

RECOMMENDATION ITU-R BS.1387-1

Method for objective measurements of perceived audio quality

(1998-2001)

The ITU Radiocommunication Assembly,

considering

- a) that conventional objective methods (e.g. for measuring signal-to-noise ratio and distortion) are no longer adequate for measuring the perceived audio quality of systems which use low bit-rate coding schemes or which employ analogue or digital signal processing;
- b) that low bit-rate coding schemes are rapidly being deployed;
- c) that not all implementations conforming to a specification or standard guarantee the best quality achievable with that specification or standard;
- d) that formal subjective assessment methods are not suitable for continuous monitoring of audio quality, e.g. under operational conditions;
- e) that objective measurement of perceived audio quality may eventually complement or supersede conventional objective test methods in all areas of measurement;
- f) that objective measurement of perceived audio quality may usefully complement subjective assessment methods;
- g) that, for some applications, a method which can be implemented in real time is necessary,

recommends

- 1** that for each application listed in Annex 1 the method given in Annex 2 be used for objective measurement of perceived audio quality.

Foreword

This Recommendation specifies a method for objective measurement of the perceived audio quality of a device under test, e.g. a low bit-rate codec. It is divided into two Annexes. Annex 1 gives the user a general overview of the method and includes four Appendices. Appendix 1 describes applications and test signals. Appendix 2 lists the Model Output Variables and discusses limitations of use and accuracy. Appendix 3 gives the outline of the model while Appendix 4 describes the principles and characteristics of objective perceptual audio quality measurement methods in general.

Annex 2 provides the implementer with a detailed description of the method using two versions of the psycho-acoustic model that were developed during the integration phase where six models were combined. In Appendix 1 of Annex 2 the validation process of the objective measurement method is described. Appendix 2 of Annex 2 gives an overview of all the databases that were used in the development and validation of the method.

TABLE OF CONTENTS

	Page
Foreword	1
Table of contents	2
Annex 1 – Overview.....	7
1 Introduction.....	7
2 Applications	7
3 Versions	8
4 The subjective domain	8
5 Resolution and accuracy	10
6 Requirements and limitations	10
Appendix 1 to Annex 1 – Applications	10
1 General.....	10
2 Main applications.....	11
2.1 Assessment of implementations.....	11
2.2 Perceptual quality line up.....	11
2.3 On-line monitoring.....	11
2.4 Equipment or connection status	12
2.5 Codec identification	12
2.6 Codec development.....	12
2.7 Network planning.....	12
2.8 Aid to subjective assessment.....	13
2.9 Summary of applications	13
3 Test signals	13
3.1 Selection of natural test signals.....	14
3.2 Duration	15
4 Synchronization	15
5 Copyright issues.....	15
Appendix 2 to Annex 1 – Output variables.....	15
1 Introduction.....	15
2 Model Output Variables.....	15
3 Basic Audio Quality.....	16
4 Coding Margin	17
5 User requirements	17

	Page
Appendix 3 to Annex 1 – Model outline	17
1 Audio processing	18
1.1 User-defined settings.....	18
1.2 Psycho-acoustic model.....	18
1.3 Cognitive model.....	19
Appendix 4 to Annex 1 – Principles and characteristics of objective perceptual audio quality measurement methods.....	20
1 Introduction and history.....	20
2 General structure of objective perceptual audio quality measurement methods	21
3 Psycho-acoustical and cognitive basics	21
3.1 Outer and middle ear transfer characteristic	22
3.2 Perceptual frequency scales	22
3.3 Excitation	23
3.4 Detection	24
3.5 Masking.....	25
3.6 Loudness and partial masking.....	25
3.7 Sharpness	25
3.8 Cognitive Processing.....	26
4 Models incorporated	27
4.1 DIX.....	27
4.2 NMR.....	28
4.3 OASE	28
4.4 Perceptual Audio Quality Measure (PAQM).....	29
4.5 PERCEVAL	30
4.6 POM	30
4.7 The Toolbox Approach	31
Annex 2 – Description of the Model	32
1 Outline	32
1.1 Basic Version	33
1.2 Advanced Version.....	33
2 Peripheral Ear Model	34
2.1 FFT-based Ear Model	34
2.1.1 Overview	34
2.1.2 Time Processing	35
2.1.3 FFT	35
2.1.4 Outer and middle ear	36

2.1.5	Grouping into critical bands	36
2.1.6	Adding internal noise	43
2.1.7	Spreading	43
2.1.8	Time domain spreading	45
2.1.9	Masking Threshold	45
2.2	Filter bank-based ear model	46
2.2.1	Overview	46
2.2.2	Subsampling	47
2.2.3	Setting of Playback Level	47
2.2.4	DC-rejection-filter	47
2.2.5	Filter Bank	48
2.2.6	Outer and middle ear filtering	50
2.2.7	Frequency domain spreading	51
2.2.8	Rectification	53
2.2.9	Time domain smearing (1) – Backward masking	53
2.2.10	Adding of internal noise	53
2.2.11	Time domain smearing (2) – Forward masking	53
3	Pre-processing of excitation patterns	54
3.1	Level and pattern adaptation	54
3.1.1	Level adaptation	54
3.1.2	Pattern adaptation	55
3.2	Modulation	56
3.3	Loudness	56
3.4	Calculation of the error signal	57
4	Calculation of Model Output Variables	57
4.1	Overview	57
4.2	Modulation difference	58
4.2.1	RmsModDiff _A	58
4.2.2	WinModDiff1 _B	59
4.2.3	AvgModDiff1 _B and AvgModDiff2 _B	59
4.3	Noise Loudness	59
4.3.1	RmsNoiseLoud _A	60
4.3.2	RmsMissingComponents _A	60
4.3.3	RmsNoiseLoudAsym _A	60
4.3.4	AvgLinDist _A	60
4.3.5	RmsNoiseLoud _B	60
4.4	Bandwidth	60
4.4.1	Pseudocode	61
4.4.2	BandwidthRef _B and BandwidthTest _B	61

4.5	Noise-to-mask ratio.....	62
4.5.1	Total NMR_B	62
4.5.2	Segmental NMR_B	62
4.6	Relative Disturbed Frames $_B$	62
4.7	Detection Probability	62
4.7.1	Maximum filtered probability of detection ($MFPD_B$)	64
4.7.2	Average distorted block (ADB_B)	64
4.8	Harmonic structure of error.....	64
4.8.1	EHS_B	65
5	Averaging.....	65
5.1	Spectral averaging.....	65
5.1.1	Linear average	65
5.2	Temporal averaging	65
5.2.1	Linear average	66
5.2.2	Squared average	66
5.2.3	Windowed average	66
5.2.4	Frame selection.....	67
5.3	Averaging over audio channels.....	67
6	Estimation of the perceived basic audio quality	67
6.1	Artificial neural network.....	68
6.2	Basic Version	68
6.3	Advanced Version.....	70
7	Conformance of Implementations.....	71
7.1	General	71
7.2	Selection.....	71
7.3	Settings for the conformance test.....	71
7.4	Acceptable tolerance interval.....	72
7.5	Test items	72
	Appendix 1 to Annex 2 – Validation process.....	73
1	General.....	73
2	Competitive phase.....	74
3	Collaborative phase.....	75
4	Verification	75
4.1	Comparison of SDG and ODG values	76
4.2	Correlation	76
4.3	Absolute Error Score (AES)	79
4.4	Comparison of ODG versus the confidence interval	80
4.5	Comparison of ODG versus the tolerance interval	84

5	Selection of the optimal model versions.....	86
5.1	Pre-selection criteria based on correlation.....	86
5.2	Analysis of number of outliers.....	87
5.3	Analysis of severeness of outliers.....	88
6	Conclusion	89
	Appendix 2 to Annex 2 – Descriptions of the reference databases.....	89
1	Introduction.....	89
2	Items per database.....	91
3	Experimental conditions	91
3.1	MPEG90.....	92
3.2	MPEG91.....	92
3.3	ITU92DI.....	92
3.4	ITU92CO	92
3.5	ITU93	92
3.6	MPEG95.....	93
3.7	EIA95	93
3.8	DB2	93
3.9	DB3	93
3.10	CRC97.....	94
4	Items per condition for DB2 and DB3	94
4.1	DB2	94
4.2	DB3	97
	Glossary.....	97
	Abbreviations	98
	References	99
	Bibliography.....	100

ANNEX 1

Overview

1 Introduction

Audio quality is one of the key factors when designing a digital system for broadcasting. The rapid introduction of various bit-rate reduction schemes has led to significant efforts in establishing and refining procedures for subjective assessments, simply because formal listening tests have been the only relevant method for judging audio quality. The experience gained was the foundation for Recommendation ITU-R BS.1116, which then became the basis for most listening tests of this type.

Since subjective quality assessments are both time consuming and expensive, it is desirable to develop an objective measurement method in order to produce an estimate of the audio quality. Traditional objective measurement methods, like Signal-to-Noise-Ratio (S/N) or Total-Harmonic-Distortion (THD) have never really been shown to relate reliably to the perceived audio quality. The problems become even more evident when the methods are applied on modern codecs which are both non-linear and non-stationary.

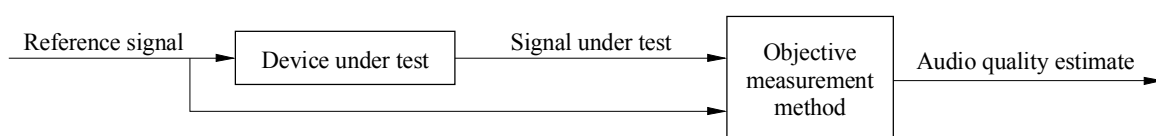
A number of methods for making objective perceptual measurements of perceived audio quality have been introduced during the last decade. But none of the methods were thoroughly validated, and consequently neither standardized nor widely accepted. In 1994, ITU-R identified an urgent need to establish a standard in this area and the work was initiated. An open call for proposals was issued and the following six candidates for measurement methods were received: Disturbance Index (DIX), Noise-to-Mask Ratio (NMR), Perceptual Audio Quality Measure (PAQM), Perceptual Evaluation (PERCEVAL), Perceptual Objective Measure (POM) and The Toolbox Approach. The methods are described in Appendix 4 to Annex 1.

The measurement method in this Recommendation is the result of a process where the performance of each of the above six methods was studied, and the most promising tools extracted and integrated into one single method. The recommended method has been carefully validated at a number of test sites. It has proven to generate both reliable and useful information for several applications. One must, however, keep in mind that the objective measurement method in this Recommendation is not generally a substitute for arranging a formal listening test.

2 Applications

The basic concept for making objective measurements with the recommended method is illustrated in Fig. 1 below.

FIGURE 1
Basic concept for making objective measurements



1387-01

The measurement method in this Recommendation is applicable to most types of audio signal processing equipment, both digital and analogue. It is, however, expected that many applications will focus on audio codecs.

The following 8 classes of applications have been identified:

TABLE 1

Applications

	Application	Brief description	Version
1	Assessment of implementations	A procedure to characterize different implementations of audio processing equipment, in many cases audio codecs	Basic/Advanced
2	Perceptual quality line up	A fast procedure which takes place prior to taking a piece of equipment or a circuit into service	Basic
3	On-line monitoring	A continuous process to monitor an audio transmission in service	Basic
4	Equipment or connection status	A detailed analysis of a piece of equipment or a circuit	Advanced
5	Codec identification	A procedure to identify the type and implementation of a particular codec	Advanced
6	Codec development	A procedure which characterizes the performance of the codec in as much detail as possible	Basic/Advanced
7	Network planning	A procedure to optimize the cost and performance of a transmission network under given constraints	Basic/Advanced
8	Aid to subjective assessment	A tool for screening critical material to include in a listening test	Basic/Advanced

3 Versions

In order to achieve an optimal fit to different cost and performance requirements, the objective measurement method recommended in this Recommendation has two versions. The Basic Version is designed to allow for a cost-efficient real-time implementation, whereas the Advanced Version has a focus on achieving the highest possible accuracy. Depending on the implementation, this additional accuracy increases the complexity approximately by a factor of four compared to the Basic Version.

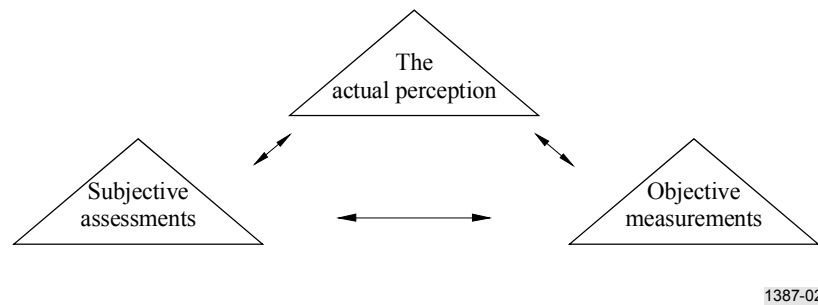
Table 1 gives some guidance on which version to apply for each of the applications.

4 The subjective domain

Formal subjective listening tests, e.g. those based on Recommendation ITU-R BS.1116, are carefully designed to come as close as possible to a reliable estimate of the judgement of the audio quality. One could, however, not expect the result from a subjective listening test to fully reflect the actual perception. Figure 2 illustrates the imperfections implicit in both the subjective and the objective domain.

It is obviously not possible to validate an objective method directly. Instead, objective measurement methods are validated against subjective listening tests.

FIGURE 2
Validation concepts

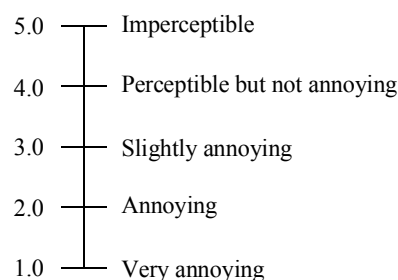


The objective measurement method in this Recommendation has been focused on applications which are normally assessed in the subjective domain by applying Recommendation ITU-R BS.1116. The basic principle of that particular test method can be briefly described as follows: the listener can select between three sources (“A”, “B” and “C”). The known Reference Signal is always available as source “A”. The hidden Reference Signal and the Signal Under Test are simultaneously available but are “randomly” assigned to “B” and “C”, depending on the trial.

The listener is asked to assess the impairments on “B” compared to “A”, and “C” compared to “A”, according to the continuous five-grade impairment scale. One of the sources, “B” or “C”, should be indiscernible from source “A”; the other one may reveal impairments. Any perceived differences between the reference and the other source must be interpreted as an impairment. Normally, only one attribute, “Basic Audio Quality”, is used. It is defined as a global attribute that includes any and, all detected differences between the reference and the Signal Under Test.

The grading scale shall be treated as continuous with “anchors” derived from the ITU-R five-grade impairment scale given in Recommendation ITU-R BS.562 as shown below.

FIGURE 3
The ITU-R five-grade impairment scale



The analysis of the results from a subjective listening test is in general based on the Subjective Difference Grade (SDG) defined as:

$$SDG = Grade_{Signal\ Under\ Test} - Grade_{Reference\ Signal}$$

The SDG values should ideally range from 0 to –4, where 0 corresponds to an imperceptible impairment and –4 to an impairment judged as very annoying.

5 Resolution and accuracy

The Objective Difference Grade (ODG) is the output variable from the objective measurement method and corresponds to the SDG in the subjective domain. The resolution of the ODG is limited to one decimal. One should however be cautious and not generally expect that a difference between any pair of ODGs of a tenth of a grade is significant. The same remark is valid when looking at results from a subjective listening test.

There is no single figure which fully describes the accuracy of the objective measurement method. Instead, one has to consider a number of different figures of merit. One of them is the correlation between SDGs and ODGs. It is important to understand that there is no guarantee that the correlation will exceed a pre-defined value. The performance of the measurement method will most likely vary with, for example, the type and level of the introduced degradation.

Another figure of merit of interest is the number of outliers. An outlier is defined as a measured value which does not meet a pre-defined tolerance scheme. According to the user requirements, the measurement method should deliver the highest possible accuracy for the upper end of the grading scale (i.e. high audio quality). Consequently, the obtained accuracy is allowed to be lower in the middle and lower range of the grading scale.

Although the correlation normally gives a good estimate of the accuracy of the objective measurement method, it is important to keep in mind that even a relatively high correlation figure could hide an unacceptable performance (from the perspective of outliers) of a measurement method.

A third figure of merit which has been used during the validation process is the Absolute Error Score (AES), which reflects the average of the relation between the size of the SDG confidence interval and the distance between SDG and ODG.

More details about the expected performance of the measurement method as well as the performance during the validation process can be found in Appendix 1 to Annex 2.

6 Requirements and limitations

The signal from the Device Under Test and the Reference Signal must be time aligned with an accuracy of 24 samples during the complete measurement interval. The synchronization mechanism is not a part of this Recommendation and is expected to be different from implementation to implementation.

APPENDIX 1

TO ANNEX 1

Applications

1 General

This Appendix provides the definitions and specific requirements for the main applications for which the recommended objective measurement method of perceived audio quality is intended.

Some of the applications require a real-time implementation of the objective measurement method while, for other applications, non real-time measurement is sufficient. For real-time implementations, it is recommended that the maximum delay through the measurement equipment does not exceed 200 ms and more than 1 s is not acceptable.

Furthermore, a distinction has to be made between on-line and off-line measurements. In off-line measurements, the measurement procedure has full access to the equipment or connection while on-line measurement implies that a programme is running, which must not be interrupted by the measurement.

2 Main applications

2.1 Assessment of implementations

Broadcasters, network operators and others have a need to assess different implementations of equipment, in particular audio codecs, when selecting such equipment for purchase or when acceptance tests are conducted.

For these kind of applications, high accuracy is required especially to assess small impairments and correctly rank different implementations. Concerning output variables, a simple output such as the ODG is sufficient for users, but developers of audio codecs can do a more thorough analysis by using a suitable set of Model Output Variables (MOVs).

Both model versions can be used, but the Advanced Version is recommended.

2.2 Perceptual quality line up

This is a fast procedure which takes place prior to taking a piece of equipment or a circuit into service. The aim is to check functionality and quality. Measurement equipment will be handled by operational staff. Any kinds of distortion may be present.

Real-time measurement is required. Test signals or pre-defined audio signals may be used. The ODGs should be properly displayed and should be given at least twice a second or, if a special test signal is used, directly after the end of the test signal.

Using the Basic Version is sufficient.

2.3 On-line monitoring

This is a continuous process, which takes place during an ongoing audio transmission. The programme must not be interrupted by the measurement procedure. Hence, the programme signal itself or a pre-defined audio fragment must be used for the measurement. The latter may be a station signal or a jingle. The measurement equipment will be handled by operational staff.

Real-time measurement is required. The ODGs must be properly displayed and should be given at least twice a second or directly after the end of the pre-defined signal. A display of MOVs is not desired.

Using the Basic Version is sufficient.

2.4 Equipment or connection status

To ensure the functionality of audio connections or equipment, an extensive quality check is required from time to time. In contrast to on-line monitoring or perceptual line up, this application requires a check of several technical parameters.

The measurement system should give detailed information about the influence of the equipment or connection status on perceived audio quality by displaying the complete set of MOVs in addition to the ODGs. Real-time measurement is not required.

Use of the Advanced Version is recommended.

2.5 Codec identification

In order to identify codecs (different algorithms or different implementations of the same algorithm), the measurement system must be able to store, retrieve and compare patterns of characteristics. Similarity between patterns can be taken as a measure of the similarity of different codec implementations. Such a procedure is used to identify the type and implementation of a particular codec.

The measurement system must record as much information about the patterns as possible. The consideration of the ODGs only may not provide enough information.

Use of the Basic Version is sufficient, even though real-time measurement is not required.

NOTE 1 – Only little experience with the recommended method exists. Furthermore, no single measure for the similarity between patterns is yet defined.

2.6 Codec development

For this application the measurement method must characterize the performance of the codec under test as accurately and with as much detail as possible, in particular for small impairments.

Continuous monitoring tests require real-time processing which is not necessarily supported by the Advanced Version. However, small degradations and detailed information will require the Advanced Version. The measurement system must be able to display the outputs at the same rate at which they are calculated. Direct access to the history of the outputs over a period of 4 s is desired.

Use of the Advanced Version is recommended. However, for real-time measurement the Basic Version is sufficient. Real-time as well as non real-time and frame-by-frame analysis is required. Any severe distortion has to be indicated, e.g. by a peak-display. Access to the complete set of MOVs is desirable.

2.7 Network planning

The planning of networks requires assessment of the expected quality at various points during the planning process. A software simulation of the network components, which allows combining different audio processing stages, can be used to examine different configurations in order to optimize the audio quality. In a later stage, the actual audio processing components can be tested in the chosen configuration.

Network planning is done by system engineers who should retrieve detailed information about the influence of network characteristics on the audio quality. Ranking of different possible network configurations should be based on a suitable set of MOVs depending on the specific application of the network. A display of the ODGs only is thus not sufficient. Real-time measurement is not required for the assessment in this application.

Both model versions can be used, but the Advanced Version is recommended.

2.8 Aid to subjective assessment

The objective measurement method provides a tool for screening critical audio material to be used in subjective listening tests. The whole set of MOVs can be used for the categorization of the critical material.

The highest possible accuracy is required and use of the Advanced Version is recommended. However, real-time measurement is desirable in order to reduce the time required to select the critical material.

2.9 Summary of applications

Table 2 summarizes the requirements on the measurement method for the main applications.

TABLE 2

Requirements on the measurement method

	Application	Category	Real-time	Min, ROV ⁽¹⁾ (Hz)	On/Off-line	Model version
1	Assessment of implementations	Diagnostic	No	–	Off	Both
2	Perceptual quality line up	Operational	Y/N	2	Off	Basic
3	On-Line monitoring	Operational	Yes	2	On	Basic
4	Equipment or connection status	Diagnostic	Y/N	–	On/Off	Advanced
5	Codec identification	Diagnostic	No	–	Off	Both
6	Codec development	Development	Y/N	–	Off	Both
7	Network planning	Development	Y/N	–	Off	Both
8	Aid to subjective assessment	Development	Y/N	–	Off	Advanced

⁽¹⁾ Rate of output values (per second).

3 Test signals

Test signals can be divided into two groups: natural and synthetic. The list of natural test signals provided here consists of critical audio sequences already used in listening tests performed, both by ITU-R and by other organizations, for the evaluation of audio quality. The signals have to be available both at the transmitting site and at the measurement site. Thus, memory in the measurement device is required.

The synthetic signals are mathematically defined and can be varied in a controlled way. These signals can be generated at the transmitting and measurement sites. Extra memory is not required in the measurement device. Due to the nature of such signals it is difficult, if not impossible, to derive subjective gradings for them. Therefore, the measurement method has not been validated against subjective results for these signals.

3.1 Selection of natural test signals

Table 3 provides a list with a subset of test signals that were used during the verification procedure that led to this Recommendation. The type of artefacts, which these signals typically unveil due to low bit-rate coding, is also indicated.

TABLE 3

List with a subset of test signals

No.	Item	File name	Remarks
1	Castanets	cas	(1)
2	Clarinet	cla	(2)
3	Claves	clv	(1)
4	Flute	flu	(2)
5	Glockenspiel	glo	(1), (2), (5)
6	Harpsichord	hrp	(1), (2), (4)
7	Kettle drum	ket	(1)
8	Marimba	mar	(1)
9	Piano Schubert	pia	(2)
10	Pitch Pipe	pip	(4)
11	Ry Cooder	ryc	(2), (4)
12	Saxophon	sax	(2)
13	Bag Pipe	sb1	(2), (4), (5)
14	Speech Female Engl.	sfe	(3)
15	Speech Male Engl.	sme	(3)
16	Speech Male German	smg	(3)
17	Snare drums	sna	(1)
18	Soprano Mozart	sop	(4)
19	Tamborine	tam	(1)
20	Trumpet	tpt	(2)
21	Triangle	tri	(1), (2), (5)
22	Tuba	tub	(2)
23	Susanne Vega	veg	(3), (4)
24	Xylophone	xyl	(1), (2)

- (1) Transients: pre-echo sensitive, smearing of noise in temporal domain.
- (2) Tonal structure: noise sensitive, roughness.
- (3) Natural speech (critical combination of tonal parts and attacks): distortion sensitive, smearing of attacks.
- (4) Complex sound: stresses the Device Under Test.
- (5) High bandwidth: stresses the Device Under Test, loss of high frequencies, programme-modulated high frequency noise.

3.2 Duration

The duration of a natural test signal should be about the same as if it were to be used in a listening test. The duration is typically in the order of 10 to 20 s. It is very likely that the critical part of the test signal, which unveils most of the artefacts, is limited to only a short part of the duration.

The duration of synthetic test signals should be long enough to stress the codec under test, which may contain a buffer for the coded audio signal. Considering these buffer lengths and the time constants present in the measurement method, the duration of each single test item in a sequence shall be more than 500 ms. The duration can be limited to such a short value because it is not expected that these signals will be used in subjective listening tests.

4 Synchronization

For the measurement procedure, the Signal Under Test and the Reference Signal shall be synchronized in time. This applies both for natural and synthetic test signals.

5 Copyright issues

The test signals given in Table 3 can be used free of copyright only for measuring purposes together with the method for objective measurements, described in Annex 2 of this Recommendation.

NOTE 1 – Clearance of copyright has to be obtained for all sequences, mainly from the EBU (EBU SQAM disc).

APPENDIX 2

TO ANNEX 1

Output variables

1 Introduction

The objective measurement method described in this Recommendation measures audio quality and outputs a value intended to correspond to perceived audio quality. The measurement method models fundamental properties of the auditory system. Several intermediate stages model physiological and psycho-acoustical effects.

These intermediate outputs can be used to characterize artefacts. The parameters are called Model Output Variables (MOV). The final stage of the measurement model combines the MOV values to produce a single output value that directly corresponds to an expected result from a subjective quality assessment.

2 Model Output Variables

Table 4 contains a description of the MOVs used to predict the objective difference grades. Subscripts_A are derived from the filter bank part of the model, while subscripts_B are derived from the FFT part of the model. The objective difference grades can be predicted either from the FFT

part only (Basic Version) or from a combination of FFT and filter bank parts (Advanced Version). Averaging is always performed over time.

3 Basic Audio Quality

The most well-known parameter from subjective listening tests is Basic Audio Quality (BAQ). BAQ is measured as a Subjective Difference Grade (SDG) which is calculated as the grade given to the reference subtracted from the grade given to the Signal Under Test in a subjective test¹. The SDG normally has a negative value. The corresponding output parameter from the model is called the Objective Difference Grade (ODG). Mapping of the MOVs to an ODG is based on a large number of reliable test items, see Annex 2, Appendix 2.

TABLE 4

Description of the Model Output Variables

Model Output Variable	Description
$WinModDiff_B$	<i>Windowed averaged difference in modulation (envelopes) between Reference Signal and Signal Under Test</i>
$AvgModDiff1_B$	<i>Averaged modulation difference</i>
$AvgModDiff2_B$	<i>Averaged modulation difference with emphasis on introduced modulations and modulation changes where the reference contains little or no modulations</i>
$RmsModDiff_A$	<i>Rms value of the modulation difference</i>
$RmsMissingComponents_A$	<i>Rms value of the noise loudness of missing frequency components, (used in $RmsNoiseLoudAsym_A$)</i>
$RmsNoiseLoud_B$	<i>Rms value of the averaged noise loudness with emphasis on introduced components</i>
$RmsNoiseLoudAsym_A$	<i>$RmsNoiseLoud_A + 0.5RmsMissingComponents_A$</i>
$AvgLinDist_A$	<i>A measure for the average linear distortions</i>
$BandwidthRef_B$	<i>Bandwidth of the Reference Signal</i>
$BandwidthTest_B$	<i>Bandwidth of the output signal of the device under test</i>
$TotNMR_B$	<i>logarithm of the averaged Total Noise to Mask Ratio</i>
$RelDistFrames_B$	<i>Relative fraction of frames for which at least one frequency band contains a significant noise component</i>
$AvgSegmNMR_B$	<i>the Segmentally Averaged logarithm of the Noise to Mask Ratio</i>
$MFPD_B$	<i>Maximum of the Probability of Detection after low pass filtering</i>
ADB_B	<i>Average Distorted Block (=Frame), taken as the logarithm of the ratio of the total distortion to the total number of severely distorted frames</i>
EHS_B	<i>Harmonic structure of the error over time</i>

The ODG is the objectively measured parameter that corresponds to the subjectively perceived quality. As the task of the listener in a listening test is to assess the BAQ of a test item, the ODG is also a measure of BAQ.

¹ See Recommendation ITU-R BS.1116.

4 Coding Margin

Another parameter which in the future may prove to be valuable is Coding Margin (CM), a way of describing inaudible artefacts. Subjective Coding Margin (SCM) may be assessed by amplifying the artefacts until they become audible for a test person. SCM describes the headroom to the threshold of audibility of artefacts.

In order to find the threshold, the artefacts have to be amplified or attenuated during the listening test. A suitable method is the difference method. The difference signal of the time synchronous original and coded signal is amplified and added to the original signal. Detection of the threshold of audibility is best performed with a forced choice method. SCM is obtained by averaging the threshold values for amplification or attenuation obtained from the test persons. Negative CM values represent audible artefacts while positive CM values represent inaudible artefacts. Unlike BAQ, Coding Margin is a measure of when (at what level) artefacts become audible and not how annoying the artefacts are. The definition and validation of the method to measure the SCM is described in [Feiten, 1997].

Objective Coding Margin (OCM) is also derived from the MOVs. Presently, only a few test items for the subjective coding margin have been assessed. Mapping to OCM from the model in this Recommendation has not yet been investigated.

5 User requirements

User requirements with respect to the output variables from the measurement method differ depending on the application. For some applications, for example numbers 2 and 3 (see Appendix 1 to Annex 1), the measurement is part of an operational procedure. In these cases it is very important that the output from the method is both easy to read and easy to interpret for persons with no in-depth knowledge about the measurement technique. This is best achieved if the method outputs only **one single value** that corresponds to a perceived audio quality.

The same may apply also to other applications, for example, applications 1 and 4. However, for these, as well as for applications 5-8, more sophisticated output variables may be beneficial for users with a deeper knowledge about the mechanisms in the measurement method.

APPENDIX 3

TO ANNEX 1

Model outline

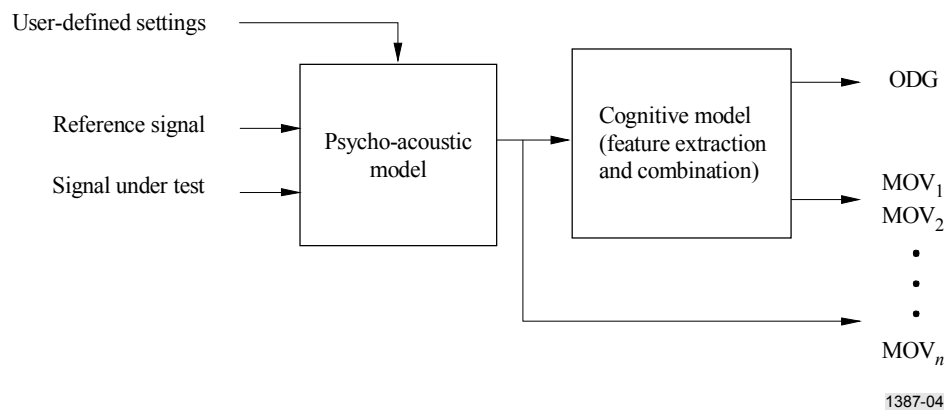
According to Recommendation ITU-R BS.1116, an SDG is obtained for an audio test item in a listening test, and the mean SDG over a number of listeners represents the item's subjective quality. The item may contain different types of audio distortions, so variations in quality are integrated over time. Therefore, prediction of the SDG based on physical measurements requires an accurate model of the peripheral auditory system as well as cognitive aspects of audio quality judgements.

The recommended model for objective measurement produces a number of Model Output Variables (MOVs) based on comparisons between the Reference Signal and the Signal Under Test. These MOVs are mapped to an ODG using an optimization technique that minimizes the squared difference between the ODG distribution and the corresponding distribution of mean SDGs for a sufficiently large data set.

Two variations of the model are described – a DFT-based version that could be used for real-time monitoring, and another version, based on both a filter bank and the DFT, that was expected to give more accurate results. The DFT-based version is called the Basic Version, while the combined version is called the Advanced Version.

The high level structure of both the Basic Version and the Advanced Version is shown in Fig. 4.

FIGURE 4
Stages of processing implemented in the model



1 Audio processing

As in the subjective listening tests, the quality of the test signal is judged relative to the Reference Signal. Both Reference Signal and Signal Under Test (monaural or stereo signals) are transformed into a psycho-acoustical representation. These representations are compared in order to derive an ODG. These operations are performed by the processing stages shown in Fig. 4.

1.1 User-defined settings

The measurement method requires the assumed listening level as a parameter. Therefore, the user has to supply the sound pressure level in dB SPL produced by a full scale sine wave of 1019.5 Hz. In case the exact listening level is unknown it is recommended to assume a listening level of 92 dB SPL.

1.2 Psycho-acoustic model

The psycho-acoustic model transforms successive frames of the time-domain signal to a basilar membrane representation. This process begins using both a DFT and a filter bank. The DFT transforms the data to the frequency domain, and the result is mapped from the frequency scale to a pitch scale, the psycho-acoustic equivalent of frequency. In the filter bank part of the model, the frequency to pitch mapping is directly taken into account by the bandwidths and spacing of the bandpass filters.

Two different concepts are used to achieve simultaneous masking. Some MOVs are calculated using the *masked threshold concept*, whereas others are based on a *comparison of internal representations*. The first concept directly calculates a masked threshold using psycho-physical masking functions. Model Output Variables are based on the distance of the physical error signal to this masked threshold. In the comparison of internal representations, the energies of both the Signal Under Test (SUT) and the Reference Signal are spread to adjacent pitch regions in order to obtain excitation patterns. Model Output Variables are based on a comparison between these excitation patterns. Non-simultaneous masking is implemented by smearing the signal representations over time.

The absolute threshold is modelled partly by applying a frequency dependent weighting function and partly by adding a frequency dependent offset to the excitation patterns. This threshold is an approximation of the minimum audible pressure [ISO 389-7, Acoustics – Reference zero for the calibration of audiometric equipment – Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions, 1996].

The main outputs of the psycho-acoustic model are the excitation and the masked threshold as a function of time and frequency. The output of the model at several levels is available for further processing.

1.3 Cognitive model

The cognitive model condenses the information from a sequence of frames produced by the psycho-acoustic model. The most important sources of information for making quality measurements are the differences between the Reference Signal and the Signal Under Test in both the frequency and pitch domain. In the frequency domain, the spectral bandwidths of both signals are measured, as well as the harmonic structure in the error. In the pitch domain error measures are derived from both the excitation envelope modulation and the excitation magnitude.

The calculated features are weighted, so that their combination results in an ODG that is sufficiently close to the SDG for the particular audio distortion of interest. The Basic Version uses 11 features to produce an ODG, while the Advanced Version uses 5 features. The optimization was performed using the back-propagation neural network learning algorithm (see Annex 2, § 6). Training data consisted of all of Databases 1 and 2, and part of Database 3. Generalization test data were obtained from the remainder of Database 3 and all of the CRC97 data set (see Appendix 2 to Annex 2).

APPENDIX 4

TO ANNEX 1

Principles and characteristics of objective perceptual audio quality measurement methods**1 Introduction and history**

The digital transmission and storage of audio signals are increasingly based on data reduction algorithms, which are adapted to the properties of the human auditory system and particularly rely on masking effects. Such algorithms do not aim mainly at minimizing the distortions but rather attempt to handle these distortions in a way that they are perceived as little as possible. The quality of these perceptual coders can no longer be assessed by conventional measurement methods, which normally determine the overall value of the distortion. An example which is often mentioned to illustrate these limitations is the so-called 13 dB miracle: Superimposed noise with a spectral structure adapted to that of the audio signal is almost inaudible even if the resulting unweighted S/N declines to 13 dB.

For this reason the evaluations of perceptual codecs require listening tests in order to assess the audio quality. Sufficient reliability and repeatability of listening tests require a large expenditure of time and work.

Objective measurement schemes that incorporate properties of the human auditory system can help to overcome these problems. This idea was first published by [Schroeder *et al*, 1979]. In this paper, which is mainly about speech coding, the measurement scheme “noise loudness (NL)” is described.

In this paper, the perceived loudness of the noise signal of the speech codec, which is the difference between its input and output signal, is estimated for each time frame of approximately 20 ms. If the noise signal is completely masked, the perceived loudness is zero. Partial masking reduces the loudness of the non-masked noise signal. The masked threshold used is optimized for tone-masking noise and the final speech degradation is calculated for each frame. No summary of the total quality of a speech sample is computed.

In 1985 Karjalainen published the measurement scheme “Auditory Spectral Difference (ASD)” [Karjalainen, 1985]. He started with several ideas from Schroeder, Atal and Hall but replaced the frame based analysis by a filter bank with overlapping filters, changed the way the absolute threshold is included and added a model for temporal masking. Both input signals to the measurement scheme are processed in exactly the same way, producing a kind of internal representation. These internal representations are compared to each other to explain perceived differences between input and output signal of a speech coding scheme. No summary of the total quality of a speech sample is computed. The temporal resolution of ASD is better adapted to the properties of the human auditory system but increases the complexity of the algorithm.

In 1987 Brandenburg published the measurement scheme “Noise to Mask Ratio (NMR)” [Brandenburg, 1987], which was intended to be used as a tool for the development of audio coding schemes. The complexity of the scheme was reduced compared to NL by calculating the spreading on perceptual bands using a spreading function that was designed as a worst-case curve. The masked threshold used is optimized for noise-masking-tone. A simple scheme of modelling post-masking and several ways to evaluate the perceived quality of longer excerpts of audio were added. This scheme was the first one implemented in real-time hardware.

In 1989 Moore and Glasberg [Moore, 1989] presented a perceptual model but did not present a way to judge the perceived quality of impaired audio signals.

2 General structure of objective perceptual audio quality measurement methods

All perceptual measurement schemes work with two input signals: one is called the Reference Signal (REF), the other the Signal Under Test (SUT). In situations where the reference cannot be transmitted to the measurement equipment, but the signal is well known, the Reference Signal can be an internal reference stored in the measurement equipment itself. It is essential, that the input signals are time-aligned.

Incorporating psycho-acoustics into measurement schemes can be done in two different ways. The first possibility is very similar to the structure of audio coding schemes: the Reference Signal is used to calculate an estimate of the actual masked threshold (see below). The difference between the Signal Under Test and the Reference Signal is compared to this masked threshold. This method is called “masked threshold concept” and is used in Noise Loudness and NMR. The difference between the input signals can be calculated either in the time domain or as the difference between the short-time energy spectra. The latter provides a better robustness against time-alignment errors but decreases the temporal resolution. The difference in the time domain usually is too sensitive to phase distortions and is therefore not used anymore.

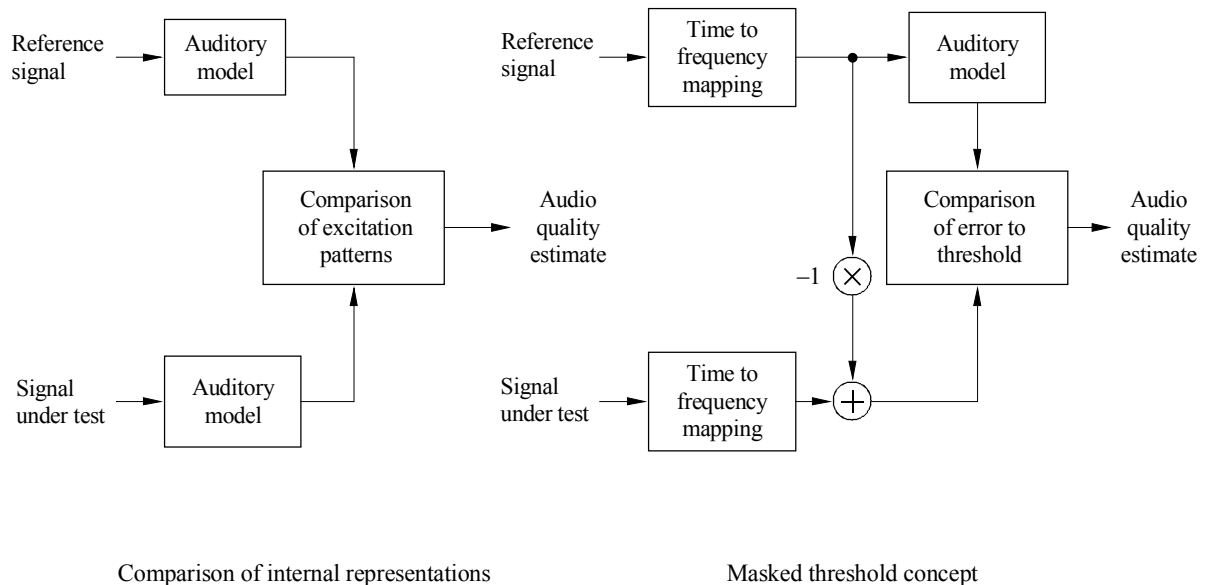
The second approach is closer to the physiological processes in the human auditory system: a so-called internal representation of both the Reference Signal and the Signal Under Test is calculated. This internal representation is an estimate of the information that is available to the human brain for comparison of signals. This method is called “comparison of internal representations” and is used in ASD.

3 Psycho-acoustical and cognitive basics

This section discusses the properties of the human auditory system that are the most prominent in the evaluation of the perceived quality of audio signals. The main emphasis is on how these properties may be modelled.

FIGURE 5

Psycho-acoustic concepts used in different approaches in perceptual measurement schemes



1387-05

3.1 Outer and middle ear transfer characteristic

In general, sound signals have to pass the outer and middle ear until they come to the inner ear where the sound detection and analysis processes are performed. The outer and middle ear perform a band pass filter operation on the input signal. Noise which is present in the auditory nerve, together with noise caused by the flow of blood, is added to the input signal. The amplitude of this noise increases with low frequencies. The outer and middle ear transfer function together with the internal noise limit the ability to detect small audio signals, and have the most influence on the absolute threshold of hearing.

3.2 Perceptual frequency scales

The receptors of sound pressure in the human ear are the hair-cells. They are located in the inner ear, more precisely in the cochlea. In the cochlea, a frequency to position transform is performed. The position of the maximum excitation depends on the frequency of the input signal. Each hair-cell at a given position on the cochlea is responsible for an overlapping range on the frequency scale. The perceptual impression of pitch is correlated with a constant distance of hair-cells.

Depending on the psycho-acoustic experiment used, different transform functions from frequency to pitch have been found:

In [Zwicker and Feldtkeller, 1967] a table is given which splits the frequency scale in Hz into 24 non-overlapping bands, the so-called critical bands. The upper cut-off frequencies of these bands are given in Table 5. The Table also contains a definition of the Bark-scale: 1 Bark corresponds to 100 Hz, 24 Bark corresponds to 15 500 Hz.

TABLE 5

Critical band scale as defined by Zwicker

Critical band	1	2	3	4	5	6	7	8	9	10	11	12
Upper cut-off frequency (Hz)	100	200	300	400	510	630	770	920	1 080	1 270	1 480	1 720

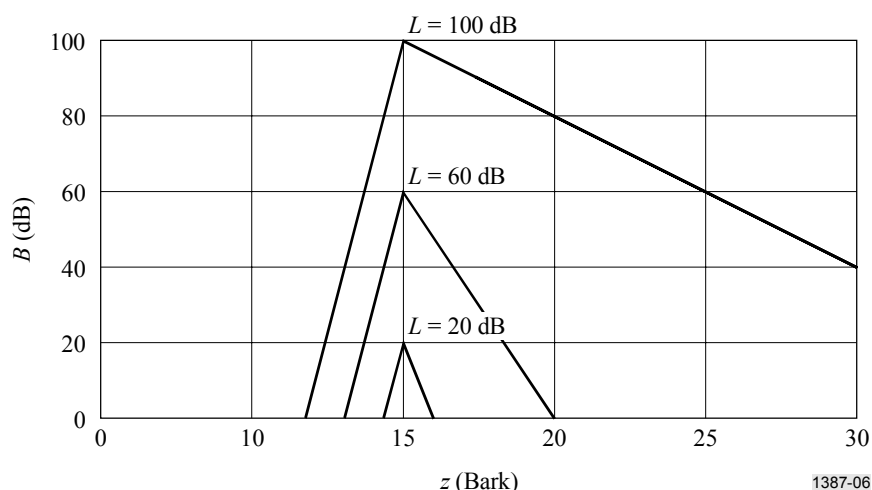
Critical band	13	14	15	16	17	18	19	20	21	22	23	24
Upper cut-off frequency (Hz)	2 000	2 320	2 700	3 150	3 700	4 400	5 300	6 400	7 700	9 500	12 000	15 500

Several approximations to the Bark scale were found in the past. A detailed discussion of different scales can be found in [Cohen and Fielder, 1992]. In the context of objective measurement of perceived audio quality, the best results were achieved using the Bark scale.

3.3 Excitation

Each hair-cell reacts to a range of frequencies that can be described by a filter characteristic. The slope of the filters can be expressed best on a perceptual scale as described above. The shape of the filters on such a scale is nearly independent of the centre frequency. The lower slope of the excitation is independent of the level L of the input signal (about 27 dB/Bark). The upper slope is steeper for lower levels than for higher levels of the input signal (–5 to –30 dB/Bark). This steep characteristic is caused by a feedback mechanism between two different kinds of hair-cells and needs some time to settle. Therefore the best auditory frequency resolution is achieved for stationary signals several milliseconds after the onset of the signal. The excitation patterns of signals consisting of several components are added in a non-linear way.

FIGURE 6
Level dependencies of excitation according to Terhardt [1979]



After exposure to a signal the hair-cells and the neural processing need some time to recover until full sensitivity is reached again. The duration of the recovery process depends on the level and the duration of the signal and can last up to several hundred milliseconds. High level signals are processed faster than low level signals on the way between hair-cell and brain. Therefore, the onset of a loud signal can mask a preceding softer signal.

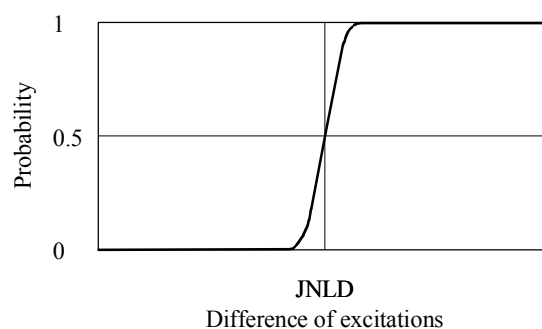
Another approach to model excitation is based on the ERB scale [Moore, 1986]. This approach uses the so-called ROEX filters [Moore, 1986]. In the context of objective measurement of perceived audio quality, better results were achieved with models based on [Zwicker and Feldtkeller, 1967] and [Terhardt, 1979].

3.4 Detection

The excitations of different audio signals are transferred to the human brain. There are three different kinds of memory that differ by the degree of detail and by the duration that the information is present: long term memory, short term memory and ultra-short term memory. In the context of listening tests, the ultra-short term memory plays the most prominent role. Most details of a signal are preserved if the duration of an audio excerpt is less than five to eight seconds depending on the listener and the audio excerpt. This is taken into account in the assessment procedure defined in Recommendation ITU-R BS.1116 where subjects are allowed to select very short parts of an audio excerpt to listen to more closely. At the detection threshold the probability of detection is 50%. Around the threshold, the probability of detection of differences increases smoothly from 0% to 100%.

The Just-Noticeable Level Difference (JNLD) is the detection threshold of level differences. The JNLD is influenced by the level of the input signals. For small signals, large differences are required for detection (level: 20 dBSPL, JNLD: 0.75 dB). For loud signals the sensitivity to small differences is much higher (level: 80 dBSPL, JNLD: 0.2 dB). These numbers are based on amplitude modulation experiments.

FIGURE 7
Principle of detection probability



1387-07

3.5 Masking

A signal that is clearly audible if presented alone can be completely inaudible in the presence of another signal, the masker. This effect is called masking and the masked signal is called the maskee. Two situations have to be distinguished:

— *Simultaneous masking*

In this situation masker and maskee are presented at the same time and are quasi-stationary. If the masker has a discrete bandwidth, the threshold of hearing is raised even for frequencies below or above the masker. The amount of masking depends on the structures of the masker and the maskee. In the situation where a noise-like signal is masking a tonal signal, the amount of masking is almost frequency independent. If the sound pressure level of the maskee is about 5 dB below the level of the masker it becomes inaudible. In the situation where a tonal signal is masking a noise like signal, the amount of masking depends on the frequency of the masker. It can be estimated by the formula

$\left(15.5 + \frac{z}{\text{Bark}}\right)$ dB, where z is the critical band rate of the masker. In addition, at high signal

levels non-linear effects reduce the masked threshold near the masker. Similar effects occur with tone-masking-tone. The masked thresholds of several signals add in a non-linear way. In general the resulting masked threshold is above the masked threshold produced by each individual signal.

— *Temporal masking*

In this situation, masker and maskee are presented at different times. Shortly after the decay of a masker the masked threshold is closer to simultaneous masking of this masker than to the absolute threshold. Depending on the duration of the masker, the decay time of the threshold can be between 5 ms (masker: Gaussian impulse with a duration of about 0.05 ms) and more than 150 ms (masker: pink noise with a duration of 1 s). Weak signals just before louder signals are masked. The duration of this backward masking effect is about 5 ms. If the maskee is just above the threshold it is not perceived before the masker but as a change of the masker. Backward masking shows large deviations from listener to listener.

3.6 Loudness and partial masking

The perceived loudness of audio signals depends on their frequency, their duration, and their sound pressure level. Due to auto-masking the loudness of a complex signal is less than the sum of the loudness of all its components. In the context of audio quality measurement, the loudness of the unwanted distortion added to the Reference Signal, the noise loudness, is reduced by the partial masking caused by the Reference Signal.

3.7 Sharpness

Sharpness, one of the basic values of sensation, is related to timbre. A sound is perceived to be sharp if it contains mainly high frequency components. For example, a sine-tone or a band-limited noise at high frequencies, or a high-pass noise with a cut-off frequency in the frequency range

above about 3 kHz is said to be sharp. The detailed frequency structure of the audio signal, however, has no major influence on the sharpness. The fundamental research concerning sharpness has been conducted by Von Bismarck [1974].

Additional investigations regarding sharpness were conducted by [Aures, 1984]. The result of these investigations is a slightly modified weighting function compared with the weighting function defined by Bismarck. It contributes less to the attribution of sharpness at very low and at very high critical band rates, and more at critical band rates between 14 and 20 Bark. In addition, these investigations have shown that the sharpness of audio signals with a high variation of the audio signal sound pressure level and strong high frequency contents cannot be based only on the overall loudness, but on a weighting function, which depends on the overall loudness.

3.8 Cognitive processing

It is clear that perceived audio quality is strongly influenced by cognitive effects. This may be demonstrated by a simple experiment.

A Reference Signal with a clearly audible background noise is processed by some audio equipment that is not capable of transmitting this background noise. Since the noise is an undesired distortion, the Reference Signal would be rated worse than the processed signal in a listening test. On the other hand, the same processed signal would score worse if the most important part of the Reference Signal was the soft background noise.

Listing all possible cognitive effects is outside the scope of this Appendix, but some examples are:

Example 1: Separation of linear from non-linear distortions

Linear distortions are less objectionable than non-linear distortions. Separation of linear from non-linear distortions can be implemented fairly easily by using adaptive inverse filtering of the output signal. The method specified in this Recommendation uses a separation of linear from non-linear distortions.

Example 2: Auditory scene analysis

Auditory scene analysis [Bregman, 1990] is a cognitive process that allows listeners to separate different auditory events and group them into different objects. A pragmatic approach, as given in [Beerends and Stemerink, 1994], was useful for quantifying an auditory scene analysis effect. If a time-frequency component is not coded by a codec, the remaining signal still forms one coherent auditory scene, while the introduction of a new unrelated time frequency component leads to two different perceptions. Because of the split into two different perceptions, the distortion will be more objectionable than one would expect on the basis of the loudness of the newly introduced distortion component. This leads to an asymmetry between the perceived disturbance of a distortion that is caused by not coding a time-frequency component versus the disturbance caused by the introduction of a new time-frequency component.

Example 3: Informational masking

Informational masking can be modelled by defining an entropy-like spectral-temporal complexity measure. The effect is most probably dependent on the amount of training that subjects are exposed to before the subjective evaluation is carried out. A first attempt to model this effect is given in [Beerends *et al*, 1996] where a local complexity estimate over a time window of about 100 ms is calculated. If this local complexity is high, then distortions within this time window are more difficult to hear than when the local complexity is low. Training can reduce the masked threshold by several 10 dB [Leek and Watson, 1984].

Example 4: Spectral-temporal weighting

Some spectral-temporal regions in the audio signal carry more information, and may therefore be more important than others. Spectral-temporal weighting was found to be important in quality judgements on speech codecs. In speech, some spectral-temporal components, such as formants, clearly carry more information than others [Beerends and Stemerdink, 1994]. In music, however, all spectral-temporal components in the signal, even silences, may carry information.

4 Models incorporated

4.1 DIX

The perceptual measurement method DIX (Disturbance Index) [Thiede and Kabet, 1996] is based on an auditory filter bank that yields a high temporal resolution and thus allows (compared to FFT-based approaches) a more precise modelling of temporal effects like pre- and post-masking. The temporal fine structure of the envelopes at each auditory filter is preserved and is used to obtain additional information about the signals and the introduced distortions.

The centre frequencies of the individual filters are equally distributed over a perceptual pitch scale. The top of the filter shape is slightly rounded to ensure that the chosen number of filters covers the full frequency range without ripples in the overall frequency response. In order to model masked thresholds, the filter slopes decrease exponentially over the Bark scale. The steepness of the filter slopes depends on the level of the input signals. The audible frequency range was covered by 80 filters in the first version of DIX and was later reduced to 40 filters, i.e. the frequency resolution corresponds to approximately 0.6 Bark. The filter bank algorithm is rather fast as compared to other filter banks with individual filters, but is still much more time consuming than block-based transforms like FFT and wavelet-package-transforms.

DIX dynamically adapts the levels and spectra between the Signal Under Test and the Reference Signal in order to separate linear from non-linear distortions. It evaluates the structure of the temporal envelopes at the filter outputs in order to model the increased amount of masking caused by modulated and noise-like maskers as compared to pure tones.

By a comparison of the internal representations of the Signal Under Test and the Reference Signal, numerous output parameters are calculated, including the partial loudness of non-linear distortions, indicators for the amount of linear distortions and measures for temporal and binaural effects. However a good estimation of basic audio quality can be achieved by using only two of the output

parameters: The partial loudness of non-linear distortions together with one of the indicators for the amount of linear distortions are mapped to an estimate for the expected basic audio quality of the Signal Under Test.

4.2 NMR

The measurement scheme NMR (Noise-to-Masked-Ratio) [Brandenburg, 1987] evaluates the level-difference between the masked threshold and the noise signal. A DFT with a Hann window of about 20 ms is used to analyse the frequency content of the signal. The transform coefficients are combined to bands according to the Bark scale. The masked threshold is estimated for each band. The slope of the masked threshold is derived using a worst-case approach taking into account the fact that the slopes are steeper for weak signals but run into the absolute threshold at higher levels. The absolute threshold is adapted to the resolution of the input signal (usually 16 bits), but not to psycho-acoustic demands. Due to these facts NMR is robust to changes of the reproduction level. The pitch scale resolution is about 1 Bark. Since the required computational power is low it was possible to implement NMR as a real-time system at an early stage of its development.

The model has been in use since 1987 and has proven its basic reliability.

The most important output values of NMR are the masking flag rate, giving the percentage of frames with audible distortions, as well as the total and mean NMR which are different ways of averaging the distance between the error energy and the masked threshold.

4.3 OASE

The measurement scheme OASE (Objective Audio Signal Evaluation) [Sporer, 1997] uses a filter bank with 241 filters to analyse the input signals. The centre frequencies are equally spaced on the Bark scale with a distance of 0.1 Bark. The filters overlap each other. Each of the filters is adapted to the frequency response of a point on the basilar membrane. The level dependency of the slopes is included via a worst-case approach as done in NMR. The filters at low centre frequencies require calculation at the full sampling rate while the filters at higher centre frequencies can be calculated at a reduced sampling rate. After the filters, a model of the temporal effects of the human auditory system, as done in ASD, is calculated. Following this step, a reduction of the sampling rate in all filter bands is possible. This leads to a temporal resolution of the filter bank of 0.66 ms at a sampling rate of 48 kHz. The output of matching filters of reference and Signal Under Test are compared with a probability of detection function. This function uses the loudness of the input signals as input to calculate the JNLD. The total probability of detection is derived from the probability of detection of each band. This operation is done for both input channels and also for the so-called centre channel. The probability of detection in the centre channel for each band is the worst case of the probability of detection of the left and the right channel. For each frame of 0.66 ms the sum of the steps above threshold is calculated, too.

Several ways of temporal averaging of the probability of detection and the steps above threshold are used:

- the temporal average of the probability of detection;
- the frequency of frames with a probability of detection above 0.5;
- the maximum of a low pass filtered probability of detection;
- the maximum of a low pass filtered probability of detection with forgetting;
- the average number of steps above threshold for frames with a probability of detection above 0.5;
- the average number of steps above threshold;
- the maximum number of steps above threshold;
- the average of the number of steps above the threshold of the 10% worst frames.

4.4 Perceptual Audio Quality Measure (PAQM)

The basic principle of PAQM [Beerends and Stemerdink, 1992] is to subtract the internal representations (representations inside the head of the subject) of the reference and degraded signal and map the difference with a cognitive mapping to the subjectively perceived audio quality. The transformation from the physical, external domain to the psycho-physical, internal domain is performed by way of four operations:

- a time-frequency mapping which is done via a DFT with a Hann window of about 40 ms duration;
- frequency warping using the Bark scale;
- time-frequency spreading (non-linear convolution);
- intensity warping (compression).

The combination of smearing and compression allows modelling of the masking behaviour of the human auditory system at and above the masked threshold. The optimization of the compression is performed by using subjective results of the first MPEG audio codec evaluation [ISO/IEC JTC 1/SC 2/WG 11 MPEG/Audio test report, Document MPEG90/N0030, October 1990] [ISO/IEC JTC 1/SC 2/WG 11 MPEG/Audio test report, Document MPEG91/N0010, June 1991]. The difference in internal representation is expressed in terms of the noise disturbance. In the latest PAQM versions, as submitted to ITU-R, two cognitive effects were included in the mapping from the noise disturbance to the subjective quality, perceptual streaming [Beerends and Stemerdink, 1994] and informational masking [Beerends *et al*, 1996].

A simplified version of PAQM, the Perceptual Speech Quality Measure, PSQM [Beerends and Stemerdink, 1994] was developed using a cognitive model as presented in [Beerends and Stemerdink, 1994] but extended with a weighting of silent intervals. During the development of PSQM it turned out that in judging speech quality in a telephony context the noise that occurs during the silent intervals is of less importance than the noise that occurs during speech active intervals. In a benchmark by the ITU-T the PSQM proposal showed the highest correlation between objective and subjective quality (Telecommunication Standardization Study Group 12, COM 12-74 – Review of validation tests for objective speech quality measures). It was standardized as ITU-T Recommendation P.861 “Objective quality measurement of telephone band (300-3400 Hz) speech codecs”.

4.5 PERCEVAL

PERCEVAL (PERCeptual EVALuation) [Paillard *et al*, 1992] models the transfer characteristics of the middle and inner ear to form an internal representation of the signal. The input signal is decomposed into a time-frequency representation using a DFT. Typically, a Hann window of approximately 40 ms is applied to the input data, with a 50% overlap between successive windows. The energy spectrum is multiplied by a frequency dependent function which models the effect of the ear canal and the middle ear. The attenuated spectral energy values are mapped from the frequency scale to a pitch scale that is more linear with respect to both the physical properties of the inner ear and observed psycho-physical effects. The transformed energy components are then convolved with a spreading function to simulate the dispersion of energy along the basilar membrane. Finally, an intrinsic frequency-dependent energy is added to each pitch component to account for the absolute threshold of hearing. Conversion of the energy to decibels results in a basilar membrane representation of the signal.

In simulations of auditory masking experiments, a basilar membrane representation is formed for each stimulus, and the difference between the representations is the information available for performing the task. One representation is of the masker alone, and the other is of the masker and test signal combined. Their difference represents the component of the signal that is not masked. PERCEVAL calculates the probability of detecting this difference. The probability of non-detection of the difference for each detector along the simulated basilar membrane is estimated using a sigmoidal probability function. With the assumption that the detectors are statistically independent, the global detection probability for the whole set of detectors is calculated as the complement of the product of the individual non-detection probabilities. Several masking experiments were successfully simulated using this approach, and the model was used to evaluate the feasibility of modelling individual listeners [Treurniet, 1996].

As a tool for estimating audio quality, PERCEVAL calculates the difference between the representations of the Reference Signal and the Signal Under Test. By applying reasonable assumptions about higher level perceptual and cognitive processes, a number of perceptually relevant variables are computed and mapped to a measure of the objective quality of the Signal Under Test. The mapping was optimized by minimizing the difference between the objective quality distribution and the corresponding distribution of mean subjective quality ratings for the available data set.

4.6 POM

The purpose of the Perceptual Objective Measurement (POM) [Colomes *et al*, 1995] is to quantify a certain amount of degradation that may occur between a Reference Signal and its “degraded” version. This is accomplished by comparing the internal basilar representation of both signals, whatever the degradation is produced by. The basilar representation models the different processes undergone by an audio signal when travelling through the human ear. Therefore, the first stage of POM is the calculation of the internal representation of an audio signal. The excitation pattern (given in dB), spread over the basilar membrane, has been chosen to model the firing rate in the neurones along the basilar membrane.