INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.861

(02/98)

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

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, 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 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 revised by ITU-T Study Group 12 (1997-2000) and was approved under the WTSC Resolution No. 1 procedure on the 27th of February 1998.

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.

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.

INTELLECTUAL PROPERTY RIGHTS

The ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. The 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, the 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.

CONTENTS

Page

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

(revised in 1998)

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 – Relationship of coding technologies, experimental factors and applications to this Recommendation *(concluded)*

Coding technologies	Note
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

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

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

NOTE 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.

NOTE 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.

NOTE 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).

NOTE 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 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 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.

– 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)*.

– 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 Supplement No. 13 (1988) to the P-Series Recommendations.

3 Abbreviations

This Recommendation uses the following abbreviations:

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

This Recommendation defines the following term:

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.

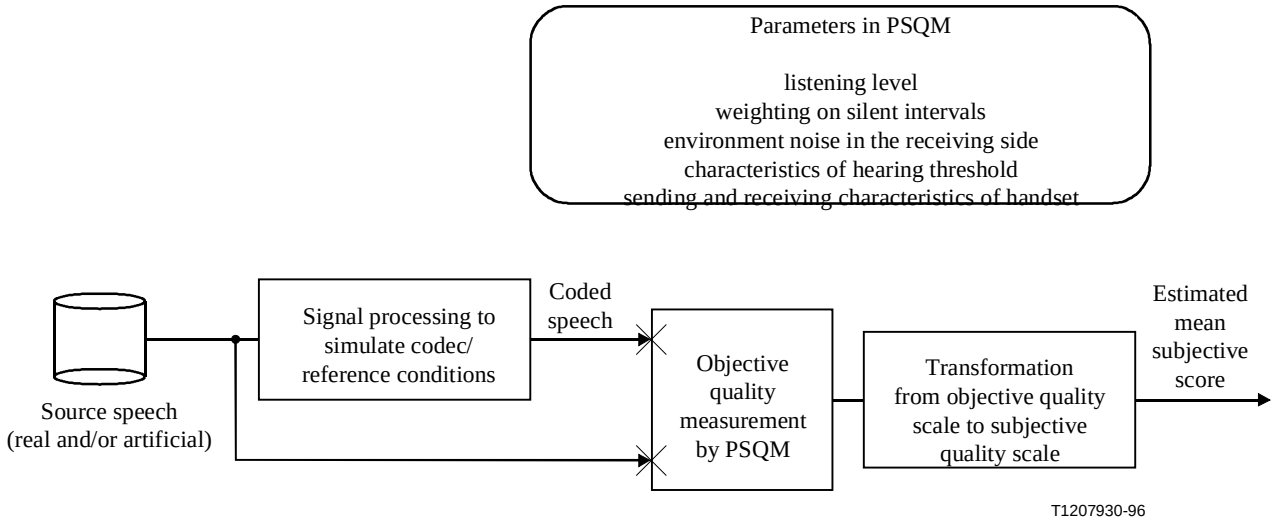

Figure 1/P.861 – Objective quality measurement procedure

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] and [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.

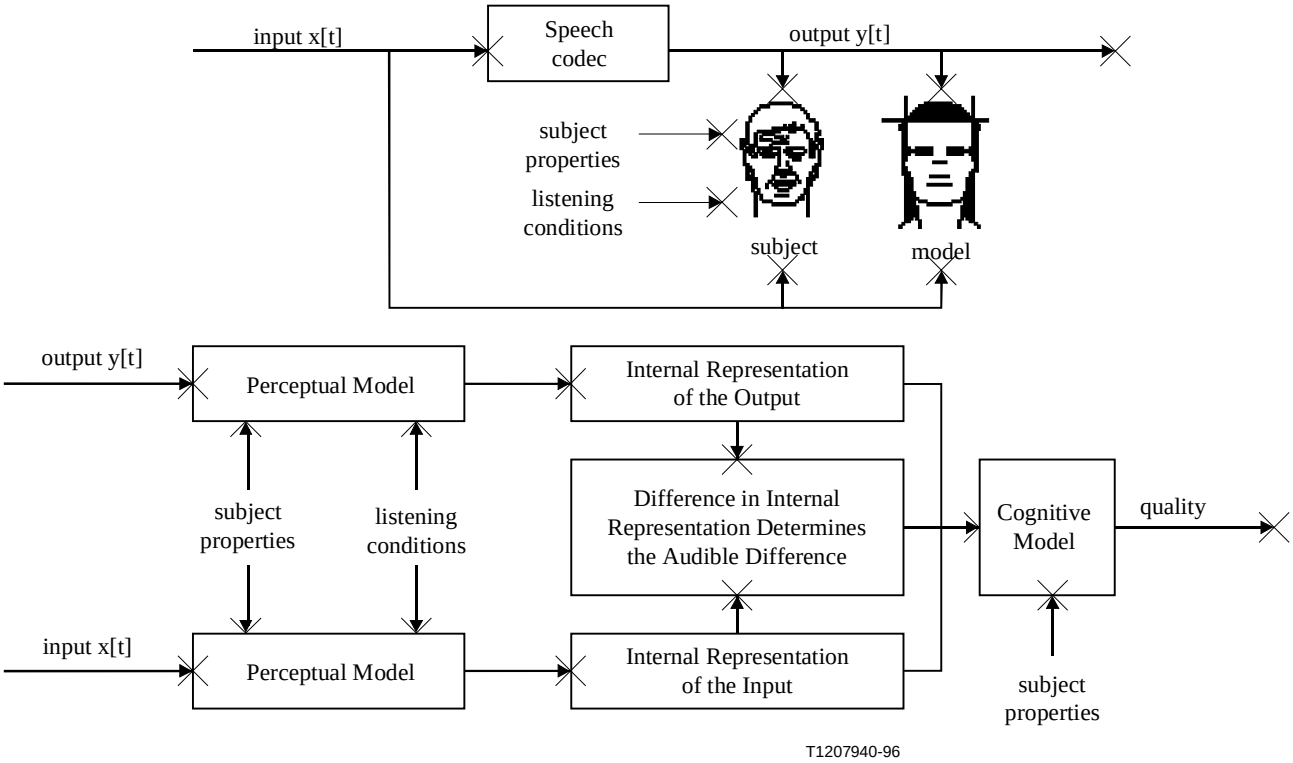

Figure 2/P.861 – Overview of the basic philosophy used in the development of the PSQM – A computer model of the subject, consisting of a perceptual and a cognitive model, is used to compare the output of the speech codec with the input

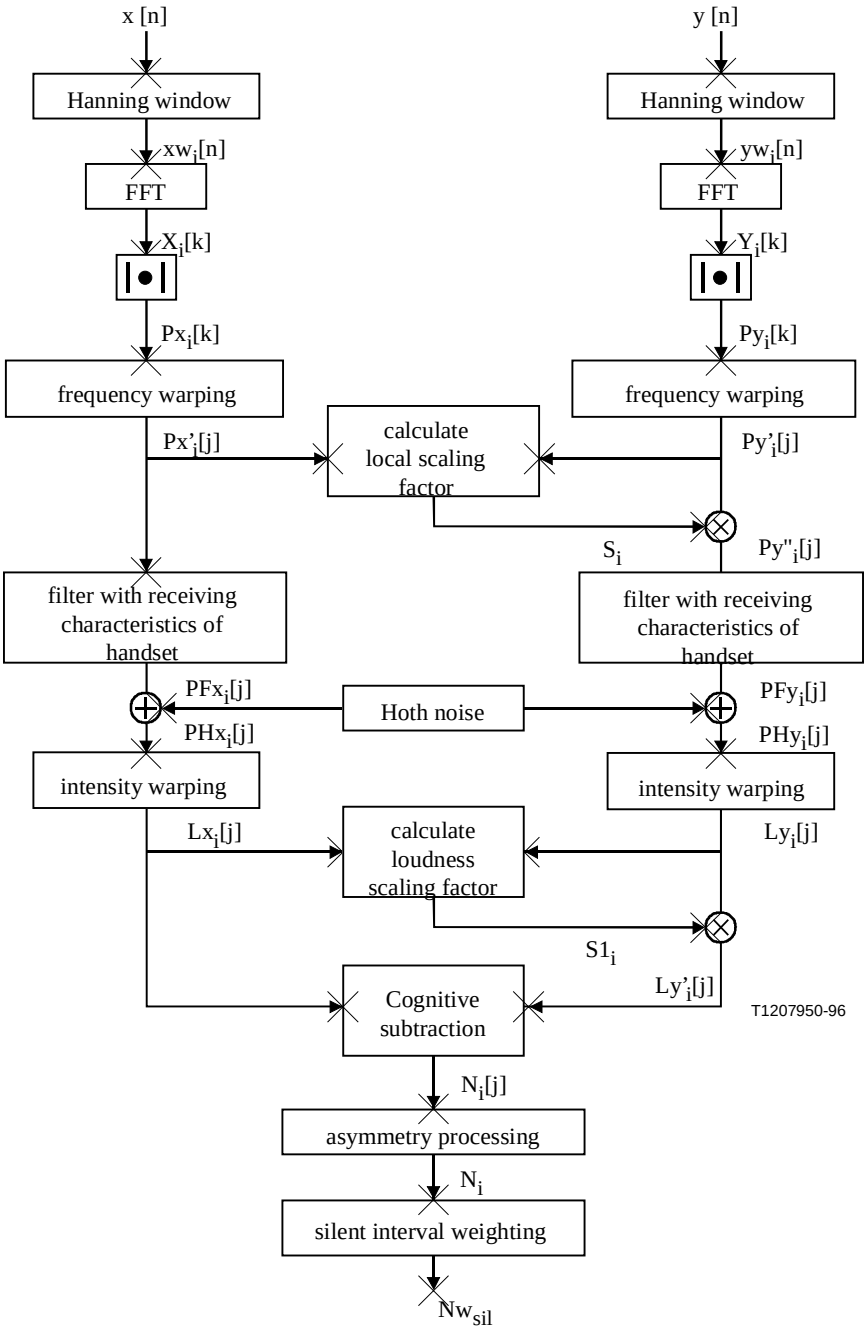

Figure 3/P.861 – Block diagram of PSQM algorithm

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 $P_0[j]$)
$\Delta f[j]$	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)

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 $yw_i[n]$

Table 3/P.861 – Variables in PSQM *(concluded)*

Name	Description
$I_f[j]$	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
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
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	1051.0	32	33	9.87E–01	3.98E+00	6.87E+02
29	1108.0	34	35	9.90E–01	3.98E+00	6.57E+02
30	1168.0	36	37	9.91E–01	3.98E+00	6.49E+02
31	1231.0	38	39	9.93E–01	3.98E+00	6.17E+02
32	1297.0	40	41	9.95E–01	4.07E+00	5.95E+02
33	1366.0	42	43	1.00E+00	4.27E+00	5.68E+02
34	1437.0	44	45	1.01E+00	4.47E+00	5.37E+02
35	1509.0	46	48	1.02E+00	4.68E+00	5.04E+02

Table 4/P.861 – Critical band frequency allocations and filter characteristics (Based on a 16 kHz sampling rate) *(concluded)*

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
36	1582.0	49	50	1.04E+00	5.01E+00	4.80E+02
37	1658.0	51	53	1.06E+00	5.37E+00	4.51E+02
38	1736.0	54	55	1.07E+00	5.62E+00	4.37E+02
39	1817.0	56	58	1.09E+00	5.89E+00	4.20E+02
40	1902.0	59	60	1.10E+00	6.31E+00	4.05E+02
41	1991.0	61	63	1.11E+00	6.61E+00	3.97E+02
42	2084.0	64	66	1.12E+00	6.92E+00	3.86E+02
43	2184.0	67	69	1.12E+00	7.24E+00	3.82E+02
44	2289.0	70	73	1.12E+00	7.59E+00	3.74E+02
45	2401.0	74	76	1.11E+00	7.76E+00	3.67E+02
46	2520.0	77	80	1.10E+00	7.94E+00	3.63E+02
47	2647.0	81	84	1.08E+00	7.94E+00	3.56E+02
48	2781.0	85	88	1.01E+00	7.94E+00	3.46E+02
49	2922.0	89	93	8.62E–01	7.94E+00	3.37E+02
50	3069.0	94	98	6.86E–01	8.13E+00	3.25E+02
51	3225.0	99	103	5.16E–01	8.13E+00	3.16E+02
52	3392.0	104	108	3.12E–01	8.32E+00	2.92E+02
53	3572.0	109	114	1.55E–01	8.32E+00	2.69E+02
54	3765.0	115	120	3.02E–02	8.32E+00	2.47E+02
55	3971.0	121	127	2.03E–03	8.32E+00	2.25E+02
56	4193.0	128	134	1.52E–04	8.32E+00	2.06E+02

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

NOTE 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:

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

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:

$$S_p = \frac{10\ 000}{\max_j(Px_i'[j])}$$

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:

$$xw_i[n] = w[n].\ x_i[n]$$
$$yw_i[n] = w[n].y_i[n]$$

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

The windowing function can be computed as follows:

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

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:

$$Py_i''[j] = S_i \,.\, Py_i'[j]$$

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:

$$PFx_i[j] = F[j].Px_i'[j]$$
$$PFy_i[j] = F[j].Py_i''[j]$$

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:

$$PHx_i[j] = H[j].PFx_i[j]$$
$$PHy_i[j] = H[j].PFy_i[j]$$

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]:

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

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

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:

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

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:

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

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.

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

NOTE 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] and [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.

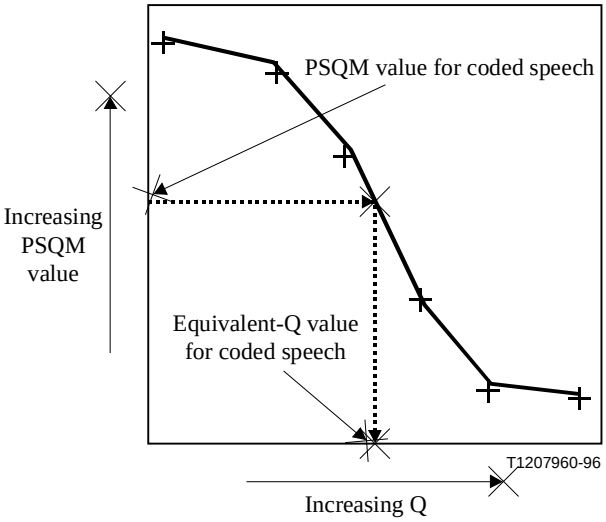

Figure 4/P.861 – Determination of equivalent-Q value of coded speech

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, test vectors and a detail of computed values have been provided on a floppy diskette included with this Recommendation. This Appendix provides a description of the files on that diskette.

I.2 \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.

APPENDIX II

Objective quality measurement of telephone-band (300-3400 Hz) speech codecs using Measuring Normalizing Blocks (MNBs)

II.1 Introduction

ITU-T has been studying objective methods for measuring the quality of coded speech. In the 1993-1996 study period, the performances of several objective quality measures for telephone-band (300-3400 Hz) speech codecs were compared. Based on the validation test results, it was concluded the method called PSQM (Perceptual Speech Quality Measure) best estimated subjective quality regardless of languages, talkers, or codecs. It was also shown that, on average, the estimation error by PSQM was comparable to the statistical reliability of subjective data. This led to the creation of Recommendation P.861 based on the PSQM algorithm.

The objective quality measure described in the body of this Recommendation is a powerful tool to estimate subjective quality of speech codecs. As described in the scope of this Recommendation (see Table 1), however, the quality factors that can be evaluated by P.861 are limited.

To extend the scope of objective quality measurement of coded speech, ITU-T is currently developing an objective quality measure which can be applied to the evaluation of speech degraded by transmission channel errors, such as bit errors in mobile networks, cell loss in ATM networks, packet loss in internet telephony, and so on.

This Appendix provides an objective quality measure whose algorithm is currently available in detail to ITU-T, and is expected to be applicable to the evaluation of bit errors and frame erasures. It is also expected that the algorithm in this Appendix will extend the scope of objective quality measurement from the viewpoint of kinds of codecs to be handled. (Table II.1). It should be noted that test factors which are inside the scope of this Recommendation should be evaluated by Recommendation P.861 itself.

Although the information on the performance of the objective quality measure in this Appendix can be found in [1], [2] and [3], more information is needed. In particular, the applicability of the measure to other kinds of channel degradation such as cell/packet loss must be investigated in comparison with other objective quality measures that have also been proposed in ITU-T [4], [5] and [6]. For these reasons, ITU-T does not recommend any new algorithms extending the scope of this Recommendation for the time being, but provides an example of objective quality measures which may have wider scope than Recommendation P.861.

ITU-T will continue to compare/validate the performance of the objective quality measures including the one described in this Appendix. This will lead to the creation of a new Recommendation which is applicable to the evaluation of transmission channel errors, expanding the scope of this current Recommendation.

II.2 Computation of the objective measure

This subclause describes the computation of an objective measure based on Measuring Normalizing Blocks (MNBs). MNBs were developed in response to the observations that listeners adapt and react differently to spectral deviations that span different time and frequency scales. Thus, for the speech quality estimation application, maximal perceptual consistency over a wide range of distortion types requires a family of analyses at multiple frequency and time scales. As spectral deviations are measured, the deviations at one scale must be removed so they are not counted again as part of the deviations at other scales. It is also observed that working from larger to smaller scales is most likely to emulate listeners' patterns of adaptation and reaction to spectral deviations. This observation has led to a hierarchical structure of MNBs.

Table II.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	3
Talker dependencies	1
Multiple simultaneous talkers	3
Bit errors and frame erasures	2
Cell and packet loss	3
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	3
Environmental noise in the sending side	3
Network information signals as input to a codec	3
Music as input to a codec	3
Delay	4
Short-term time warping of audio signal	3
Long-term time warping of audio signal	5
Temporal clipping of speech	3
Amplitude clipping of speech	3
Coding technologies	
Waveform	1
CELP and hybrids ≥ 4 kbit/s	1
CELP and hybrids < 4 kbit/s	2
VOCODERs	2
Other coders	2
Coder optimization	1
Coder evaluation	1
Coder selection	3
Network planning	6
Live network testing	7
In-service non-intrusive measurement devices	4

Notes relative to Table II.1/P.861

NOTE 1 – If your testing conditions consist of only those items labelled with the number 1, you should use the algorithm described in the body of this Recommendation.

NOTE 2 – If your testing conditions consist of only those items labelled with the number 2, or a combination of items labelled 1 and 2, the algorithm described in this Appendix should provide satisfactory results.

NOTE 3 – Insufficient information is available about the accuracy of the objective measures with regard to this variable.

NOTE 4 – The objective measures are known to provide inaccurate predictions when used in conjunction with this variable, or are otherwise not intended to be used with this variable.

NOTE 5 – The objective measure in the body of this Recommendation 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 described in this Appendix is for further study. The applicability of both objective measures when there is a small amount of wander is for further study.

NOTE 6 – With caution, the objective measures 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).

NOTE 7 – With caution, the objective measures 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 or this Appendix is not applicable (see the "Test factors" and "Coding technologies" sections of this Table).

Two types of measuring normalizing blocks are considered here. The first is the Time Measuring Normalizing Block (TMNB) and the second is the Frequency Measuring Normalizing Block (FMNB). Each of these blocks takes perceptually transformed reference [R(t,f)] and test [T(t,f)] signals as inputs and returns them and a set of measurements as outputs. These two building blocks are defined by Figures II.1 and II.2 respectively. The TMNB integrates over some frequency scale, then measures differences and normalizes the test signal at multiple times. Finally, the positive and negative portions of the measurements are integrated over time. In an FMNB the converse is true. An FMNB integrates over some time scale, then measures differences and normalizes the test signal at multiple frequencies. Finally, the positive and negative portions of the measurements are integrated over frequency. By design, both types of MNBs are idempotent. This important property is illustrated in Figure II.3 and simply means that a second pass through a given MNB will not further alter the test signal, and that second pass will result in a measurement vector of zeros. The idempotency of MNBs allows them to be cascaded and yet still measure the deviation at a given time or frequency scale once and only once.

In order to measure spectral deviations at multiple time and frequency scales, a hierarchical structure of TMNBs and FMNBs, operating at decreasing scales has been formed (Figure II.4). When used as a distance measure in conjunction with a perceptual transformation (described below), this structure appears to do a good job of emulating listeners' patterns of adaptation and reaction to spectral deviations. The structure shown results in 12 measurements. Because of the hierarchical nature of these structures, measurements from other than the top layer mean little individually, but a linear combination of the measurements has been found to be a good indicator of the perceptual distance between the two signals. The value that results from this linear combination is called Auditory Distance (AD):

$$AD = \sum_{i=1}^{N} \text{weight}_i \cdot m_i .$$

Auditory distance is a positive quantity. When the reference and test signals are similar, AD is small. As the reference and test signals move apart perceptually, AD increases. A logistic function or some other "limiter function" can be used to map AD into a finite interval. This allows AD to correlate better with subjective quality or impairment judgements, which usually cover a finite range. The weights, $weight_i$, have been selected to maximize this correlation.

NOTE – This Appendix contains sufficient information to implement the algorithms in a computer programming language. New implementations can be validated by using information available in the file ftp://ftp.its.bldrdoc.gov/dist/voice/verify.zip.

II.2.1 Input-output specifications

The input to the algorithm is a pair of speech files called reference and test. The file called reference contains a digital representation of the reference signal, which is typically the input to the Device Under Test (DUT), and is referred to as *x* in the equations. The file called test contains a digital representation of the test signal, which is typically the output from the DUT and is referred to as *y* in the equations. The sample rate is 8000 samples per second, and the recommended precision is at least 16 bits per sample. Lower precision may be used, if the user is willing to accept the associated loss of sensitivity. In addition, higher precision and higher sample rates (e.g. 16 000 samples per second) may be used if care is given to ensure that the transformation to the frequency domain produces the results specified in II.2.4. Also, if higher-sample-rate files are available, they can be down-sampled using the programs available in the Software Tools Library of ITU-T and published in Recommendation G.191 (1996). The input files must contain at least one second of telephone bandwidth speech. (Files that contain only pauses in a natural conversation are not useful.) Files used in the development of these algorithms ranged from 3 to 9 seconds in duration.

The algorithm generates a single, non-negative output value called Auditory Distance (AD). AD is an estimate of the perceptual distance between the reference and test signals. Thus, when the DUT and the test set-up are transparent, the reference and test signals will be identical, and AD will be zero. As the DUT introduces more and more distortion, the reference and test signals will move apart perceptually, and AD will increase.

II.2.2 Time delay

It is assumed that the two files have the same length, and are synchronized. That is, any delay in the DUT, or the test set-up has been removed. If these delays are known a priori, they may be removed by proper timing during data acquisition. If these delays are not known a priori, they may be estimated using the technique described in Section 7 of the ANSI T1 Standard on Multimedia Communications Delay, Synchronization, and Frame Rate Measurement (ANSI Standard T1.801.04-1997), and then removed by editing one or both of the files.

II.2.3 Signal preparation

The contents of reference are read into the vector x, and the contents of test are read into the vector y. The mean value is then removed from each of the N1 entries in each of these vectors:

$$x(i) = x(i) - \frac{1}{N1} \cdot \sum_{j=1}^{N1} x(j), \quad y(i) = y(i) - \frac{1}{N1} \cdot \sum_{j=1}^{N1} y(j), \quad 1 \le i \le N1.$$

NOTE – The notation *Array(i) = Array(i) +/–/·/÷ Normalization Factor* is used throughout this Appendix. While not mathematically consistent, it is indicative of how this algorithm might be implemented in code. The original array is modified (normalized) in the manner indicated, and the modified value is used in future computations.

This eliminates any DC component that may be present in the test and reference signals. (The DC component of a signal is inaudible.)

Next, each of the vectors is normalized to a common RMS level:

$$x(i) = x(i) \cdot \left[\frac{1}{N1} \sum_{j=1}^{N1} x(j)^2 \right]^{-1/2}, \; y(i) = y(i) \cdot \left[\frac{1}{N1} \sum_{j=1}^{N1} y(j)^2 \right]^{-1/2}, \; 1 \le i \le N1$$

This approximately removes any fixed gain in the DUT or the test set-up. Thus a fixed gain will not influence the values of AD produced by this algorithm.

II.2.4 Transformation to frequency domain

The signals are then transformed to the frequency domain using the FFT. The frame size is 16 ms (128 samples for speech sampled at 8000 Hz), and the frame overlap is 50%. Any samples beyond the final full frame are discarded. Each frame of samples is multiplied (sample-by-sample) by the length 128 Hamming window:

$$w(i) = 0.54 - 0.46 \cos\left(\frac{2\pi(i-1)}{127} \right), 1 \le i \le 128.$$

After multiplication by the Hamming window, each frame is transformed to a 128 point frequency domain vector using the FFT. For each frame, the squared-magnitude of frequency samples 1 through 65 (DC through Nyquist) are retained. The results are stored in the matrices X and Y. These matrices contain 65 rows, and N2 columns, where N2 is the number of frames that are extracted from the N1 original samples in *x* and *y*.

Note that in these matrices, rows represent the frequency axis (indexed by *i* in the algorithmic description), and columns represent the time axis (indexed by *j*).

Because FFT scaling is not standardized, care should be taken to ensure that the FFT used in this algorithm is scaled so that the following condition is met: When a frame of 128 real-valued samples, each with value 1 is input to the FFT without windowing, then the complex value in the DC bin of the FFT output must be 128 + 0·j.

II.2.5 Frame selection

Only frames that meet or exceed energy thresholds in both X and Y are used in the calculation of AD. For X, that energy threshold is set to 15 dB below the energy of the peak frame in X:

$$xenergy(j) = \sum_{i=1}^{65} X(i, j), \; xthreshold = 10^{\frac{-15}{10}} \cdot \max_j(xenergy(j)).$$

For Y, the energy threshold is set to 35 dB below the energy of the peak frame in Y:

$$yenergy(j) = \sum_{i=1}^{65} Y(i, j), \; ythreshold = 10^{\frac{-35}{10}} \cdot \max_j(yenergy(j)).$$

Frames that meet or exceed both of these energy thresholds are retained:

$$\{xenergy(j) \ge xthreshold\} \text{ AND } \{yenergy(j) \ge ythreshold\} \Rightarrow \text{frame } j \text{ is retained.}$$

If any frame contains one or more samples that are equal to zero, that frame is eliminated from both X and Y. These matrices now contain 65 rows, and N3 columns, where N3 is the number of frames that have been retained. If N3 = 0, the input files do not contain suitable signals and the algorithm is terminated.

Note that the threshold levels are different for X and Y. This is to provide for the inclusion of frames in Y that have reduced power due to a perceptually significant artifact (e.g. temporal clipping), yet still retain enough power to be successfully compared to the original speech material.

II.2.6 Perceived loudness approximation

Each of the frequency domain samples in X and Y are now logarithmically transformed to an approximation of perceived loudness:

$$X(i,j) = 10 \cdot \log_{10}(X(i,j)),\ \ Y(i,j) = 10 \cdot \log_{10}(Y(i,j)),\ \ 1 \le i \le 65, 1 \le j \le N3.$$

II.2.7 Frequency Measuring Normalizing Block (FMNB)

A FMNB is applied to X and Y at the longest available time scale, defined by the length of the input files. Four measurements are extracted and stored in the measurement vector m. These measurements cover the lower and upper band edges of telephone-band speech (where it is easiest to detect noise and changes in frequency response). Temporary vectors f1, f2, and f3 are used:

$$f1(i) = \frac{1}{N3}\sum_{j=1}^{N3} Y(i,j) - \frac{1}{N3}\sum_{j=1}^{N3} X(i,j), 1 \le i \le 65, \text{ Measure}$$

$$f2(i) = f1(i) - f1(17), 1 \le i \le 65, \text{ Normalize measurement to 1 kHz}$$

$$Y(i,j) = Y(i,j) - f2(i), 1 \le i \le 65, 1 \le j \le N3, \text{ Normalize Y}$$

$$f3(i) = \frac{1}{4}\sum_{j=1}^{4} f2(1 + 4 \cdot (i-1) + j), 1 \le i \le 16, \text{ Smooth the measurement}$$

$$m(1)=f3(1),\ m(2)=f3(2), m(3) = f3(13), m(4), \text{ Save 4 measurements}$$

II.2.8 Computing Time Measuring Normalizing Blocks

In the MNB structure, the middle portion of the band undergoes two additional levels of binary band splitting, resulting in bands that are approximately 2-3 Bark wide. The extreme top and bottom portions of the band are each treated once by a separate TMNB. Finally a residual measurement is made. There are a total of 9 TMNBs in the structure and a graphical representation is given in Figure II.4. The MNB structure generates 8 measurements in addition to those generated by the initial FMNB (described in II.2.7). Temporary variables $t0$, $t1$, ... $t9$ are used.

TMNB-0 (Bottom of band, 1.9 Bark Wide):

$$t0(j) = \frac{1}{5}\sum_{i=2}^{6} Y(i,j) - \frac{1}{5}\sum_{i=2}^{6} X(i,j), 1 \le j \le N3, \text{ Measure}$$

$$Y(i,j) = Y(i,j) - t0(j), 2 \le i \le 6, 1 \le j \le N3, \text{ Normalize } Y$$

$$m(5) = \frac{1}{N3}\sum_{j=1}^{N3} \max(t0(j), 0), \text{ Save positive portion of measurement}$$

TMNB-1 (Middle of band, top layer, 10 Bark wide):

$$t1(j) = \frac{1}{36}\sum_{i=7}^{42} Y(i,j) - \frac{1}{36}\sum_{i=7}^{42} Y(i,j), 1 \le j \le N3, \text{ Measure}$$

$$Y(i,j) = Y(i,j) - t1(j), 7 \le i \le 42, 1 \le j \le N3, \text{ Normalize } Y$$

$$m(6) = \frac{1}{N3}\sum_{j=1}^{N3} \max(t1(j),0), \text{ Save positive portion of measurement}$$

$$m(7) = -\frac{1}{N3}\sum_{j=1}^{N3} \min(t1(j),0), \text{ Save negative portion of measurement}$$

TMNB-2 (Top of Band, 3 Bark wide):

$$t2(j) = \frac{1}{23}\sum_{i=43}^{65} Y(i,j) - \frac{1}{23}\sum_{i=43}^{65} X(i,j), 1 \le j \le N3, \text{ Measure}$$

$$Y(i,j) = Y(i,j) - t2(j), 43 \le i \le 65, 1 \le j \le N3, \text{ Normalize } Y$$

$$m(8) = \frac{1}{N3}\sum_{j=1}^{N3} \max(t2(j),0), \text{ Save positive portion of measurement}$$

TMNB-3 (Middle of band, middle layer, 5 Bark wide):

$$t3(j) = \frac{1}{12}\sum_{i=7}^{18} Y(i,j) - \frac{1}{12}\sum_{i=7}^{18} X(i,j), 1 \le j \le N3, \text{ Measure}$$

$$Y(i,j) = Y(i,j) - t3(j), 7 \le i \le 18, 1 \le j \le N3, \text{ Normalize } Y$$

$$m(9) = \frac{1}{N3}\sum_{j=1}^{N3} \max(t3(j),0), \text{ Save positive portion of measurement}$$

TMNB-4 (Middle of band, middle layer, 5 Bark wide):

$$t4(j) = \frac{1}{24}\sum_{i=19}^{42} Y(i,j) - \frac{1}{24}\sum_{i=19}^{42} X(i,j), 1 \le j \le N3, \text{ Measure}$$

$$Y(i,j) = Y(i,j) - t4(j), 19 \le i \le 42, 1 \le j \le N3, \text{ Normalize } Y$$

TMNB-5 (Middle of band, bottom layer, 2.5 Bark wide):

$$t5(j) = \frac{1}{5}\sum_{i=7}^{11} Y(i,j) - \frac{1}{5}\sum_{i=7}^{11} X(i,j), 1 \le j \le N3, \text{ Measure}$$

$$Y(i,j) = Y(i,j) - t5(j), 7 \le i \le 11, 1 \le j \le N3, \text{ Normalize } Y$$

$$m(10) = \frac{1}{N3}\sum_{j=1}^{N3} \max(t5(j),0), \text{ Save positive portion of measurement}$$

TMNB-6 (Middle of band, bottom layer, 2.5 Bark wide):

$$t6(j) = \frac{1}{7}\sum_{i=12}^{18} Y(i,j) - \frac{1}{7}\sum_{i=12}^{18} X(i,j), 1 \le j \le N3, \text{ Measure}$$

$$Y(i,j) = Y(i,j) - t6(j), 12 \le i \le 18, 1 \le j \le N3, \text{ Normalize } Y$$

TMNB-7 (Middle of band, bottom layer, 2.5 Bark wide):

$$t7(j) = \frac{1}{10}\sum_{i=19}^{28} Y(i,j) - \frac{1}{10}\sum_{i=19}^{28} X(i,j), 1 \le j \le N3, \text{ Measure}$$

$$Y(i,j) = Y(i,j) - t7(j), 19 \le i \le 28, 1 \le j \le N3, \text{ Normalize } Y$$

$$m(11) = \frac{1}{N3}\sum_{j=1}^{N3} \max(t7(j),0), \text{ Save positive portion of measurement}$$

TMNB-8 (Middle of band, bottom layer, 2.5 Bark wide):

$$t8(j) = \frac{1}{14}\sum_{i=29}^{42} Y(i,j) - \frac{1}{14}\sum_{i=29}^{42} X(i,j), 1 \le j \le N3, \text{ Measure}$$

$$Y(i,j) = Y(i,j) - t8(j), 29 \le i \le 42, 1 \le j \le N3, \text{ Normalize } Y$$

Residual Measurement:

$$t9(i,j) = Y(i,j) - X(i,j), 1 \le i \le 65, 1 \le j \le N3, \text{ Measure residual}$$

$$m(12) = \frac{1}{N364}\sum_{i=2}^{65}\sum_{j=1}^{N3} \max(t9(i,j),0), \text{ Save positive portion of residual measurement}$$

Note that if two measurements (positive part and negative part) were retained for each of the 9 TMNBs in the structure, a total of 18 measurements would result. The hierarchical nature of the MNB structure, along with the idempotency property of the MNB leads to linear dependence among these 18 measurements. Only 7 linearly independent MNB measurements are available. These combine with the single residual measurement and the 4 FMNB measurements for a total of 12 measurements.

II.2.9 Linear combination of measurements for MNB structure

The 12 measurements now are combined linearly to generate an AD value. The weights used in this linear combination are given in Table II.2:

$$\text{AD} = \sum_{i=1}^{12} m(i) \cdot weight_i .$$

Note that when all 12 measurements are zero, AD is zero.

Table II.2/P.861 – Weights for MNB structure linear combination

i	$weight_i$
1	0.0000
2	–0.0023
3	–0.0684
4	0.0744
5	0.0142
6	0.0100
7	0.0008
8	0.2654
9	0.1873
10	2.2357
11	0.0329
12	0.0000
NOTE – It may seem inefficient to compute measurement values and multiply them by a weighting factor of zero (i.e. $weight_1$ and $weight_{12}$). However, there are two reasons for maintaining the computation of these measurements. First, the set of measurements, m_1 to m_{12}, represents the complete set of linearly independent measurement values available as a part of the MNB structure. Second, this set of weightings, $weight_i$, represents the best weightings for the applications and conditions set forth in Table II.1. For other applications and conditions, a different weighting function may be required that has non-zero values for these measurements.	

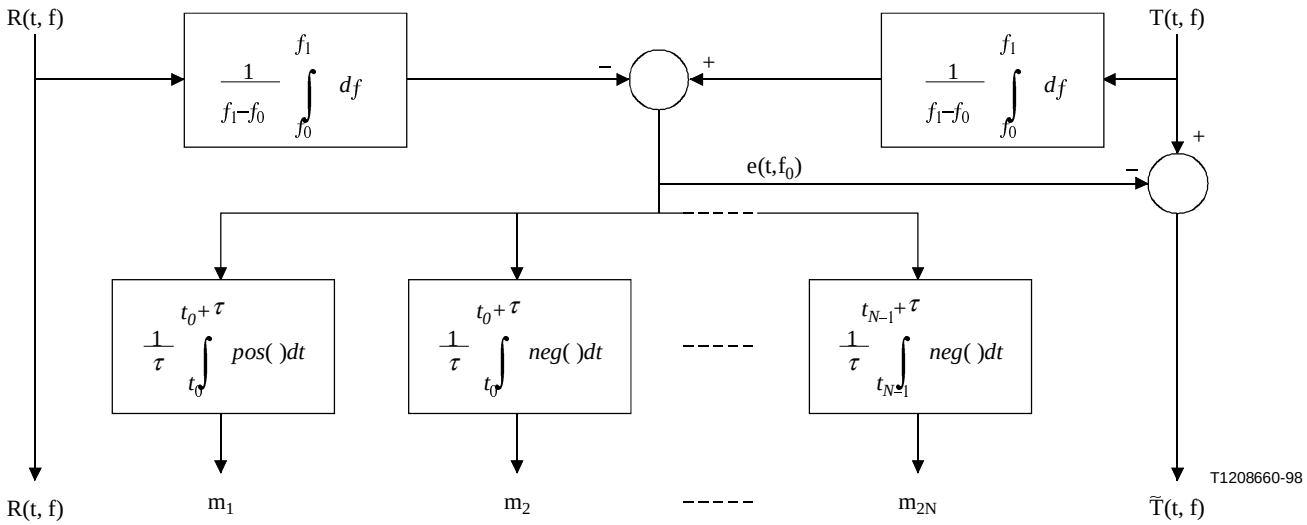

Figure II.1/P.861 – Time Measuring Normalizing Block (TMNB)

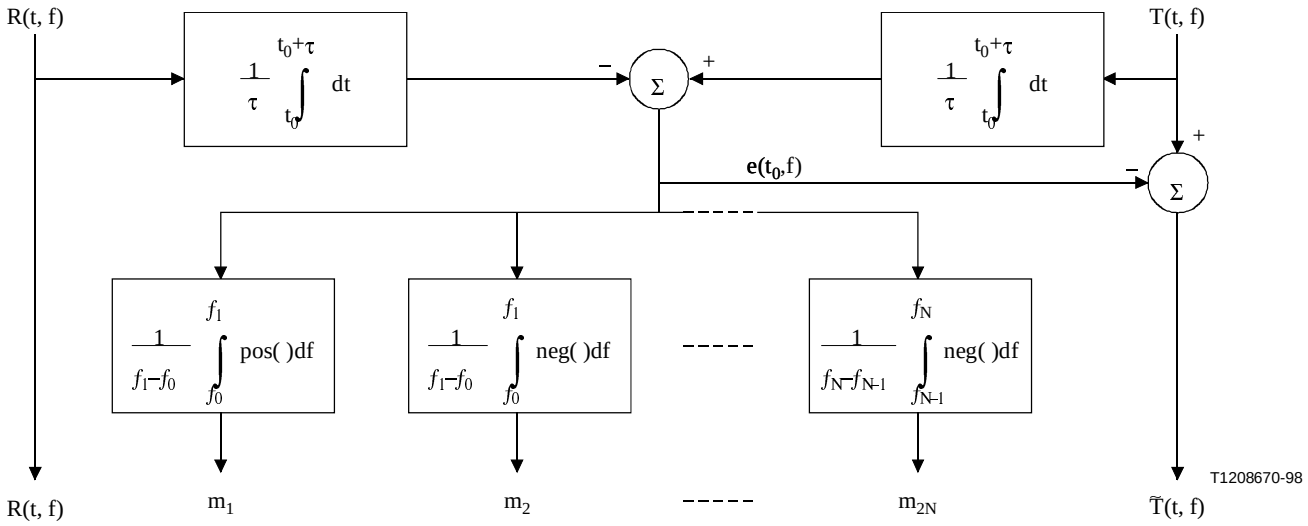

Figure II.2/P.861 – Frequency Measuring Normalizing Block (FMNB)

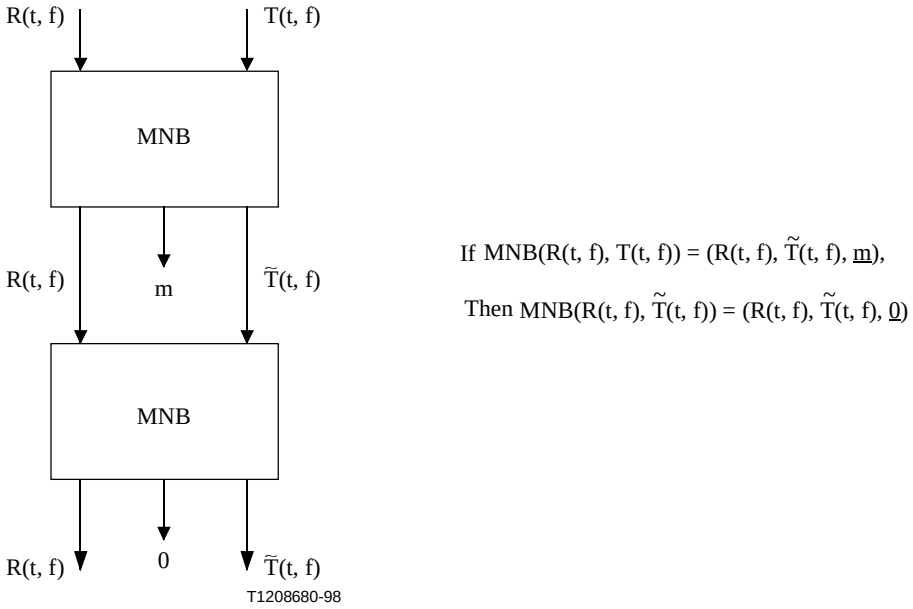

Figure II.3/P.861 – MNBs are Idempotent

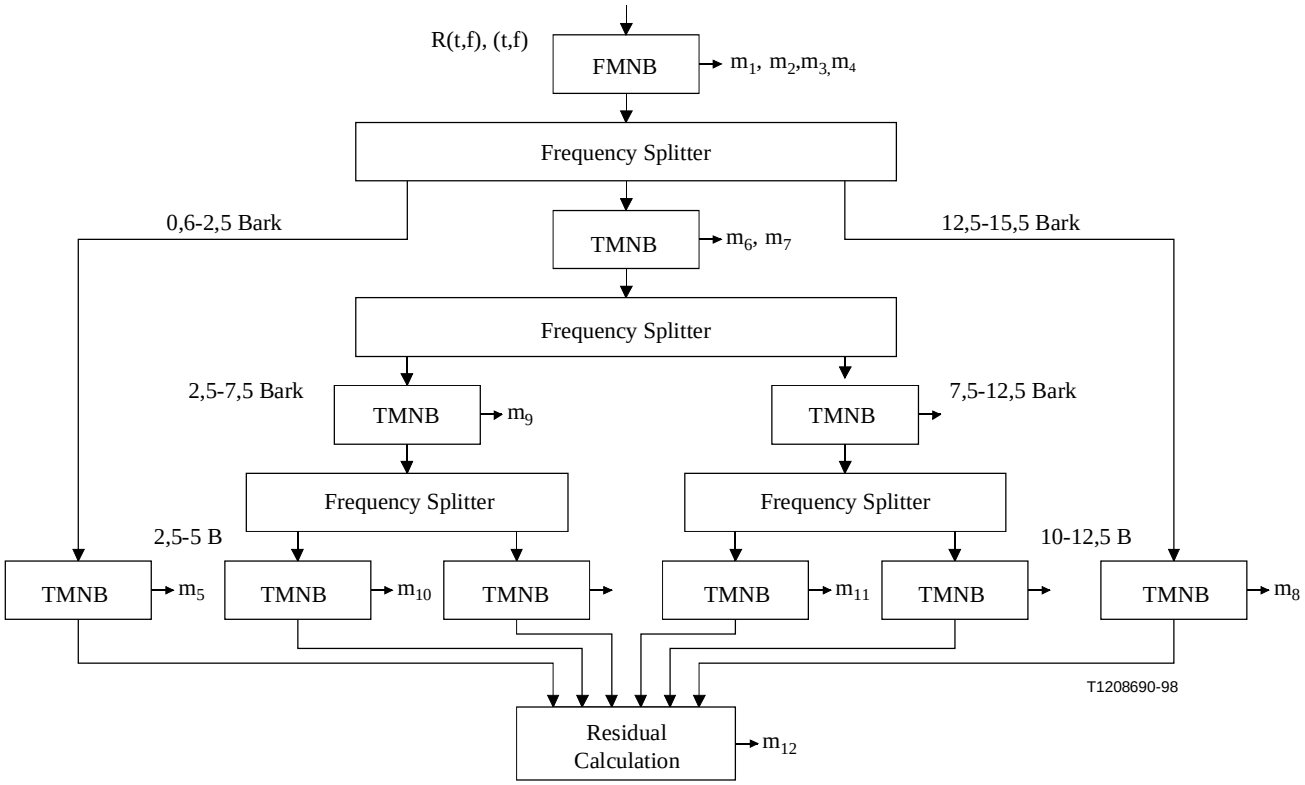

Figure II.4/P.861 – MNB hierarchical structure

Bibliography

[1] ITU-T Contribution COM 12-25: Summary of available performance results for the MNB voice quality measurement algorithm, USA, February 1998.

[2] ITU-T Contribution COM 12-26: Comparison of two methods for objective assessment of the quality of coded speech, *AT&T* February 1998.

[3] ITU-T Contribution COM 12-27: Results of objective voice quality experiments, USA, February 1998.

[4] ITU-T Contribution COM 12-20: Improvement of the P.861 Perceptual Speech Quality Measure, KPN, February 1998.

[5] ITU-T Contribution COM 12-21: An experimental investigation of the accumulation of perceived error in time-varying speech distortions, *BT*, February 1998.

[6] ITU-T Contribution COM 12-34: TOSQA – Telecommunication objective speech quality assessment, FRG, February 1998.

ITU-T RECOMMENDATIONS SERIES

Series A	Organization of the work of the ITU-T
Series B	Means of expression: definitions, symbols, classification
Series C	General telecommunication statistics
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	TMN and network 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, telephone installations, local line networks**
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks and open system communication
Series Y	Global information infrastructure
Series Z	Programming languages