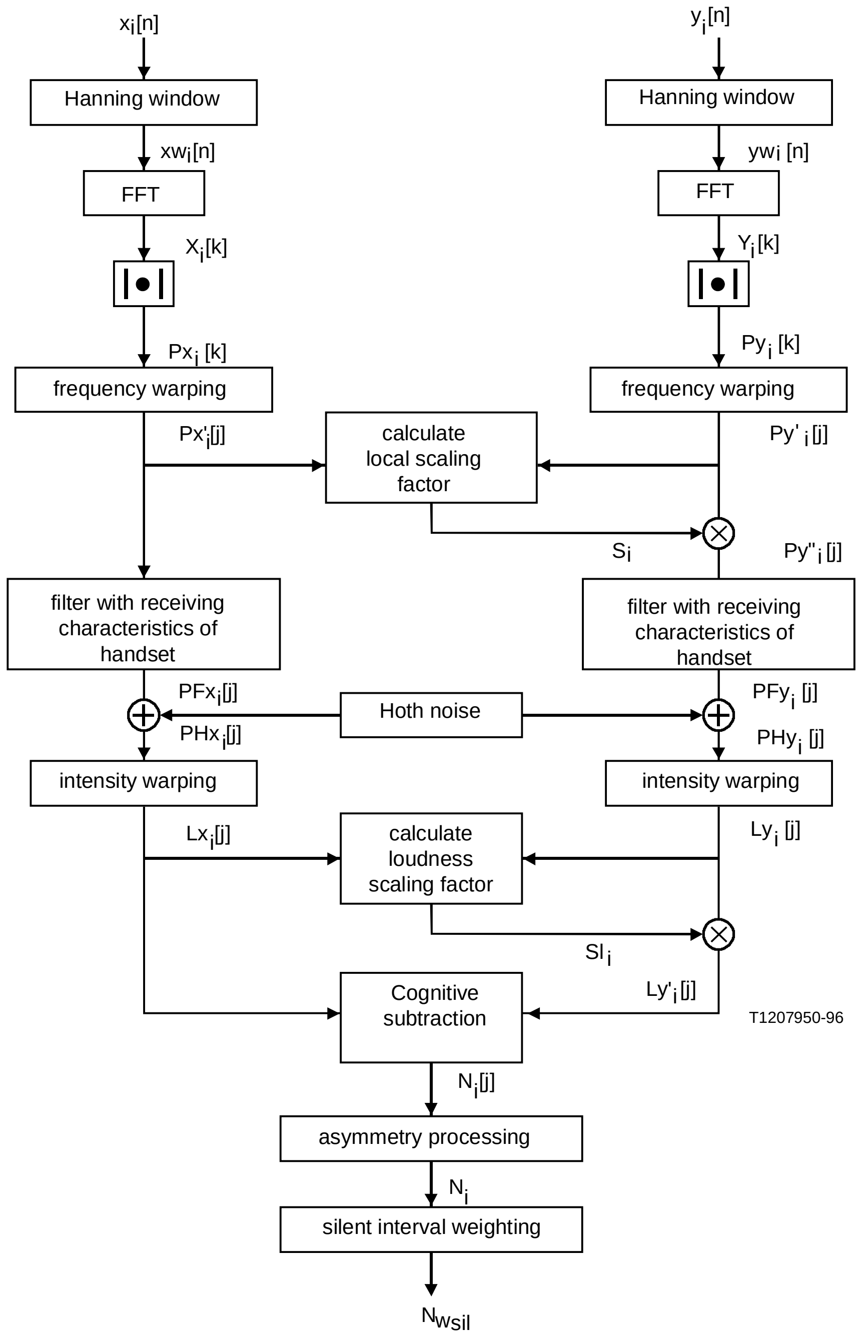

FIGURE 3/P.861

Block diagram of PSQM algorithm

TABLE 3/P.861

Variables in PSQM

Name	Description
m	index in time domain
n	index in time domain in a frame (n: 1, 2, 3, … , Nf)
i	index for frames
j	index in warped-frequency domain (critical band domain) (j: 1, 2, 3, … , Nb)
k	index in frequency domain (Hz) (k: 1, 2, 3, … , Nf/2)
x[m]	time-aligned and global-calibrated version of sampled source speech signal
y[m]	time-aligned, global-scaled, and global-calibrated version of sampled coded speech signal
S_{global}	scaling factor in global scaling
S_p	pitch power calibration factor
S_l	pitch loudness calibration factor
$x_i[n]$	x[m] in frame I
$y_i[n]$	y[m] in frame I
$xw_i[n]$	windowed version of $x_i[n]$
$yw_i[n]$	windowed version of $y_i[n]$
$X_i[k]$	FFT of $xw_i[n]$
$Y_i[k]$	FFT of $yw_i[n]$
$Px_i[k]$	SPD of $xw_i[n]$
$Py_i[k]$	SPD of $yw_i[n]$
$I_f[j]$	FFT index of first value of k of $Px_i[k]$ and $Py_i[k]$ in band j
$I_l[j]$	FFT index of last value of k of $Px_i[k]$ and $Py_i[k]$ in band j
$Px'_i[j]$	sampled Pitch Power Density of $xw_i[n]$
$Py'_i[j]$	sampled Pitch Power Density of $yw_i[n]$
Px'_i	power of source speech signal in frame i
Py'_i	power of coded speech signal in frame i
$Py''_i[j]$	local-scaled version of $Py'_i[j]$
$PFx_i[j]$	telephone-band filtered version of $Px'_i[j]$
$Pfy_i[j]$	telephone-band filtered version of $Py''_i[j]$
$Phx_i[j]$	$PFx_i[j]$ plus Hoth noise as environmental noise (receiving)
$Phy_i[j]$	$PFy_i[j]$ plus Hoth noise as environmental noise (receiving)
S_i	scaling factor in local scaling in frame i
S_{av}	average (arithmetic mean) of S_i
$Lx_i[j]$	sampled Compressed Loudness Density of source speech signal in frame i and band j
$Ly_i[j]$	sampled Compressed Loudness Density of coded speech signal in frame i and band j

TABLE 3/P.861 (concluded)

Variables in PSQM

Name	Description
Lx_i	momentary Compressed Loudness of source speech signal in frame i
Ly_i	momentary Compressed Loudness of coded speech signal in frame i
Sl_i	scaling factor in loudness scaling in frame i
$Ly'_i[j]$	loudness-scaled version of $Ly_i[j]$
$N_i[j]$	sampled noise disturbance density in frame i and band j
$C_i[j]$	asymmetry-effect factor in frame i and band j
N_i	noise disturbance in frame i
N_{wsil}	average noise disturbance with weighting on silent frames
M_{sp}	number of active speech frames
M_{sil}	number of silent frames
N_{spav}	average of N_i over active speech frames
N_{silav}	average of N_i over silent frames

TABLE 4/P.861

Critical band frequency allocations and filter characteristics (Based on a 16 kHz sampling rate)

Band number, j	Upper frequency [Hz]	First FFT Bin in Band j, I_f	Last FFT Bin in Band j, I_l	Receiving Characteristic, F	Hearing Threshold, P_0	Hoth noise, H
0	15.6	0	0	discarded in processing		
1	46.9	1	1	2.45E-06	3.89E+07	1.72E+04
2	78.1	2	2	9.24E-06	1.12E+06	1.72E+04
3	109.4	3	3	3.56E-05	1.26E+05	1.72E+04
4	140.6	4	4	2.59E-04	1.86E+04	1.22E+04
5	171.9	5	5	1.18E-03	6.17E+03	8.49E+03
6	203.1	6	6	7.48E-03	2.29E+03	6.31E+03
7	234.4	7	7	3.19E-02	9.33E+02	4.91E+03
8	265.6	8	8	7.31E-02	4.37E+02	3.95E+03
9	296.9	9	9	1.37E-01	2.29E+02	3.26E+03
10	328.1	10	10	2.09E-01	1.29E+02	2.74E+03
11	359.4	11	11	2.93E-01	7.76E+01	2.35E+03
12	390.6	12	12	4.25E-01	4.27E+01	2.04E+03
13	421.9	13	13	5.23E-01	3.02E+01	1.79E+03
14	453.1	14	14	5.98E-01	2.19E+01	1.59E+03
15	484.8	15	15	6.51E-01	1.66E+01	1.44E+03
16	519.2	16	16	6.94E-01	1.32E+01	1.39E+03

TABLE 4/P.861

Critical band frequency allocations and filter characteristics (Based on a 16 kHz sampling rate)

Band number, j	Upper frequency [Hz]	First FFT Bin in Band j, I_f	Last FFT Bin in Band j, I_l	Receiving Characteristic, F	Hearing Threshold, P_0	Hoth noise, H
17	553.6	17	17	7.31E-01	1.07E+01	1.25E+03
18	590.8	18	18	7.66E-01	8.91E+00	1.22E+03
19	631.2	19	20	7.98E-01	7.59E+00	1.19E+03
20	672.9	21	21	8.37E-01	6.31E+00	1.10E+03
21	716.6	22	22	8.63E-01	5.62E+00	1.04E+03
22	760.4	23	24	8.88E-01	5.13E+00	9.45E+02
23	804.6	25	25	9.12E-01	4.68E+00	8.69E+02
24	851.4	26	27	9.35E-01	4.37E+00	8.41E+02
25	898.3	28	28	9.56E-01	4.17E+00	7.68E+02
26	947.0	29	30	9.71E-01	4.07E+00	7.33E+02
27	997.0	31	31	9.80E-01	3.98E+00	6.90E+02
28	1051.	32	33	9.87E-01	3.98E+00	6.87E+02
29	1108.	34	35	9.90E-01	3.98E+00	6.57E+02
30	1168.	36	37	9.91E-01	3.98E+00	6.49E+02
31	1231.	38	39	9.93E-01	3.98E+00	6.17E+02
32	1297.	40	41	9.95E-01	4.07E+00	5.95E+02
33	1366.	42	43	1.00E+00	4.27E+00	5.68E+02
34	1437.	44	45	1.01E+00	4.47E+00	5.37E+02
35	1509.	46	48	1.02E+00	4.68E+00	5.04E+02
36	1582.	49	50	1.04E+00	5.01E+00	4.80E+02
37	1658.	51	53	1.06E+00	5.37E+00	4.51E+02
38	1736.	54	55	1.07E+00	5.62E+00	4.37E+02
39	1817.	56	58	1.09E+00	5.89E+00	4.20E+02
40	1902.	59	60	1.10E+00	6.31E+00	4.05E+02
41	1991.	61	63	1.11E+00	6.61E+00	3.97E+02
42	2084.	64	66	1.12E+00	6.92E+00	3.86E+02
43	2184.	67	69	1.12E+00	7.24E+00	3.82E+02
44	2289.	70	73	1.12E+00	7.59E+00	3.74E+02
45	2401.	74	76	1.11E+00	7.76E+00	3.67E+02
46	2520.	77	80	1.10E+00	7.94E+00	3.63E+02
47	2 647.	81	84	1.08E+00	7.94E+00	3.56E+02
48	2781.	85	88	1.01E+00	7.94E+00	3.46E+02
49	2922.	89	93	8.62E-01	7.94E+00	3.37E+02
50	3069.	94	98	6.86E-01	8.13E+00	3.25E+02
51	3225.	99	103	5.16E-01	8.13E+00	3.16E+02
52	3392.	104	108	3.12E-01	8.32E+00	2.92E+02

TABLE 4/P.861

Critical band frequency allocations and filter characteristics
(Based on a 16 kHz sampling rate)

Band number, j	Upper frequency [Hz]	First FFT Bin in Band j, I_f	Last FFT Bin in Band j, I_l	Receiving Characteristic, F	Hearing Threshold, P_0	Hoth noise, H
53	3572.	109	114	1.55E-01	8.32E+00	2.69E+02
54	3765.	115	120	3.02E-02	8.32E+00	2.47E+02
55	3971.	121	127	2.03E-03	8.32E+00	2.25E+02
56	4193.	128	134	1.52E-04	8.32E+00	2.06E+02

NOTES

1 The absolute threshold, P_0, uses the calibration 0 dB SPL=1.0.

2 The first upper frequency (15.6 Hz) is equivalent to 0.156 of a critical band. The bandwidth Δz is 0.312 of a critical band.

9.1 Global initializations

Before starting the computation of the noise disturbance, which is the output of the PSQM algorithm, the following global initializations should be carried out for each pair of source and coded speech:

– time alignment;

– global scaling for compensation of the system gain;

– global calibration for setting the loudness of the speech.

Since telephone-band speech codecs usually adopt an input sampling frequency of 8 kHz, this Recommendation assumes both the source and coded speech have a sampling frequency of 8 kHz or 16 kHz (i.e. up-sampled by a factor of 2).

9.1.1 Time alignment

The first global initialization that should be carried out is the time alignment of the source signal x[m] and the coded signal y[m]. If the signals are not aligned properly, PSQM cannot be applied.

When the time lag in the coded signal relative to the source signal is unknown theoretically, the time lag that gives the maximum of the cross-correlation between source and coded signals can be used as an estimate. For signals that show group delay distortion, the delay that leads to the minimum PSQM value is the correct one.

In the processing, leading and trailing zeros in the speech file are discarded and the start point and stop point are calculated by detecting speech activity using only the source signal. The algorithms for the determination of the first and last active speech sample are as follows.

When determining the start of active speech in a file, the first sample to be declared active is the one in which the magnitude (i.e. absolute value) of that sample, plus the magnitudes of the four preceding samples total 200 or more. (For the purposes of testing the first four samples for the start of speech activity, samples preceding the first sample are considered to have a value of 0.)

When determining the end of active speech in a file, the last sample to be declared active is the last sample for which the magnitude (i.e. absolute value) of that sample, plus the magnitudes of the four following samples total 200 or more. (For the purposes of testing the last four samples for the end of speech activity, samples following the last sample are considered to have a value of 0.)

9.1.2 Global scaling

After the time-alignment process, the coded signal y[m] is scaled in order to compensate for the overall gain of the system. The scaling factor S_{global} is defined by:

$$S_{global} = \sqrt{\frac{\sum_{start\ point}^{stop\ point} x^2[m]}{\sum_{start\ point}^{stop\ point} y^2[m]}}$$

The coded signal y[m] is then multiplied by S_{global}.

9.1.3 Global calibration

In order to ensure optimum accuracy of the objective measure, it is necessary to provide a calibration between the listening level and the compressed loudness. The values in Table 4 are based on the assumption that 0 dB SPL is equivalent to a maximum value of 1.0 in the pitch power domain as computed in 9.3.1 [i.e. $\max_j (Px_i'[j])$ = 1.0 for a given frame]. Also assumed is that the optimum listening level of 78 dB SPL is used in conjunction with speech files that have an active speech level of −26 dBov, as indicated in Recommendation P.830.

The calibrations are performed with a 1 kHz sine wave at a level of 40 dB SPL (i.e. –64 dBov). This is best performed with a real (i.e. non-integer) sine wave, to avoid quantization artefacts in the calibration function. A level of 40 dB SPL corresponds to a zero-to-peak amplitude of 29.54.

The first calibration is to scale the maximum value of the pitch power representation of the calibration tone to 10 000 [i.e. if the $\max_j (Px_i'[j]) = 1.0$ for 0 dB SPL, the $\max_j (Px_i'[j]) = 10\,000$ for 40 dB SPL]. This calibration factor, S_p, is calculated by:

$$S_p = \frac{10000}{\max_j (Px_i'[j])}$$

when $Px'_i[j]$ (see 9.3.1) is calculated for the calibration tone. For an implementation of PSQM where the FFT is scaled by *n*, as in the commercially available routine "four1" from *Numerical Recipes in C* [13],

$$S_p = 6.4661\,\mathrm{e}^{-06}$$

The second calibration sets the compressed loudness of the calibration tone, as calculated in 9.4, to 1.0 Compressed Sone. The calibration factor is calculated by:

$$S_l = \frac{1}{Lx_i}$$

when Lx_i is calculated for the calibration tone. If the first calibration is performed correctly, S_l = 240.05.

NOTE 1 – The calibration tone should not be filtered through the receiving characteristic, F, nor should Hoth noise be added to the calibration tone prior to the computation of Lx_i and S_l. This exception is for calibration purposes only.

If the active speech level in the digital file is not –26 dBov, or the listening level is not 78 dB SPL, the input data should be scaled accordingly.

NOTE 2 – In a 16-bit digital file, 0 dBov is represented by a DC level of 32 767. Therefore, a sinusoid with a 0-to-peak amplitude of 32 767 would have an RMS level of –3.01 dBov. With the assumptions in this subclause, that would correspond to approximately 101 dB SPL.

9.2 Time-frequency mapping

The mapping from time domain to time-frequency domain is implemented with a short-term Fourier transform with a Hanning window resulting in a time-frequency representation with a constant resolution in both time and frequency domains.

9.2.1 Windowing

The source signal $x_i[n]$ and coded signal $y_i[n]$ in frame i are windowed using a Hanning ($\sin^2$) window:

$$xw_i[n] = w[n] \cdot x_i[n]$$

$$yw_i[n] = w[n] \cdot y_i[n]$$

with $w[n]$ as the window function.

The windowing function can be computed as follows:

$$w[n] = 0.5\left(1 - \cos\left(\frac{2\pi i}{Nf}\right)\right) \quad for\ 0 \le i \le Nf - 1$$

Throughout clause 9, all calculations are defined on a frame-by-frame basis. A frame length of 256 samples for 8-kHz sampling and 512 samples for 16-kHz sampling, which approximately correspond to the window length of the ear, should be used and adjacent frames should overlap each other by 50%.

9.2.2 Sampled Spectral Power Density (SPD)

The sampled SPDs of $xw_i[n]$ and $yw_i[n]$, denoted by $Px_i[k]$ and $Py_i[k]$, are calculated by using the Fast Fourier Transforms (FFTs):

$$xw_i[n] \Rightarrow FFT \Rightarrow X_i[k]$$

$$yw_i[n] \Rightarrow FFT \Rightarrow Y_i[k]$$

$$Px_i[k] = (\mathrm{Re}\, X_i[k])^2 + (\mathrm{Im}\, X_i[k])^2$$

$$Py_i[k] = (\mathrm{Re}\, Y_i[k])^2 + (\mathrm{Im}\, Y_i[k])^2$$

9.3 Frequency warping and filtering

This subclause first describes the warping from the Hertz scale to the critical band scale, leading to a sampled pitch power density representation within each frame. The sampled pitch power density of the coded signal is scaled within each frame after which both the source and coded signals are telephone-band filtered and Hoth noise is added to simulate the listening environment. Finally, the signals are filtered with the transfer function from the outer- to inner-ear characteristic.

9.3.1 Sampled Pitch Power Density

The frequency index k in Hz is transformed to a pitch index j in the critical band domain by a frequency-scale warping. First the critical band scale is divided into equal interval bands, and for each band a pitch power density value (a sample) is computed from the samples (usually more than 1) of the spectral power density in the corresponding band on the Hertz scale. The sampled pitch power densities $Px'_i[j]$ and $Py'_i[j]$ for band j in frame i are found by:

$$Px_i'[j] = S_p \cdot \frac{\Delta f_j}{\Delta z} \cdot \frac{1}{I_l[j] - I_f[j] + 1} \cdot \sum_{I_f[j]}^{I_l[j]} Px_i[k]$$

$$Py_i'[j] = S_p \cdot \frac{\Delta f_j}{\Delta z} \cdot \frac{1}{I_l[j] - I_f[j] + 1} \cdot \sum_{I_f[j]}^{I_l[j]} Py_i[k]$$

where $I_f[j]$ is the index of the first and $I_l[j]$ is index of the last sample on the Hertz scale for band *j*, with Δf_j as the bandwidth in band *j* in Hertz, Δz as the bandwidth of each subband in the critical band domain, and S_p as the pitch power calibration factor as indicated in 9.1.3.

9.3.2 Local scaling

The source and coded signals should be scaled within each frame, as a compensation for slow gain variations. Only audible time-frequency components are taken into account (above the absolute threshold of audibility for each band $P_0[j]$ defined in Table 4). The total powers of the source and coded signals in a frame *i*, Px'_i and Py'_i, are computed using the warped frequency representation:

$$Px_i' = \sum_{j=1}^{Nb} Px_i'[j]$$

$$Py_i' = \sum_{j=1}^{Nb} Py_i'[j]$$

with *Nb* as the total number of bands.

When both the power of the source signal Px'_i and the power of the coded signal Py'_i are above 40 dB SPL, the power of the coded signal for band *j*, $Py'_i[j]$, is multiplied by a scaling factor S_i:

$$Py_i''[j] = S_i \cdot Py_i'[j]$$

where:

$$S_i = \frac{Px_i'}{Py_i'}$$

When either the power of the source signal Px'_i or the power of the coded signal Py'_i is below 40 dB SPL, the power of the coded signal for band *j*, $Py'_i[j]$, is multiplied by a scaling factor S_{av} which is the average of all factors S_i calculated earlier.

9.3.3 Telephone-band filtering

$Px'_i[j]$ and Py''_i[j] should be filtered using receiving characteristics appropriate for a telephone handset:

$$PFx_i[j] = F[j] \cdot Px_i'[j]$$

$$PFy_i[j] = F[j] \cdot Py_i''[j]$$

where *F*[*j*] is the frequency response in band j of the receiving characteristics of a handset. ITU-T recommends the use of the modified IRS receiving characteristics defined in Annex D/P.830 as receiving frequency characteristics of a telephone handset. The values of *F*[*j*] for these characteristics are given in Table 4.

9.3.4 Hoth noise

In normal telephone use, the speech signal is disturbed by surrounding sounds in the receiving environment. Within PSQM, this effect is modelled by adding Hoth noise to both the source and coded signals. The Hoth noise [8] is added to the sampled pitch power density for every value of *j*, using the spectral power density function as given in Recommendation P.800:

$$PHx_i[j] = H[j] \cdot PFx_i[j]$$

$$PHy_i[j] = H[j] \cdot PFy_i[j]$$

where *H*[*j*] is the power of Hoth noise in band *j* given in Table 4.

NOTE – All validations of the PSQM method within the ITU-T were performed using Hoth noise at a level of 45 dBA.

9.4 Intensity warping

After calculating the sampled pitch power densities that take into account telephone-band filtering and Hoth noise, the intensity scale is warped to a loudness scale leading to a sampled compressed loudness density function.

From the pitch power densities $PHx_i[j]$ and $PHy_i[j]$, the sampled compressed loudness densities $Lx_i[j]$ and $Ly_i[j]$ are calculated using a compression function given by Zwicker [9]:

$$Lx_i[j] = S_l \cdot \left(\frac{P_0[j]}{0.5}\right)^{\gamma} \cdot \left[\left(0.5 + 0.5 \cdot \frac{PHx_i[j]}{P_0[j]}\right)^{\gamma} - 1\right]$$

$$Ly_i[j] = S_l \cdot \left(\frac{P_0[j]}{0.5}\right)^{\gamma} \cdot \left[\left(0.5 + 0.5 \cdot \frac{PHy_i[j]}{P_0[j]}\right)^{\gamma} - 1\right]$$

with $P_0[j]$ as the internal threshold as given in Table 4, and S_l as the pitch loudness calibration factor as explained in 9.1.3. Negative values of $Lx_i[j]$ and $Ly_i[j]$ are set to zero.

The optimal value of γ found in optimizations using databases that resulted from different speech quality evaluation experiments is 0.001.

The (total) momentary compressed loudnesses Lx_i and Ly_i (in Compressed Sone) are computed by summation of the sampled compressed loudness densities $Lx_i[j]$ and $Ly_i[j]$:

$$Lx_i = \sum_{j=1}^{Nb} Lx_i[j] \cdot \Delta z$$

$$Ly_i = \sum_{j=1}^{Nb} Ly_i[j] \cdot \Delta z$$

with Δz as the bandwidth in the critical band domain. The momentary compressed loudnesses Lx_i and Ly_i are used in the cognitive modelling.

9.5 Cognitive modelling

Within the PSQM context, all operations that cannot be performed on the source signal alone or on the coded signal alone are defined as cognitive operations. Four cognitive effects are discussed in this subclause:

- loudness scaling;
- internal cognitive noise;
- asymmetry processing;
- silent interval processing.

9.5.1 Loudness scaling

Within PSQM, the sampled compressed loudness density of the coded signal is scaled, within each frame, relative to the loudness of the source signal:

$$Ly_i^{'}[j] = Sl_i \cdot Ly_i[j]$$

where the scaling factor Sl_i is computed from the (total) momentary compressed loudnesses Lx_i and Ly_i:

$$Sl_i = \frac{Lx_i}{Ly_i}$$

When Lx_i or Ly_i is below 0.02 Compressed Sone, Sl_i is set equal to 1.

9.5.2 Sampled noise disturbance density

The sampled noise disturbance density $N_i[j]$ in band j in frame i is computed as the absolute difference between $Lx_i[j]$ and $Ly'_i[j]$:

$$N_i[j] = \left| Ly_i^{'}[j] - Lx_i[j] \right| - 0.01$$

where the factor 0.01 Compressed Sone represents the internal cognitive noise. If, because of this factor, $N_i[j]$ becomes negative, then $N_i[j]$ is set equal to zero.

9.5.3 Asymmetry processing

When a new time-frequency component is introduced in the speech signal, the subjective quality turns out to be more degraded than when an equally loud component is left out by the speech codec. This asymmetry is most prominent during the silent intervals. Noise that is present in the source signal may be suppressed, leading to an increase in quality. If there is no noise during the silent intervals in the source signal, any difference between source and coded speech leads to a decrease in quality.

Furthermore, when a time-frequency component is left out of the source signal (not coded by the codec), the remaining signal is still one coherent auditory event. If a new unrelated time-frequency component is introduced into the coded signal (a distortion), the newly-formed signal can be decomposed in two parts, the original signal and the distortion. This decomposition of the auditory stream makes the noise more annoying.

The asymmetry effect is quantified by $C_i[j]$ and taken into account in the noise disturbance in frame i, N_i:

$$N_i = \sum_{j=1}^{Nb} N_i[j] \cdot C_i[j] \cdot \Delta z$$

where:

$$C_i[j] = \left(\frac{PHy_i[j] + 1}{PHx_i[j] + 1} \right)^{0.2}$$

with $PHx_i[j]$ and $PHy_i[j]$ as the powers of the source and coded signals (after IRS filtering and the addition of Hoth noise), respectively, within frame i and band j. When $PHx_i[j]$ and $PHy_i[j]$ are less than 20 dB above the absolute threshold of audibility in band j (i.e. $PHx_i[j]$ and $PHy_i[j]$ are less than $100 * P_0[j]$), $C_i[j]$ is set equal to 1. The maximum value of $C_i[j]$ must be limited to 2.0.

9.5.4 Noise disturbance including silent interval processing

In PSQM, the silent intervals are taken into account using a weighting factor, W_{sil}, that depends on the context of subjective experiments. Silent frames are defined as frames for which the source signal has a total power Px'_i (i.e., $\sum_j Px'_i[j]$) below 70 dB SPL. If the global calibration factor, S_p, has been computed correctly the silence threshold is $Px'_i = 1.0*10^7$. Frames with Px'_i less than this value are considered silent.

The average noise loudnesses, N_{spav} and N_{silav}, can now be computed over active speech frames and over silent frames, respectively:

$$N_{spav} = \frac{1}{M_{sp}} \sum_{i \text{ for active speech frames}} N_i$$

$$N_{silav} = \frac{1}{M_{sil}} \sum_{i \text{ for silent frames}} N_i$$

where M_{sp} is the number of active speech frames and M_{sil} is the number of silent frames.

The influence of silent intervals depends directly on the length of these intervals. If the source speech does not contain any silent intervals, the influence is zero. If the source speech contains a certain percentage of silent frames, the influence is proportional to this percentage. Using a set of trivial boundary conditions, one can show that the correct weighting is:

$$N_{w_{sil}} = \frac{W_{sp} \cdot p_{sp}}{W_{sp} \cdot p_{sp} + p_{sil}} \cdot N_{spav} + \frac{p_{sil}}{W_{sp} \cdot p_{sp} + p_{sil}} \cdot N_{silav}$$

with p_{sil} as the fraction of silent frames, p_{sp} as the fraction of active speech frames ($p_{sil} + p_{sp} = 1.0$), W_{sil} as the weighting factor on silent intervals, $W_{sp} = \frac{1-W_{sil}}{W_{sil}}$, and $N_{w_{sil}}$ as the noise disturbance corrected with a weight factor W_{sil} for the silent interval.

This noise disturbance $N_{w_{sil}}$, called the PSQM value in the following subclauses, can be used to predict the subjectively perceived speech quality obtained in the Absolute Category Rating (ACR) method using the Listening Quality scale.

NOTES

1 The value of $N_{w_{sil}}$ should have an upper limit of 6.5.

2 For speech material having long periods of silent intervals, the weighting is different from that for speech material having only short periods of silent intervals. Furthermore, the noise in the recording of the source materials also has an influence on the silent interval weighting. For speech material having no silent intervals, the weighting is not relevant and $N_{w_{sil}}$ becomes equal to N_{spav}. A number of speech databases were examined for determining the optimal weighting on the silent intervals. These databases consisted of speech material with about 50% silent intervals. The optimal weighting that was found varied between 0.0 and 0.5 [10] [11] [12]. Determination of the value of W_{sil} for speech with silent intervals is still under study. Provisionally, a weighting of 0.2 is recommended for speech materials with about 50% silent intervals.

10 Transformation from the objective quality scale to the subjective quality scale

The output of the algorithm described in clause 9, which is called the PSQM value, indicates the degree of subjective quality degradation due to speech coding. Therefore, when estimation of subjective quality on a specific scale is not necessary, e.g. in optimizing parameters of a codec or in simply comparing the performance of codecs, the PSQM value itself is quite useful. To estimate subjective quality on quality scales such as Mean Opinion Scores (MOSs) and the equivalent-Q values, however, the PSQM value is transformed as described below.

10.1 Mean opinion scores

In subjective assessment of the performance of codecs, the ACR method using the Listening Quality scale specified in Recommendation P.800 is often used, giving subjective quality in terms of MOS. Since the relationship between the MOS and PSQM values is not necessarily the same for different languages or even for different subjective tests within a language, it is difficult to determine a unique function which transforms the PSQM value to the estimated MOS value. In practice, therefore, it is necessary to derive such transformation functions for individual languages and individual subjective tests in advance.

NOTE – The absolute value of the MOS depends on the context of the subjective experiment. The estimated MOS obtained by a predetermined transformation function predicts the subjective quality in the subjective experiment with context equivalence to those used in deriving the transformation function.

When the results are presented in the estimated-MOS domain, the transformation function from the PSQM value to the MOS value should be reported.

10.2 Equivalent-Q values

It is difficult to compare the MOSs obtained in different subjective experiments since subjective judgement is affected by the experimental settings, e.g. the range of speech quality in the experiment. Therefore, the equivalent-Q value is sometimes used as a subjective quality scale. The equivalent-Q value is determined as the Q value of MNRU defined in Recommendation P.810 for which the MOS is equivalent to that of coded speech.

In the objective measurement, the equivalent-Q value can be estimated directly from the PSQM values for coded speech and MNRU conditions, without transforming the PSQM value to the MOS (see Figure 4). When the results are presented in the estimated equivalent-Q domain, the Q versus PSQM-value characteristics illustrated in Figure 4 should be reported.

NOTE – The equivalent-Q value becomes relatively unreliable in the regions of high- and low-Q value because the Q versus PSQM curve becomes almost flat in these regions. Accordingly, care should be taken when working in the Q domain with very high- and very low-quality speech.

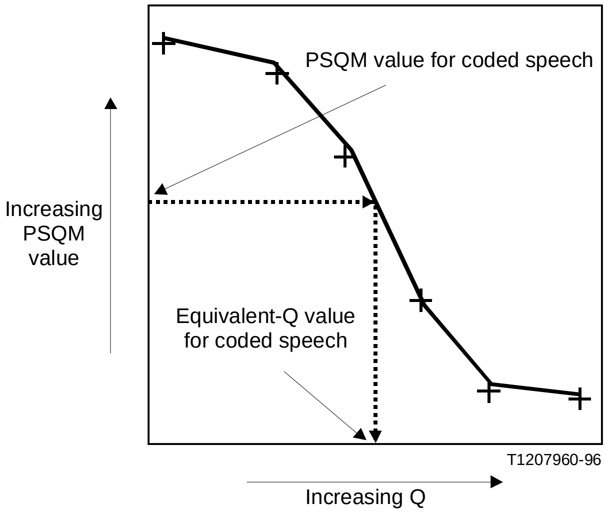

FIGURE 4/P.861

Determination of equivalent-Q value of coded speech

11 Analysis of results

The analysis of objective measurement results should be carried out based on the PSQM value, the estimated MOS, or the estimated equivalent-Q.

For each testing condition, the mean scores over male talkers, female talkers, and their average should be calculated separately and reported.

Calculation of separate standard deviations for each testing condition is not recommended. Confidence limits should be evaluated by taking into account the variation of objective quality over talkers and sentences and significance tests performed by conventional analysis-of-variance techniques.

NOTE – The statistical analyses described here are different from those in subjective assessment where the means of subjective quality are statistically evaluated by taking into account the variations over subjects as well as talkers and sentences. Since the PSQM cannot estimate the distributions of subjective votes but only the mean of them, it is impossible to perform the analysis over subjects. Estimating the distributions of subjective votes is still under study. Therefore, when the analysis over subjects is necessary, subjective experiments conforming to Recommendation P.830 should be conducted.

Bibliography

[1] NTT: Transmission performance objective evaluation model for fundamental factors, *CCITT Contribution COM XII-174*, November 1983.

[2] LALOU (J): The information index: an objective measure of speech transmission performance, *Annales des Télécommunications*, Volume 45, No. 1-2, pp. 47-65, CNET/France, 1990.

[3] BNR: Evaluation of non-linear distortion via the coherence function, *CCITT Contribution COM XII-60*, April 1982.

[4] KUBICHEK (R.F.), QUINCY (E.A.), KISER (K.L.): Speech Quality Assessment Using Expert Pattern Recognition Techniques, *Proceedings of the IEEE Pacific Rim Conference on Computers, Communication, and Signal Processing*, June 1989.

[5] Royal PTT, Netherlands: Measuring the quality of audio devices, *CCITT Contribution COM XII-114*, Geneva, December 1991.

[6] BEERENDS (J.G.), STEMERDINK (J.A.): A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation, *J. Audio Eng. Soc.*, Vol. 42, No. 3, pp. 115-123, March 1994.

[7] BEERENDS (J.G.): Modelling Cognitive Effects that Play a Role in the Perception of Speech Quality, *Speech Quality Assessment*, Workshop papers, Bochum, pp. 1-9, November 1994.

[8] CCITT Supplement No. 13 to P-Series Recommendations, *Noise Spectra clause 2, Blue Book*, Fasc. 5, ITU, Geneva 1988.

[9] ZWICKER (Feldtkeller): *Das Ohr als Nachrichtenempfänger*, *S. Hirzel Verlag,* Stuttgart, 1967.

[10] Royal PTT, The Netherlands: Correlation of a perceptual quality speech measure with the subjective quality of the CCITT LD-CELP (G.728) speech codec, *ITU-T Contribution COM 12-10*, Geneva, March 1993.

[11] Royal PTT, The Netherlands: Correlation between the PSQM and the subjective results of ITU-T 8 kbit/s 1993 speech codec test, *ITU-T Contribution COM 12-31*, Geneva, September 1994.

[12] NTT: Review of validation tests for objective speech quality measures, *ITU-T contribution COM 12-74*, Geneva, May 1996.

[13] PRESS (W.H.) *et al.*: *Numerical Recipes in C, The Art of Scientific Computing, Cambridge University Press*, Cambridge, Massachusetts, 1988.

Appendix I

Contents of floppy diskette accompanying Recommendation P.861

I.1 Introduction

In order to assist the readers of this Recommendation in the development of their own implementation of the PSQM, a floppy diskette has been included with this Recommendation. This appendix provides a description of the files on that diskette.

I.2 \src directory

This directory contains an example of a "C" code implementation of the algorithm described in this Recommendation. The code is designed to work with 16-bit linear PCM speech files sampled at 16 kHz and stored with the least significant byte (LSB) first. The following files are included in this directory:

```
psqm.c            psqm.h              psqmprot.h
psqmvals.c        readsamp.c      spchopn.c
spchread.c
```

The contents of the files are described below.

`psqm.c` This file contains the majority of the logic for the computation of the PSQM. The FFT algorithm used was not included due to copyright restrictions. That algorithm was from *Numerical Recipes in C*. (Press, W.H., *et al.* [1988] *Numerical Recipes in C, The Art of Scientific Computing,* Cambridge University Press, Cambridge, Massachusetts.)

`psqm.h` This file contains the declarations for the globally available variables used in computing the PSQM.

`psqmprot.h` This file contains the function prototypes for all functions used in computing the PSQM.

`psqmvals.c` This file contains the value assignments for the constants and arrays used in computing the PSQM.

`readsamp.c` Integer function read_samples. This function reads in a single frame of speech for processing.

`spchopn.c` This file has a function which opens the speech file and determines the start and stop points based on the algorithm described in 9.1.1/P.861.

`spchread.c` This file contains integer function speech_read. This serves as an interface to the function read_samples.

I.3 \test directory

This directory contains files for use in testing an implementation of the PSQM for accuracy. The files contained in this directory are:

`longs.cod` `longs.src` `outlong.txt`
`outshort.txt` `shorts.cod` `shorts.src`

The contents of the files are described below.

`longs.cod` A coded speech file intended to be used in calibrating the PSQM. Stored LSB first and delayed by 22 samples from the source speech file.

`longs.src` The source speech file used to create `longs.cod`. Stored LSB first.

`outlong.txt` Frame-by-frame output for the long test vector. In addition to the calibration variables (S_p and S_l), the following information is included in the file:

calculated global scaling factor (S_{global});
start and stop points for both source and coded speech files;
frame-by-frame noise disturbance (N_i);
frame-by-frame indicator of frame silence (1=silent, 0=not silent).

`outshort.txt` Step-by-step variable values for the short test vector. Requires the following values in the program:

delay = 0;
start point = 0;
stop point = 511;
S_{global} = 1.0.

The following intermediate values are provided in the file:

input sequence ($x_i[n]$, $y_i[n]$);
windowed version of input sequence ($xw_i[n]$, $yw_i[n]$);
sampled Spectrum Power Density ($Px_i[k]$, $Py_i[k]$;

local scaling factor (S_i);
sampled Pitch Power Density ($Px'_i[j]$, $Py'_i[j]$);
results of telephone-band filtering ($PFx_i[j]$, $PFy_i[j]$);
results of adding Hoth noise ($PHx_i[j]$, $PHy_i[j]$);
sampled compressed loudness density ($Lx_i[j]$, $Ly_i[j]$);
local loudness scaling factor (Sl_i);
sampled noise disturbance density ($N_i[j]$);
asymmetry effect factor ($C_i[j]$);
noise disturbance (N_i).

`shorts.cod` A coded speech file intended to be used in calibrating the PSQM. Stored LSB first and delayed by 0 samples from the source speech file.

`shorts.src` The source speech file used to create `shorts.cod`. Stored LSB first.

ITU-T RECOMMENDATIONS SERIES

Series A	Organization of the work of the ITU-T
Series B	Means of expression
Series C	General telecommunication statistics
Series D	General tariff principles
Series E	Telephone network and ISDN
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media
Series H	Transmission of non-telephone signals
Series I	Integrated services digital network
Series J	Transmission of sound-programme and television signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	Maintenance: international transmission systems, telephone circuits, telegraphy, facsimile and leased circuits
Series N	Maintenance: international sound-programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality**
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminal equipment and protocols for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks and open system communication
Series Z	Programming languages

97749

Printed in Switzerland

Geneva, 1997