RECOMMENDATION ITU-R BS.1387

METHOD FOR OBJECTIVE MEASURMENTS OF PERCEIVED AUDIO QUALITY

(Question ITU-R 210/10)

(1998)

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

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 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 (SNR) 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), PERCEVAL, Perceptual Objective Measure (POM) and The Toolbox Approach. The methods are described in Appendix 4 to Annex 1.

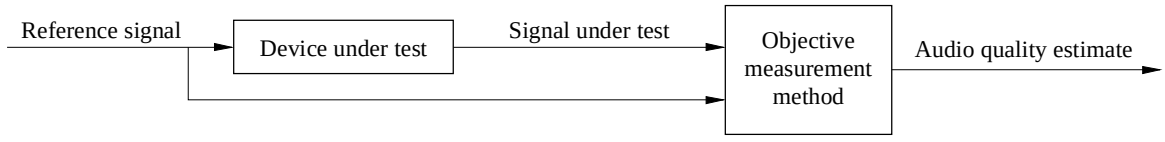
The measurement method in this Recommendation is the result of a process where the performance of each of the just mentioned six methods was studied, and the most promising tools extracted and integrated into one single method. The recommended method has been carefully validated at a number of test sites. 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 Figure 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.

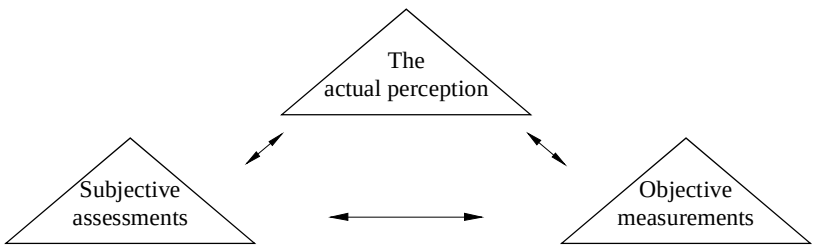
Table 1 gives some guidance on which version to apply for each of the applications.

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.

FIGURE 2

Validation concepts



1387-02

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

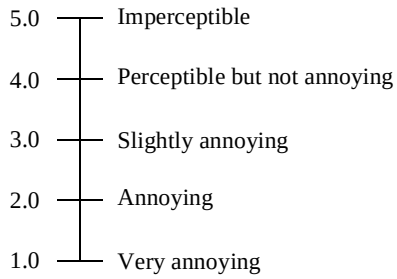
The objective measurement method in this Recommendation has been focused on applications which normally are assessed in the subjective domain by applying Recommendation ITU-R BS.1116. The basic principle of that particular test method can be briefly described as follows: the listener can select between 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



1387-03

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

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

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 is 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 two times 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 two times 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 – 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 seconds 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

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.

---

1 Rate of output values (per second).

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

The following table 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

Remarks:

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 seconds. 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 – 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 test[2]. 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.

---

2 See Recommendation ITU-R BS.1116.

TABLE 4

Description of the Model Output Variables

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

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.

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, March 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.

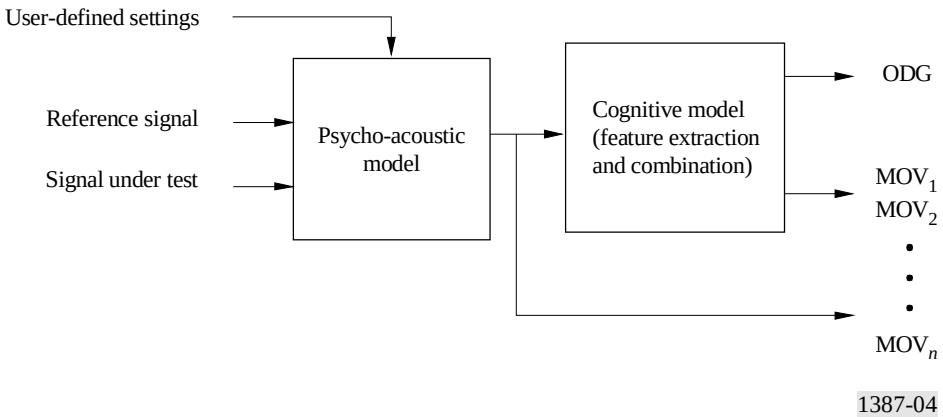
The recommended model for objective measurement produces a number of Model Output Variables (MOVs) based on comparisons between the Reference Signal and the Signal Under Test. These MOVs are mapped to an ODG using an optimization technique that minimizes the squared difference between the ODG distribution 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 Figure 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 Figure 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 1 019.5 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 mainly aim 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 signal-to-noise ratio 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.

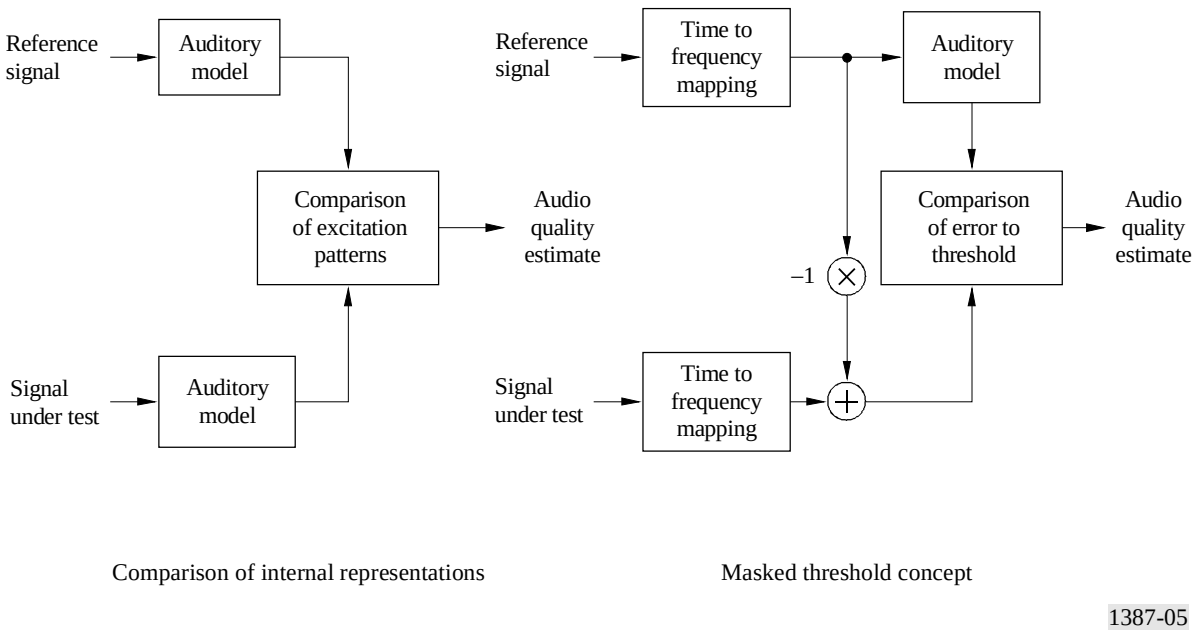
The second approach is closer to the physiological processes in the human auditory system: a so-called internal representation of both the Reference Signal and the Signal Under Test is calculated. This internal representation is an estimate of the information that is available to the human brain for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



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 6. The table also contains a definition of the Bark-scale: 1 Bark corresponds to 100 Hz, 24 Bark corresponds to 15 000 Hz.

TABLE 6

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

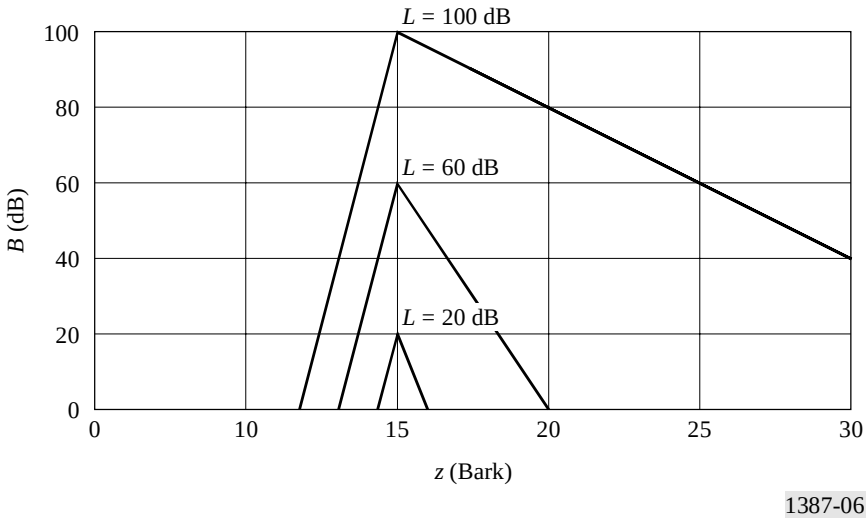
Several approximations to the Bark scale were found in the past. A detailed discussion of different scales can be found in [Cohen and Fielder, 1992]. In the context of objective measurement of perceived audio quality, the best results were achieved using the Bark scale.

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].

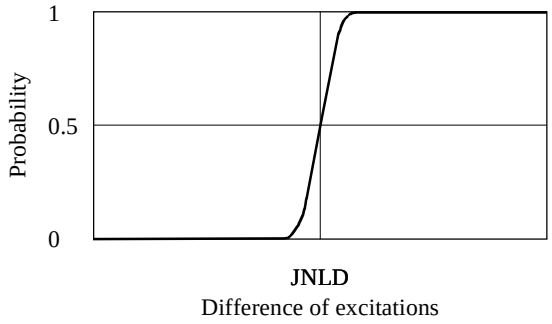
3.4 Detection

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 memories play 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 $\left(15.5+\frac{z}{\text{Bark}}\right)\text{dB}$ 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 G.v. Bismarck [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 Bismark. 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:

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.

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 Stemerdink, 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.

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].

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, March 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 Kabot, 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 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 (ITU-T 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 (300-3 400 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 per cent 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.

The process of calculating the excitation pattern is called the artificial ear. Then, once we get the two internal representations of the signals to be compared to each other, POM has to state whether the difference between their internal representation is audible or not, and if so in which way. This is called the detection process.

POM uses a DFT with a Hann window of approximately 40 ms duration (with an overlap of 50% between two Hann windows). The number of analysis basilar channels is 620. The remaining parts of the auditory model are almost identical to the ones used in both PAQM and PERCEVAL.

The spreading function is quite accurately described by a more precise approximation that takes into account both the level dependency according to [Terhardt, 1979] and the rounded shape according to [Schroeder et al, 1979].

This model outputs the probability of detecting a distortion between the two compared signals, as well as a so-called basilar distance that represents the perceptual gap between the two compared excitations.

4.7 The toolbox approach

Toolbox uses a three step approach to measure the perceived distance in audio quality of an audio test signal in relation to an audio Reference Signal, thus giving an indication of the overall subjective audio quality level of the test signal. The method is based on well-known perceptual models which are used to describe the perceptual representation of the differences between the two audio signals. Further, it includes a weighting procedure for the perceived audio quality of a stereo test signal, taking into account the results of both, left and right channels. A rigid correlation on a sample-by-sample basis of the reference and the audio Signal Under Test is not required.

The main functionality of toolbox, step 1, is based on the calculation of the specific loudness, calculated according to [Zwicker and Feldtkeller, 1967], using an FFT of 2 048 points, windowed by a Hann window, which corresponds to about 40 ms duration. The whole window is shifted by increments of 10 ms. In addition, temporal masking effects, such as post- and pre-masking, according to Zwicker, are applied. From these basic values of sensation, other perceptual parameters, such as integrated loudness, partially masked loudness, sharpness, according to [von Bismarck, 1974] and [Aures, 1984], and the amount of pre-echoes are calculated as a result of a pre-processing stage for the next steps.

The second step of toolbox includes weighting procedures which depend mainly on the amount of the perceived difference in loudness and the variation of loudness in time.

The third step of toolbox includes the generation of a set of intermediate toolbox output values which are based on a statistical analysis of the values obtained in steps 1 and 2. The output of this statistical analysis includes, the mean, maximum and r.m.s. values, as well as the standard deviation of the mean values. A weighted sum of these intermediate toolbox output values is used for the final fitting of the perceptual distance between the Signal Under Test and the Reference Signal. If necessary, this single output value can be matched to a Subjective Difference Grade, usually obtained in subjective listening tests, by sequentially fitting the output data for each increment of time using either a linear or higher order polynomial function.

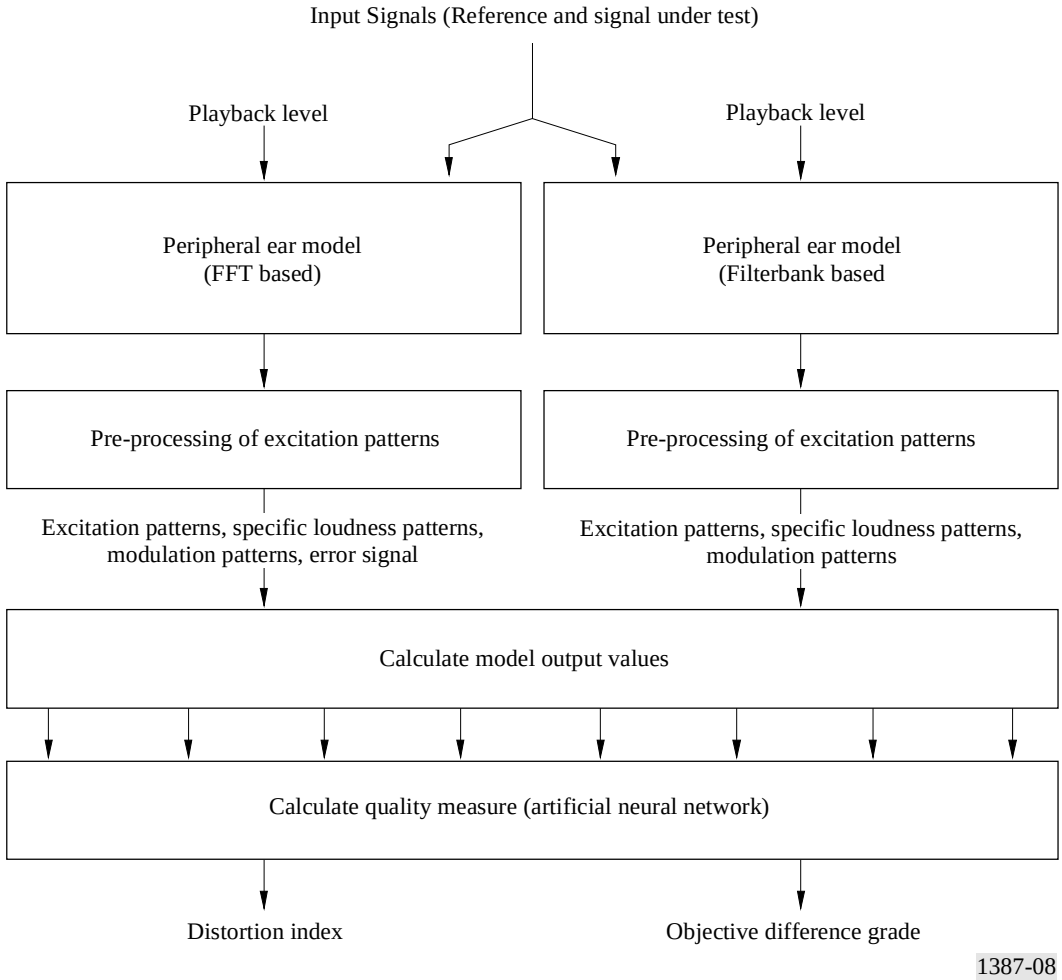
ANNEX 2

Description of the model[3]

1 Outline

FIGURE 8

Generic block diagram of the measurement scheme



The proposed *Method for Objective Measurement of Perceived Audio Quality* consists of a *peripheral ear model*, several intermediate steps (here referred as "*pre-processing of excitation patterns*"), the calculation of (mostly) psycho-acoustically based *Model Output Variables* ("MOVs") and a mapping from a set of Model Output Variables to a single value representing the *basic audio quality* of the Signal Under Test. It includes two peripheral ear models, one

[3] The proponents of the technology described in this Recommendation have submitted patent statements conforming to Annex 1 of Resolution ITU-R 1-2. The technology described within this Recommendation is protected by international patents, and like all ITU Recommendations subject to copyright. Prior consent of the owners in the form of a licence is mandatory to exploit this technology. To obtain further information regarding licensing this technology please refer to the patent database of the ITU-R, or to the BR secretariat.

based on an FFT and one based on a filter bank. Except for the calculation of the error signal (which is only used with the FFT-based part of the ear model) the general structure is the same for both peripheral ear models.

The inputs for the MOV calculation are:

– The excitation patterns for both test and Reference Signal.

– The spectrally adapted excitation patterns for both test and Reference Signal.

– The specific loudness patterns for both test and Reference Signal.

– The modulation patterns for both test and Reference Signal.

– The error signal calculated as the spectral difference between test and Reference Signal (only for the FFT-based ear model).

If not indicated differently, in the case of stereo signals all computations are performed independently and in the same manner for the left and right channel.

The description defines two setups, one called the "*Basic Version*" and one called the "*Advanced Version*".

In all given equations, the index "*Ref.*" stands for all patterns calculated from the Reference Signal, the index "*Test*" stands for all patterns calculated from the Signal Under Test. The index "*k*" stands for the discrete frequency variable (i.e. the frequency band) and "*n*" stands for the discrete time variable (i.e. either the frame counter or the sample counter). If the values for $\boldsymbol{k}$ or $\boldsymbol{n}$ are not explicitly defined, the computations are to be carried out for all possible values of $\boldsymbol{k}$ and $\boldsymbol{n}$. All other abbreviations are explained at the place they occur.

In the names of the Model Output Variables, the index "*A*" stands for all variables calculated using the filter bank-based part of the ear model and the index "*B*" stands for all variables calculated using the FFT-based part of the ear model.

1.1 Basic Version

The *Basic Version* includes only MOVs that are calculated from the FFT-based ear model. The filter bank-based part of the model is not used. The *Basic Version* uses a total of 11 MOVs for the prediction of the perceived *basic audio quality*.

1.2 Advanced Version

The *Advanced Version* includes MOVs that are calculated from the filter bank-based ear model as well as MOVs that are calculated from the FFT-based ear model. The spectrally adapted excitation patterns and the modulation patterns are computed from the filter bank-based part of the model only. The *Advanced Version* uses a total of 5 MOVs for the prediction of the perceived *basic audio quality*.

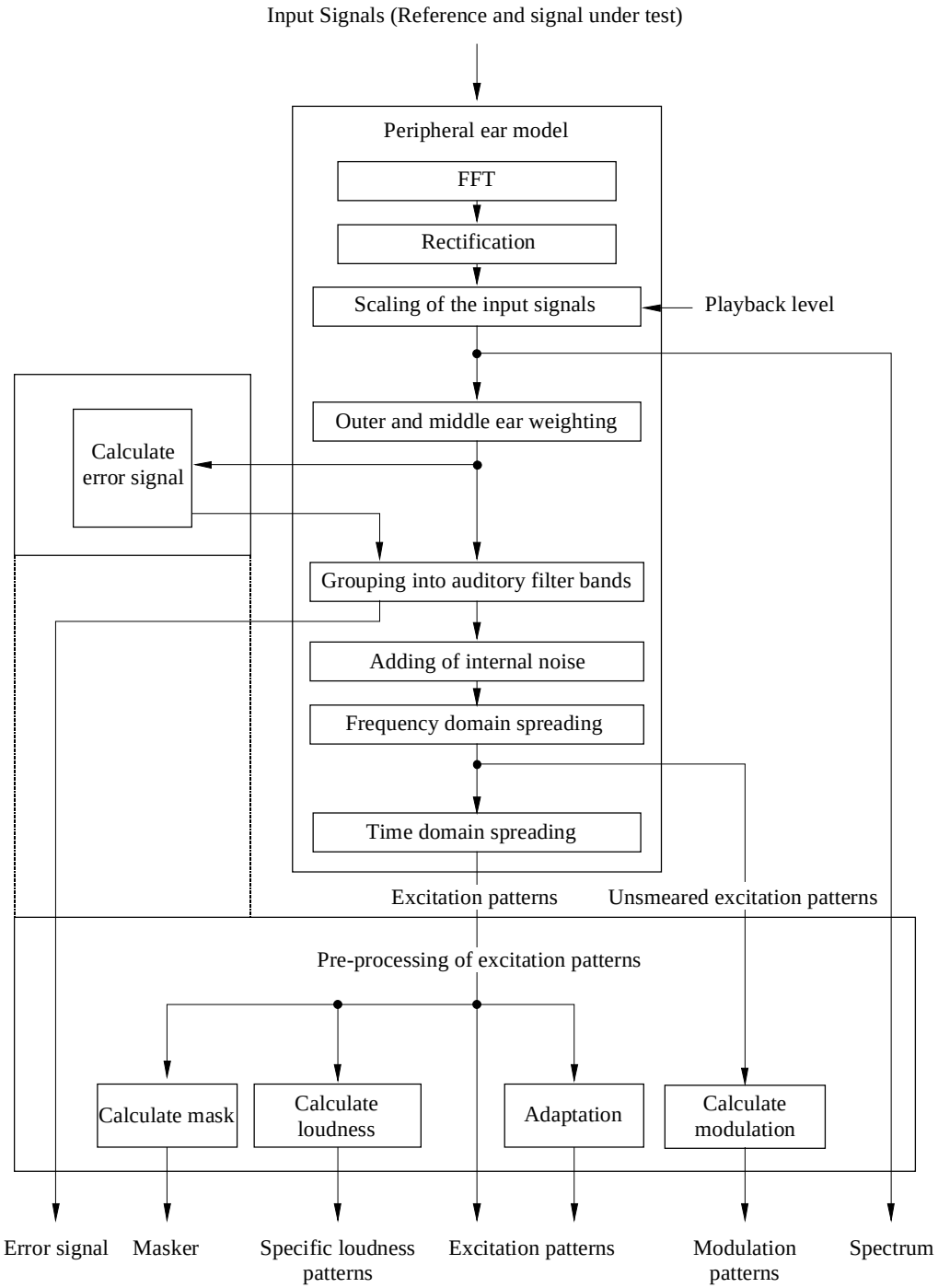
2 Peripheral ear model

2.1 FFT-based ear model

2.1.1 Overview

FIGURE 9

Peripheral ear model and pre-processing of excitation patterns for the FFT-based part of the model



The input of the FFT-based ear model, 48 kHz sampled time aligned reference and test signals, are cut into frames of about 0.042 seconds with an overlap of 50%. Each frame is transformed to the frequency domain using a Hann window and a short term FFT, and scaled to the playback level. A weighting function is applied to the spectral coefficients, which models the outer and middle ear frequency response. The transformation to the pitch representation is done by grouping the weighted spectral coefficients into critical bands. A frequency dependent offset is added to simulate the internal noise in the auditory system. A level dependent spreading function is used to model the spectral auditory filters in the frequency domain. It follows a time domain spreading that accounts for forward masking effects.

The now obtained *excitation patterns* are used to compute *specific loudness patterns* and the *Masking patterns*. The patterns before the final time domain spreading ("*unsmeared excitation patterns*") are used to calculate *modulation patterns*.

To model the error signal, the reference and test signal patterns of the output of the outer and middle ear filter are combined and mapped to the pitch scale by grouping to critical bands.

These outputs are used together with the excitation patterns to calculate the values of the Model Output Variables.

2.1.2 Time processing

The input of the FFT-based ear model, test and Reference Signal are cut into frames of 2 048 samples with an overlap of 1 024 samples:

$$t_n\left[k_t, n\right] = t\left[1024 \cdot n + k_t\right] \quad \text{n} = 0,1,2 \ldots \; \boldsymbol{k_t} = 0..2047 \tag{1}$$

where n is the number of the time-frame and $\boldsymbol{k_t}$ is a time counter inside a frame.

2.1.3 FFT

The mapping from the time domain to the frequency domain is done using a Hann window:

$$h_w[k] = \frac{1}{2}\sqrt{\frac{8}{3}}\left[1 - \cos\left(2\pi \frac{k}{N-1}\right)\right] \quad \Big| \quad N = \; 2048 \tag{2}$$

$$t_w\left[k_t, n\right] = h_w\left[k_t\right] \cdot t_n\left[k_t, n\right] \tag{3}$$

followed by a short term Fourier-transform:

$$F_f\left[k_f, n\right] = \frac{1}{2048}\sum_{k_t=0}^{2047} t_w\left[k_t, n\right] e^{-j\frac{2\pi}{2048}k_f k_t} \tag{4}$$

The scaling factor for the FFT is calculated from the assumed sound pressure level $\boldsymbol{Lp}$ of a full scale sine wave by:

$$fac = \frac{10^{\frac{\boldsymbol{Lp}}{20}}}{Norm} \tag{5}$$

$$F\left[k_f, n\right] = fac \cdot F_f\left[k_f, n\right] \tag{6}$$

Where the normalization factor *Norm* is calculated by taking a sine wave of 1 019.5 Hz and 0 dB full scale as the input signal and calculating the maximum absolute value of the spectral coefficients over 10 frames.

If the sound pressure level is unknown it is recommended to set $\boldsymbol{Lp}$ to 92 dB_{SPL}.

2.1.4 Outer and middle ear

The frequency response of the outer and middle ear is modelled by a frequency dependent weighting function:

$$W[k]/dB = -0.6 \cdot 3.64 \cdot \left(\frac{f[k]}{kHz}\right)^{-0.8} + 6.5 \cdot e^{-0.6\cdot\left(\frac{f[k]}{kHz}-3.3\right)^2} - 10^{-3} \cdot \left(\frac{f[k]}{kHz}\right)^{3.6} \quad (7)$$

where

$$f[k]/Hz = k \cdot 23.4375 \quad (8)$$

is the frequency representation at line k that is applied to the FFT outputs (equation 9).

$$F_e[k_f, n] = \left|F[k_f, n]\right| \cdot 10^{\frac{W[k_f]}{20}} \quad (9)$$

$F_e[k_f]$ are referenced as "*Outer ear weighted FFT outputs*".

2.1.5 Grouping into critical bands

The auditory pitch scale is calculated from an approximation given by [Schroeder et al, 1979].

$$z/Bark = 7 \cdot \operatorname{arsinh}\left(\frac{f/Hz}{650}\right) \quad (10)$$

The pitch units are named *Bark* (although this scale does not exactly represent the Bark-scale as defined by [Zwicker and Feldtkeller, 1967]).

The frequency borders of the filters range from 80 Hz to 18 000 Hz. The widths and spacing of the filter bands correspond to a resolution of $res=0.25$ Bark for the Basic Version and $res=0.5$ Bark for the advanced version.

This leads to the number of frequency bands $Z=109$ for the Basic Version and $Z=55$ for the Advanced Version.

TABLE 7

Frequency bands of the FFT-based ear model used in the Basic Version

Group	Lower frequency/Hz	Centre frequency/Hz	Upper frequency/Hz	Frequency width/Hz
k	$f_l[k]$	$f_c[k]$	$f_u[k]$	$f_w[k]$
0	80	91.708	103.445	23.445
1	103.445	115.216	127.023	23.577
2	127.023	138.87	150.762	23.739
3	150.762	162.702	174.694	23.932
4	174.694	186.742	198.849	24.155
5	198.849	211.019	223.257	24.408
6	223.257	235.566	247.95	24.693
7	247.95	260.413	272.959	25.009
8	272.959	285.593	298.317	25.358

TABLE 7 (*Continued*)

Group	Lower frequency/Hz	Centre frequency/Hz	Upper frequency/Hz	Frequency width/Hz
k	$f_l[k]$	$f_c[k]$	$f_u[k]$	$f_w[k]$
9	298.317	311.136	324.055	25.738
10	324.055	337.077	350.207	26.151
11	350.207	363.448	376.805	26.598
12	376.805	390.282	403.884	27.079
13	403.884	417.614	431.478	27.594
14	431.478	445.479	459.622	28.145
15	459.622	473.912	488.353	28.731
16	488.353	502.95	517.707	29.354
17	517.707	532.629	547.721	30.014
18	547.721	562.988	578.434	30.713
19	578.434	594.065	609.885	31.451
20	609.885	625.899	642.114	32.229
21	642.114	658.533	675.161	33.048
22	675.161	692.006	709.071	33.909
23	709.071	726.362	743.884	34.814
24	743.884	761.644	779.647	35.763
25	779.647	797.898	816.404	36.757
26	816.404	835.17	854.203	37.799
27	854.203	873.508	893.091	38.888
28	893.091	912.959	933.119	40.028
29	933.119	953.576	974.336	41.218
30	974.336	995.408	1016.797	42.461
31	1016.797	1038.511	1060.555	43.758
32	1060.555	1082.938	1105.666	45.111
33	1105.666	1128.746	1152.187	46.521
34	1152.187	1175.995	1200.178	47.991
35	1200.178	1224.744	1249.7	49.522
36	1249.7	1275.055	1300.816	51.116
37	1300.816	1326.992	1353.592	52.776
38	1353.592	1380.623	1408.094	54.502
39	1408.094	1436.014	1464.392	56.298
40	1464.392	1493.237	1522.559	58.167
41	1522.559	1552.366	1582.668	60.109
42	1582.668	1613.474	1644.795	62.128
43	1644.795	1676.641	1709.021	64.226
44	1709.021	1741.946	1775.427	66.406
45	1775.427	1809.474	1844.098	68.671
46	1844.098	1879.31	1915.121	71.023
47	1915.121	1951.543	1988.587	73.466
48	1988.587	2026.266	2064.59	76.003
49	2064.59	2103.573	2143.227	78.637

TABLE 7 (*Continued*)

Group	Lower frequency/Hz	Centre frequency/Hz	Upper frequency/Hz	Frequency width/Hz
k	$f_l[k]$	$f_c[k]$	$f_u[k]$	$f_w[k]$
50	2143.227	2183.564	2224.597	81.371
51	2224.597	2266.34	2308.806	84.208
52	2308.806	2352.008	2395.959	87.154
53	2395.959	2440.675	2486.169	90.21
54	2486.169	2532.456	2579.551	93.382
55	2579.551	2627.468	2676.223	96.672
56	2676.223	2725.832	2776.309	100.086
57	2776.309	2827.672	2879.937	103.627
58	2879.937	2933.12	2987.238	107.302
59	2987.238	3042.309	3098.35	111.112
60	3098.35	3155.379	3213.415	115.065
61	3213.415	3272.475	3332.579	119.164
62	3332.579	3393.745	3455.993	123.415
63	3455.993	3519.344	3583.817	127.823
64	3583.817	3649.432	3716.212	132.395
65	3716.212	3784.176	3853.348	137.136
66	3853.348	3923.748	3995.399	142.051
67	3995.399	4068.324	4142.547	147.148
68	4142.547	4218.09	4294.979	152.432
69	4294.979	4373.237	4452.89	157.911
70	4452.89	4533.963	4616.482	163.592
71	4616.482	4700.473	4785.962	169.48
72	4785.962	4872.978	4961.548	175.585
73	4961.548	5051.7	5143.463	181.915
74	5143.463	5236.866	5331.939	188.476
75	5331.939	5428.712	5527.217	195.278
76	5527.217	5627.484	5729.545	202.329
77	5729.545	5833.434	5939.183	209.637
78	5939.183	6046.825	6156.396	217.214
79	6156.396	6267.931	6381.463	225.067
80	6381.463	6497.031	6614.671	233.208
81	6614.671	6734.42	6856.316	241.646
82	6856.316	6980.399	7106.708	250.392
83	7106.708	7235.284	7366.166	259.458
84	7366.166	7499.397	7635.02	268.854
85	7635.02	7773.077	7913.614	278.594
86	7913.614	8056.673	8202.302	288.688
87	8202.302	8350.547	8501.454	299.152
88	8501.454	8655.072	8811.45	309.996
89	8811.45	8970.639	9132.688	321.237
90	9132.688	9297.648	9465.574	332.887

TABLE 7 (*end*)

Group	Lower frequency/Hz	Centre frequency/Hz	Upper frequency/Hz	Frequency width/Hz
k	$f_l[k]$	$f_c[k]$	$f_u[k]$	$f_w[k]$
91	9465.574	9636.52	9810.536	344.962
92	9810.536	9987.683	10168.013	357.477
93	10168.013	10351.586	10538.46	370.447
94	10538.46	10728.695	10922.351	383.891
95	10922.351	11119.49	11320.175	397.824
96	11320.175	11524.47	11732.438	412.264
97	11732.438	11944.149	12159.67	427.231
98	12159.67	12379.066	12602.412	442.742
99	12602.412	12829.775	13061.229	458.817
100	13061.229	13296.85	13536.71	475.48
101	13536.71	13780.887	14029.458	492.748
102	14029.458	14282.503	14540.103	510.645
103	14540.103	14802.338	15069.295	529.192
104	15069.295	15341.057	15617.71	548.415
105	15617.71	15899.345	16186.049	568.339
106	16186.049	16477.914	16775.035	588.986
107	16775.035	17077.504	17385.42	610.385
108	17385.42	17690.045	18000	614.58

TABLE 8

Frequency bands of the FFT-based ear model used in the Advanced Version

Group	Lower frequency/Hz	Centre frequency/Hz	Upper frequency/Hz	Frequency width/Hz
k	$f_l[k]$	$f_c[k]$	$f_u[k]$	$f_w[k]$
0	80	103.445	127.023	47.023
1	127.023	150.762	174.694	47.671
2	174.694	198.849	223.257	48.563
3	223.257	247.95	272.959	49.702
4	272.959	298.317	324.055	51.096
5	324.055	350.207	376.805	52.75
6	376.805	403.884	431.478	54.673
7	431.478	459.622	488.353	56.875
8	488.353	517.707	547.721	59.368
9	547.721	578.434	609.885	62.164
10	609.885	642.114	675.161	65.277
11	675.161	709.071	743.884	68.723
12	743.884	779.647	816.404	72.52
13	816.404	854.203	893.091	76.687
14	893.091	933.119	974.336	81.245

TABLE 8 *(end)*

Group	Lower frequency/Hz	Centre frequency/Hz	Upper frequency/Hz	Frequency width/Hz
k	$f_l[k]$	$f_c[k]$	$f_u[k]$	$f_w[k]$
15	974.336	1016.797	1060.555	86.219
16	1060.555	1105.666	1152.187	91.632
17	1152.187	1200.178	1249.7	97.513
18	1249.7	1300.816	1353.592	103.892
19	1353.592	1408.094	1464.392	110.801
20	1464.392	1522.559	1582.668	118.275
21	1582.668	1644.795	1709.021	126.354
22	1709.021	1775.427	1844.098	135.077
23	1844.098	1915.121	1988.587	144.489
24	1988.587	2064.59	2143.227	154.64
25	2143.227	2224.597	2308.806	165.579
26	2308.806	2395.959	2486.169	177.364
27	2486.169	2579.551	2676.223	190.054
28	2676.223	2776.309	2879.937	203.713
29	2879.937	2987.238	3098.35	218.414
30	3098.35	3213.415	3332.579	234.229
31	3332.579	3455.993	3583.817	251.238
32	3583.817	3716.212	3853.348	269.531
33	3853.348	3995.399	4142.547	289.199
34	4142.547	4294.979	4452.89	310.343
35	4452.89	4616.482	4785.962	333.072
36	4785.962	4961.548	5143.463	357.5
37	5143.463	5331.939	5527.217	383.754
38	5527.217	5729.545	5939.183	411.966
39	5939.183	6156.396	6381.463	442.281
40	6381.463	6614.671	6856.316	474.853
41	6856.316	7106.708	7366.166	509.85
42	7366.166	7635.02	7913.614	547.448
43	7913.614	8202.302	8501.454	587.84
44	8501.454	8811.45	9132.688	631.233
45	9132.688	9465.574	9810.536	677.849
46	9810.536	10168.013	10538.46	727.924
47	10538.46	10922.351	11320.175	781.715
48	11320.175	11732.438	12159.67	839.495
49	12159.67	12602.412	13061.229	901.56
50	13061.229	13536.71	14029.458	968.229
51	14029.458	14540.103	15069.295	1039.837
52	15069.295	15617.71	16186.049	1116.754
53	16186.049	16775.035	17385.42	1199.371
54	17385.42	17690.045	18000	614.58

The frequency to pitch mapping is done by the algorithm described in the following subsection, where $Fsp[k_f]$ is the energy representation of the "*Outer ear weighted FFT outputs*":

$$Fsp\left[k_f, n\right] = \left|F_e\left[k_f, n\right]\right|^2 \tag{11}$$

or the energy representation of the error signal

$$Fsp\left[k_f, n\right] = \left|F_{noise}\left[k_f, n\right]\right|^2 \tag{12}$$

respectively. See § 3.4 for calculation of the error signal.

The outputs of this stage of processing are the energies of the frequency groups, $P_e[k, n]$.

2.1.5.1 Pseudocode

/* inputs */

Fsp[ ] input energies

/* outputs */

Pe[ ] : pitch mapped energies

/* intermediate values */

i : index to frequency groups

k : \ index to fft line

Z number of frequency groups:

109 for the Basic Version
55 for the Advanced Version

fl[] : lower frequency of frequency group

fu[] : upper frequency of frequency group

Fres : constant for frequency resolution

```
Fres = 48000/2048;
for(i=0; i<Z; i++ )
{
 Pe[i]=0;
 for(k=0;k<1024;k++)
 {
  /* line inside frequency group */
  if( (( k-0.5)*Fres >= fl[i])  && ((k+0.5)*Fres <= fu[i]))
  {
    Pe[i] += Fsp[k];
  }
  /* frequency group inside*/
  else if( (( k-0.5)*Fres < fl[i])  && ((k+0.5)*Fres > fu[i]))
```

```
    {
      Pe[i] += Fsp[k]*(fl[i]-fu[i])/Fres;
    }
    /* left border */
    else if( ((k-0.5)*Fres < fl[i]) && ((k+0.5)*Fres > fl[i]))
    {
     Pe[i] += Fsp[k]*( (k+0.5)*Fres - fl[i])/Fres;
    }
    /* right border
    else if( ((k-0.5)*Fres < fu[i]) && ((k+0.5)*Fres > fu[i]);
    {
     Pe[i] += Fsp[k]*(fu[i]- (k-0.5)*Fres)/Fres;
    }
    /* line outside frequency group */
    else
    {
      Pe[i] += 0;
    }
  }

   /* limit result */
    Pe[i]=max(Pe[i],0.000000000001);
}
```

2.1.6 Adding internal noise

A frequency dependent offset P_{Thres} is added to the energies in each frequency group:

$$P_{Thres}[k] = 10^{0.4 \cdot 0.364 \cdot \left(\frac{f_c[k]}{kHz}\right)^{-0.8}} \tag{13}$$

$$P_p[k,n] = P_e[k,n] + P_{Thres}[k] \tag{14}$$

The Output of this stage of processing, $P_p[k,n]$ is referenced as "*Pitch patterns*".

2.1.7 Spreading

The *Pitch patterns* $P_p[k,n]$ are smeared out over frequency using a level dependent spreading function. The spreading function is a two sided exponential. The lower slope is always 27 dB/Bark and the upper slope is frequency and energy dependent.

The slopes are calculated according to:

$$\frac{S_u[k, L[k,n]]}{dB\,/\,Bark} = -24 - \frac{230Hz}{f_c[k]} + 0.2 \cdot L[k,n] / dB \tag{15}$$

$$S_l[k, L[k,n]] = 27\frac{dB}{Bark} \tag{16}$$

with

$$L[k,n] = 10 \cdot \log_{10}\left(P_p[k,n]\right)$$

The spreading is carried out independently for each frequency group k:

$$E_2[k,n] = \frac{1}{Norm_{SP}[k]}\left(\sum_{j=0}^{Z} E_{line}[j,k,n]^{0.4}\right)^{\frac{1}{0.4}} \tag{17}$$

where E_{line} is given by

$$E_{line}[j,k,n] = \begin{cases} \dfrac{10^{\frac{L[j,n]}{10}} \cdot 10^{\frac{-res\cdot(j-k)\cdot S_l[j,L[j,n]]}{10}}}{\sum_{\mu=0}^{j-1} 10^{\frac{-res\cdot(j-\mu)\cdot S_l[j,L[j,n]]}{10}} + \sum_{\mu=j}^{Z} 10^{\frac{res\cdot(\mu-j)\cdot S_u[j,L[j,n]]}{10}}} & \text{if } k < j \\ \dfrac{10^{\frac{L[j,n]}{10}} \cdot 10^{\frac{res\cdot(k-j)\cdot S_u[j,L[j,n]]}{10}}}{\sum_{\mu=0}^{j-1} 10^{\frac{-res\cdot(j-\mu)\cdot S_l[j,L[j,n]]}{10}} + \sum_{\mu=j}^{Z} 10^{\frac{res\cdot(\mu-j)\cdot S_u[j,L[j,n]]}{10}}} & \text{if } k \geq j \end{cases} \tag{18}$$

$Norm_{SP}[k]$ is calculated according to:

$$Norm_{SP}[k] = \left(\sum_{j=0}^{Z} \tilde{E}_{line}[j,k]^{0.4}\right)^{\frac{1}{0.4}} \tag{19}$$

with

$$\tilde{E}_{line}[j,k] = \begin{cases} \dfrac{10^{\frac{-res\cdot(j-k)\cdot S_l[j,0]}{10}}}{\sum_{\mu=0}^{j-1} 10^{\frac{-res\cdot(j-\mu)\cdot S_l[j,0]}{10}} + \sum_{\mu=j}^{Z} 10^{\frac{res\cdot(\mu-j)\cdot S_u[j,0]}{10}}} & \text{if } k < j \\ \dfrac{10^{\frac{res\cdot(k-j)\cdot S_u[j,0]}{10}}}{\sum_{\mu=0}^{j-1} 10^{\frac{-res\cdot(j-\mu)\cdot S_l[j,0]}{10}} + \sum_{\mu=j}^{Z} 10^{\frac{res\cdot(\mu-j)\cdot S_u[j,0]}{10}}} & \text{if } k \geq j \end{cases} \tag{20}$$

and res is the resolution of the pitch scale in Bark (0.25 for the Basic Version and 0.5 for the Advanced Version).

The patterns at this stage of processing, $E_2[k,n]$, are used later on for the computation of modulation patterns and are referred to as "*unsmeared excitation patterns*".

2.1.8 Time domain spreading

In order to model forward masking, the energies in each frequency group are smeared out over time by first order low pass filters. The time constants depend on the centre frequency of each group (as given in equation 10 and 7) and are calculated according to:

$$\tau = \tau_{\min} + \frac{100Hz}{f_c[k]} \cdot (\tau_{100} - \tau_{\min}) \quad \left| \begin{array}{lll} \tau_{100} & = & 0.030\ s \\ \tau_{\min} & = & 0.008\ s \end{array} \right. \tag{21}$$

The first order low pass filters are computed according to:

$$E_f[k,n] = a \cdot E_f[k,n-1] + (1-a) \cdot E_2[k,n] \tag{22}$$

$$E[k,n] = \max\left(E_f(k,n), E_2(k,n)\right) \tag{23}$$

where a is calculated from the above time constants by:

$$a = e^{-\frac{4}{187.5} \cdot \frac{1}{\tau}} \tag{24}$$

n is the actual frame number, k is the group index and $E_f[k,0] = 0$.

The patterns at this stage of processing, $E[k,n]$, are referred to as "*excitation patterns*".

2.1.9 Masking threshold

Masking describes the effect by which a fainter, but distinctly audible signal becomes inaudible when a correspondingly louder signal occurs. This threshold is calculated by weighting the excitation patterns with the weighting function m[k].

$$m[k] = \begin{cases} 3.0 & \text{k} \cdot \text{res} \le 12 \\ 0.25 \cdot k \cdot \text{res} & \text{k} \cdot \text{res} > 12 \end{cases} \tag{25}$$

$$M[k,n] = \frac{E[k,n]}{10^{\frac{m[k]}{10}}} \cdot \tag{26}$$

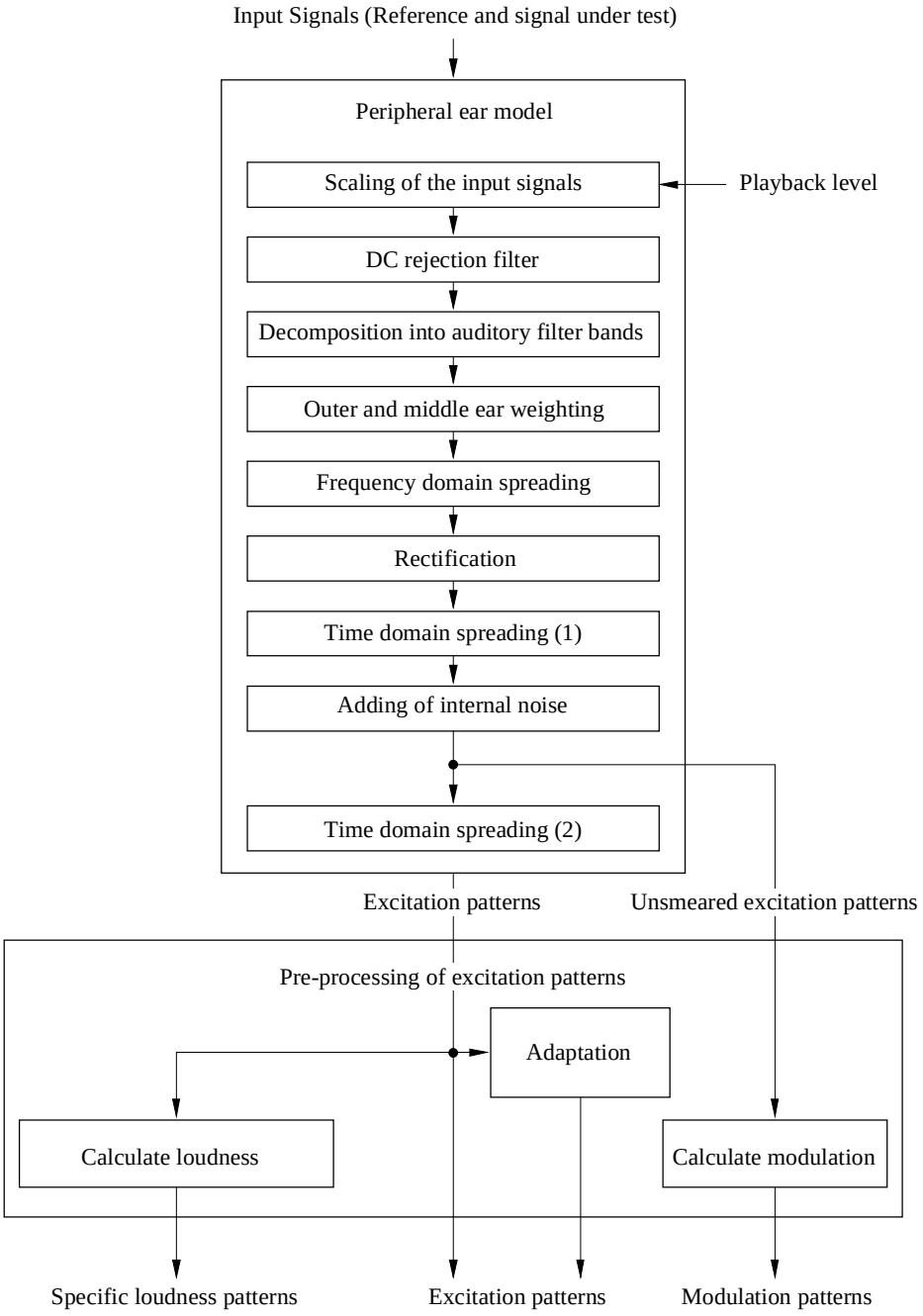
The patterns at this stage of processing, $M[k,n]$, are referred to as "*Mask patterns*".

2.2 Filter bank-based ear model

2.2.1 Overview

FIGURE 10

Peripheral ear model and pre-processing of excitation patterns for the filter bank-based part of the model



At the input of the filter bank-based ear model, Signal under Test and Reference Signal are adjusted to the assumed playback level and sent through a high pass filter in order to remove DC and subsonic components of the signals. The signals are then decomposed into band pass signals by linear phase filters that are distributed equally over a perceptual

pitch scale. A frequency dependent weighting is applied to the band pass signals in order to model the spectral characteristics of the outer and middle ear. The level dependent spectral resolution of the auditory filters is modelled by a frequency domain convolution of the outputs with a level dependent spreading function.

The envelopes of the signals are calculated using the Hilbert-transform of the band pass signals ("rectification") and a time domain convolution with a window function is applied in order to model backward masking. Then, a frequency dependent offset is added that accounts for internal noise in the auditory system and models the threshold in quiet. Finally, a second time domain convolution is carried out using an exponential spreading function that accounts for forward masking.

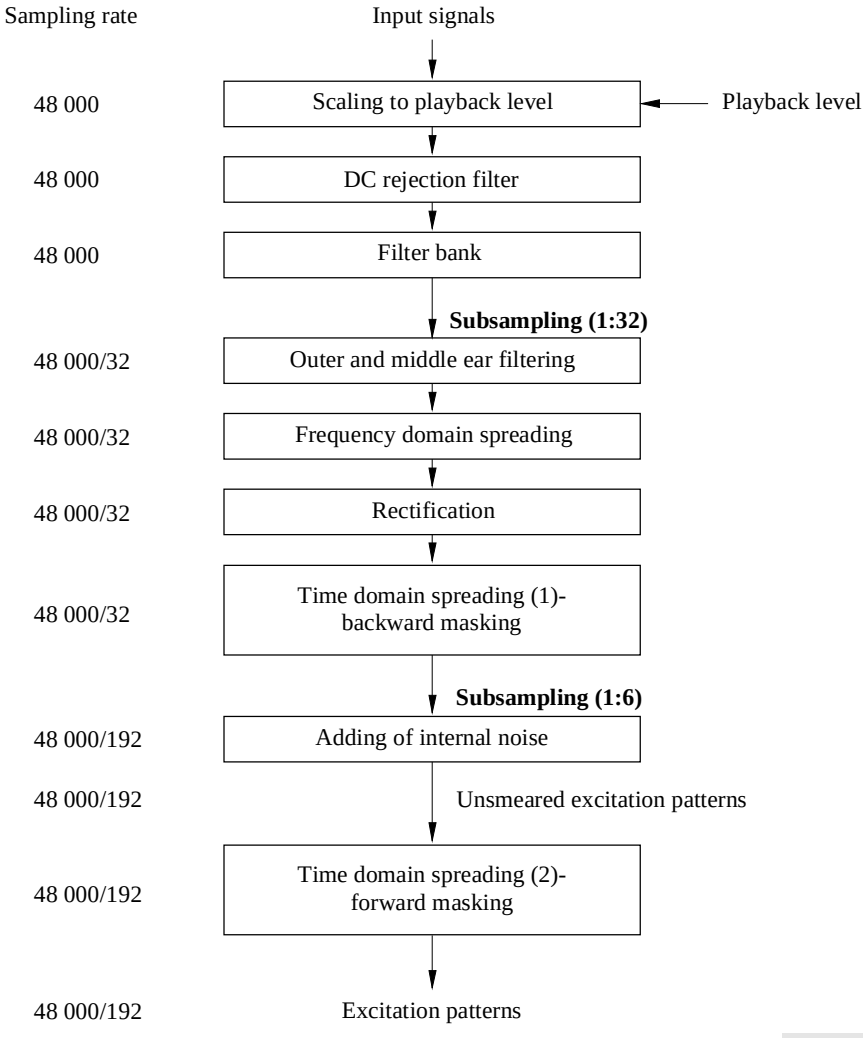
The now obtained *excitation patterns* are used to compute *specific loudness patterns* and the patterns before the final time domain spreading ("*unsmeared excitation patterns*") are used to calculate *modulation patterns*. These, together with the excitation patterns themselves, are the basis on which the model values are calculated. In order to separate the influence of the steady state frequency response of the device under test from other distortions, the excitation patterns of Signal Under Test and Reference Signal are also spectrally adapted to each other ("adaptation"). The modulation patterns and specific loudness patterns are calculated from both the adapted and the non-adapted excitation patterns.

2.2.2 Subsampling

At the output of the filterbank the signals are downsampled by a factor of 32 and after the first time domain spreading, the signals are downsampled by a factor of 6 (see Figure 11).

FIGURE 11

Subsampling in the filterbank-based peripheral ear model



2.2.3 Setting of playback level

The scaling factor for the input is calculated from the assumed playback level of a full scale input signal by:

$$fac = \frac{10^{L_{\max}/20}}{32767} \tag{27}$$

In case the exact playback level is unknown, it is recommended to set L_{MAX} to 92 dB_{SPL}.

2.2.4 DC rejection filter

As the filter bank is sensitive to subsonics in the input signals, a DC rejection filter is applied to the input signals. A fourth order Butterworth high pass filter with a cut-off frequency of 20 Hz is used. The filter is realized as a cascade of two second order IIR-filters.

$$y_n = x_n - 2x_{n-1} + x_{n-2} + b_1 y_{n-1} + b_2 y_{n-2} \tag{28}$$

where the coefficients for the first block are:

$$b_{1,2} = 1.99517,\ -0.995174$$

and the coefficients for the second block are:

$$b_{1,2} = 1.99799,\ -0.997998.$$

2.2.5 Filter bank

The filter bank consists of 40 filter pairs for each channel of Signal Under Test and Reference Signal. The filters are equally spaced and have constant absolute bandwidth when related to an auditory pitch scale. Each filter pair consists of two filters with equal frequency response but a 90° difference in the phase response. Thus, the output of the second filter represents the Hilbert-transform of the output of the first filter (or the imaginary part, if the first filter is assumed to represent the real part of a complex signal). The envelopes of their impulse responses have a $\cos^2$ shape. The filters are defined by Table 9 and equation 29 (where k is the index of the filter, n is the index of the time sample and T is the time between two samples: $T=1/48000$). They can be realized as FIR filters using the values $h_{re}(k,n)$ and $h_{im}(k,n)$ as coefficients. When the input signals are time limited, the filter outputs can also be computed by very fast recurrent algorithms.

$$\begin{aligned}
h_{re}(k,n) &= \frac{4}{N[n]} \cdot \sin^2\left(\pi \cdot \frac{n}{N[k]}\right) \cdot \cos\left(2\pi \cdot f_c[k] \cdot \left(n - \frac{N[k]}{2}\right) \cdot T\right) \\
h_{im}(k,n) &= \frac{4}{N[n]} \cdot \sin^2\left(\pi \cdot \frac{n}{N[k]}\right) \cdot \sin\left(2\pi \cdot f_c[k] \cdot \left(n - \frac{N[k]}{2}\right) \cdot T\right) \qquad 0 \le n < N[k] \\
h_{re}(k,n) &= h_{im}(k,n) = 0 \qquad n < 0,\ n \ge N[k]
\end{aligned} \tag{29}$$

The centre frequencies range from 50 Hz to 18 000 Hz. The auditory pitch scale is calculated from an approximation given by [Schroeder et al, 1979]:

$$z / Bark = 7 \cdot \operatorname{arsinh}\left(\frac{f / Hz}{650}\right) \tag{30}$$

The pitch units are named *Bark* (although this scale does not exactly represent the Bark scale as defined by [Zwicker and Feldtkeller, 1967]).

TABLE 9

Centre frequency, length of impulse response and additional delay for each filter

Index of filter	Centre frequency/Hz	Length of impulse response/samples	Additional delay/samples
k	$f_c[k]$	$N[k]$	$D[k]$
0	50.00	1456	1
1	116.19	1438	10
2	183.57	1406	26
3	252.82	1362	48
4	324.64	1308	75
5	399.79	1244	107
6	479.01	1176	141
7	563.11	1104	177
8	652.97	1030	214
9	749.48	956	251
10	853.65	884	287
11	966.52	814	322
12	1089.25	748	355
13	1223.10	686	386
14	1369.43	626	416
15	1529.73	570	444
16	1705.64	520	469
17	1898.95	472	493
18	2111.64	430	514
19	2345.88	390	534
20	2604.05	354	552
21	2888.79	320	569
22	3203.01	290	584
23	3549.90	262	598
24	3933.02	238	610
25	4356.27	214	622
26	4823.97	194	632
27	5340.88	176	641
28	5912.30	158	650
29	6544.03	144	657
30	7242.54	130	664
31	8014.95	118	670

TABLE 9 (end)

Index of filter	Centre frequency/Hz	Length of impulse response/samples	Additional delay/samples
k	$f_c[k]$	$N[k]$	$D[k]$
32	8869.13	106	676
33	9813.82	96	681
34	10858.63	86	686
35	12014.24	78	690
36	13292.44	70	694
37	14706.26	64	697
38	16270.13	58	700
39	18000.02	52	703

In order to have equal delays for all filters, the input for each filter is delayed by D samples, where D is half the difference between the length of its impulse response and the length of the impulse response of the filter with the longest impulse response[4]:

$$\mathrm{D}[\mathrm{k}] = 1 + \frac{1}{2}\left(N[0] - N[k]\right). \tag{31}$$

The filter outputs are subsampled by a factor of 32, i.e. output values are calculated each 32nd input sample for all filters[5].

2.2.6 Outer and middle ear filtering

The frequency response of the outer and middle ear is modelled by a frequency dependent weighting function that is applied to the filter outputs (equation 32).

$$W[k]/\,dB = -0.6 \cdot 3.64 \cdot \left(\frac{f_c[k]}{kHz}\right)^{-0.8} + 6.5 \cdot e^{-0.6 \cdot \left(\frac{f_c[k]}{kHz} - 3.3\right)^2} - 10^{-3} \cdot \left(\frac{f_c[k]}{kHz}\right)^{3.6} \tag{32}$$

2.2.6.1 Pseudocode

```
/* inputs */
out_re,out_im   :   filter bank outputs (real and imaginary part)
W               :   weighting function (see (32))
/* outputs */
out_re,out_im   :   filter bank outputs
/* intermediate values */
k               :   index of the filter
Wt              :   weighting factor
```

---

4 The additional delay of one sample is not necessary for implementation. Note that the reference implementation used for compliance testing includes this additional delay.

5 Actually, the envelopes of the filters in the upper frequency bands are not necessarily fulfilling the sampling theorem. Even though aliasing will only occur under very particular conditions (i.e. high frequency components modulated with frequencies larger than 1.5 kHz) and problems related to such effects never occurred within the known databases, it should be stated that aliasing problems might occur, especially when using artificial test signals.

```
/* outer and middle ear filtering */
for(k=0..39)
{
        Wt    = pow(10,W[k]/20)
        out_re[k] *= Wt;
        out_im[k]*= Wt;
}
```

2.2.7 Frequency domain spreading

The output values of the filter bank are smeared out over frequency using a level dependent spreading function. The spreading function is a two sided exponential. The lower slope is always 31 dB/Bark and the upper slope varies between -24 and -4 dB/Bark.

The upper slope, ***s[k]***, is calculated according to:

$$\frac{s[k]}{dB/Bark} = \min\left(-4, -24 - \frac{230Hz}{f_c[k]} + 0.2 \cdot L[k]/dB\right) \tag{33}$$

The level *L[k]* is calculated independently for each filter channel by taking the squared absolute value of the filter output and transforming it to the dB-scale. The centre frequencies, $f_c[k]$ are taken from Table 9. The linear representations of the slopes are smoothed over time by a first order low pass filter with a time constant of 100 ms.

The spreading is carried out independently for the filters representing the real part of the signals and the filters representing the imaginary parts of the signals (equation 29). The frequency spreading is first carried out for the (level dependent) upper slope and after that for the lower slope using a first order IIR-filter algorithm.

2.2.7.1 Pseudocode

/* inputs */		
out_re,out_im	:	filter bank outputs (real and imaginary part)
z[ ]	:	critical band rates for the centre frequencies of the filter bands in Bark (according to Table 9 and equation 30)
/* outputs */		
A_re,A_im	:	output patterns
/* intermediate values */		
j,k	:	index of the filter
a,b	:	coefficients for temporal smoothing
dist	:	constant for calculating the crosstalk
L[ ]	:	level at each filter output
s[ ]	:	local slope for upward spreading
d1,d2	:	buffers
/* static */		(values from the preceeding frame are preserved; values are initialized with zeros at the start of the measurement)
cl, cu[ ]	:	spreaded fraction of the signal

```
/* level dependent upward spreading */
dist      = pow(0.1,(z[39]-z[0])/(39*20));
/* (z[39]-z[0])/39 is the distance in Bark between two adjacent filter bands */
a         = exp(-32/(48000*0.1));
b         = 1 - a;
for(k=0..39)
{
        A_re[k]    = out_re[k];
        A_im[k]   = out_im[k];
}
for(k=0..39)
{
        /* calculate level dependent slope */
        L[k]   = 10*log10(out_re[k]* out_re[k] + out_im[k]* out_im[k];
        s[k]   = max(4,(24 + 230/fcentre[k] - 0.2*L[k]));
        /* calculate spreaded fraction and smooth it over time*/
        cu[k] = a*pow(dist,s[k])+b*cu[k];
        /* spreading of band k */
        d1     = out_re[k]
        d2     = out_im[k]
        for(j=k+1..39)
        {
         d1   *= cu[k];
         d2   *= cu[k];
         A_re[j]   += d1;
         A_im[j]  += d2;
        }
}
/* downward spreading */
cl        = pow(dist,31);
d1        = 0;
d2        = 0;
for(k=0..39)
{
        /* spreading of band k */
        d1     = d1 * cl + A_re[k];
        d2     = d2 * cl + A_im[k];
        A_re[k]    = d1;
        A_im[k]   = d2;
}
```

2.2.8 Rectification

The energies at the filter outputs are calculated by adding the squared values of the filter representing the real part of the signal and the one representing the imaginary part of the signal.

$$E_0[k,n] = A_{re}[k,n]^2 + A_{im}[k,n]^2 \tag{34}$$

All the following operations are carried out on these energies.

2.2.9 Time domain smearing (1) - Backward masking

In order to model backward masking, the energies at the filter outputs are smeared out over time by an FIR-filter with a $\cos^2$ shaped impulse response with 12 taps (which corresponds to a filter response of 384 samples at the input sampling rate of the filter bank). After the time smearing the outputs are downsampled by a factor of 6. The resulting values are multiplied by a calibration factor $cal_1 = 0.9761$ in order to get the appropriate output levels for the given playback level.

$$E_1[k,n] = \frac{0.9761}{6} \cdot \sum_{i=0}^{11} E_0[k,6n-i] \cdot \cos^2\left(\pi \cdot \frac{(i-5)}{12}\right) \tag{35}$$

2.2.10 Adding of internal noise

After the first time domain spreading, a frequency dependent offset E_{Thres} is added to the energies in each filter channel.

$$E_{Thres}[k] = 10^{0.4 \cdot 0.364 \cdot \left(\frac{f_c[k]}{kHz}\right)^{-0.8}} \tag{36}$$

$$E_2[k,n] = E_1[k,n] + E_{Thres}[k,n] \tag{37}$$

The patterns at this stage of processing, $E_2[k,n]$, are used later on for the computation of modulation patterns and are referred to as "*unsmeared excitation patterns*".

2.2.11 Time domain smearing (2) – Forward masking

In order to model forward masking, the energies in each filter channel are smeared out over time by first order low pass filters. The time constants depend on the centre frequency of each filter (as given in Table 6) and are calculated according to:

$$\tau = \tau_0 + \frac{100Hz}{f_c[k]} \cdot (\tau_{100} - \tau_{\min}) \quad \left| \begin{array}{lcl} \tau_{100} & = & 0.020\ s \\ \tau_{\min} & = & 0.004\ s \end{array} \right. \tag{38}$$

The first order low pass filters are computed according to:

$$E[k,n] = a \cdot E[k,n-1] + (1-a) \cdot E_2[k,n] \tag{39}$$

where a is calculated from the above time constants by:

$$a = e^{-\frac{192}{48000 \cdot \tau}} \tag{40}$$

The patterns at this stage of processing, $E[k,n]$, are referred to as "*excitation patterns*".

3 Pre-processing of excitation patterns

Most of the computations described in this section are used with the filter bank-based ear model as well as with the FFT-based ear model. As the subsampling factor and the number of frequency bands are different between the two ear models, constants depending on this factor are described using the ear model dependent variables ***StepSize*** and ***Z***. For the FFT-based ear model the value of ***StepSize*** is ***1024*** and ***Z*** is either ***55*** (*Advanced Version*) or ***109*** (*Basic Version)*. For the filter bank-based ear model the value of ***StepSize*** is ***192*** and ***Z*** is ***40***.

3.1 Level and pattern adaptation

In order to compensate for level differences and linear distortions between test and Reference Signal, the average levels of test and Reference Signal are adapted to each other.

In the first step, the energies in each filter channel are smoothed by first order low pass filters. The time constants depend on the centre frequencies of the filters and are chosen as:

$$\tau = \tau_{\min} + \frac{100\,Hz}{f_c[k]} \cdot (\tau_{100} - \tau_{\min}) \quad \Big| \quad \begin{array}{lll} \tau_{100} & = & 0.050\ s \\ \tau_{\min} & = & 0.008\ s \end{array} \tag{41}$$

The first order low pass filters are computed according to:

$$P_{Ref}[k,n] = a \cdot P_{Ref}[k,n-1] + (1-a) \cdot E_{Ref}[k,n] \tag{42}$$

$$P_{Test}[k,n] = a \cdot P_{Test}[k,n-1] + (1-a) \cdot E_{Test}[k,n] \tag{43}$$

where E_{Test} and E_{Ref} are the excitation patterns that are to be adapted to each other and a is calculated from the time constants by:

$$a = e^{-\frac{StepSize}{48000 \cdot \tau}} \tag{44}$$

3.1.1 Level adaptation

From the low passed input patterns P_{test} and P_{ref} a momentary correction factor ***LevCorr*** is calculated by:

$$LevCorr[n] = \left(\frac{\sum_{k=0}^{Z-1} \sqrt{P_{Test}[k,n] \cdot P_{Ref}[k,n]}}{\sum_{k=0}^{Z-1} P_{Test}[k,n]} \right)^2 \tag{45}$$

If the correction factor is larger than one, the Reference Signal is divided by the correction factor, otherwise the test signal is multiplied by the correction factor.

$$E_{L,Ref}[k,n] = E_{Ref}[k,n] / LevCorr[n] \quad | \quad LevCorr[n] > 1 \tag{46}$$

$$E_{L,Test}[k,n] = E_{Test}[k,n] \cdot LevCorr[n] \quad | \quad LevCorr[n] \leq 1 \tag{47}$$

3.1.2 Pattern adaptation

Correction factors for each channel are calculated by comparing the temporal envelopes of the filter outputs of the test and Reference Signals:

$$R[k,n] = \frac{\sum_{i=0}^{n} a[k]^i \cdot E_{L,Test}[k,n-i] \cdot E_{L,Ref}[k,n-i]}{\sum_{i=0}^{n} a[k]^i \cdot E_{L,Ref}[k,n-i] \cdot E_{L,Ref}[k,n-i]} \tag{48}$$

The values for a are calculated as above (equation 44) from the time constants given by equation 41. If $R[k,n]$ is larger than one, the correction factor for the test signal is set to $R[k,n]^{-1}$ and the correction factor for the Reference Signal is set to one. In the opposite case, the correction factor for the Reference Signal is set to $R[k,n]$ and the correction factor for the test signal is set to 1.

$$\begin{aligned} &R_{Test}[k,n] = \frac{1}{R[k,n]}, \quad R_{Ref}[k,n] = 1 && \Big| R[k,n] \geq 1 \\ &R_{Test}[k,n] = 1, \qquad\quad R_{Ref}[k,n] = R[k,n] && \Big| R[k,n] < 1 \end{aligned} \tag{49}$$

If the denominator of (48) is zero (and $R[k,n]$ thus would be undefined) and the numerator is larger than zero $R_{Test}[k,n]$ is set to zero and $R_{Ref}[k,n]$ is set to one. When the numerator of (48) is zero as well, the ratios $R_{Test}[k,n]$ and $R_{Ref}[k,n]$ are copied from the frequency band below. If there is no frequency band below (i.e. $k=0$) then the ratios $R_{Test}[k,n]$ and $R_{Ref}[k,n]$ are set to one.

The correction factors are averaged over M filter channels and smoothed over time (equation 50) using the same time constants as given above (equations 41 to 44). The width of the frequency window M is 3 for the filter bank-based ear model. For the FFT-based ear model it is 4 (Advanced Version) or 8 (Basic Version) respectively.

$$\begin{aligned} PattCorr_{Test}[k,n] &= a \cdot PattCorr_{Test}[k,n-1] + (1-a) \cdot \frac{1}{M} \cdot \sum_{i=-M_1}^{M_2} R_{Test}[k+i,n] \\ PattCorr_{Ref}[k,n] &= a \cdot PattCorr_{Ref}[k,n-1] + (1-a) \cdot \frac{1}{M} \cdot \sum_{i=-M_1}^{M_2} R_{Ref}[k+i,n] \end{aligned} \tag{50}$$

$$\begin{cases} M_1 = M_2 = \dfrac{M-1}{2} & | M \text{ odd} \\ M_1 = \dfrac{M}{2} - 1, \quad M_2 = \dfrac{M}{2} & | M \text{ even} \end{cases}$$

At the borders of the frequency scale where the frequency window would exceed the range of filter bands, the width of the frequency window is reduced accordingly:

$$M_1 = \min(M_1, k), \qquad M_2 = \min(M_2, z - k - 1), \qquad M = M_1 + M_2 + 1 \tag{51}$$

The level adapted input patterns are weighted with the corresponding correction factors $PattCorr_{Test/Ref}[k,n]$ in order to obtain the spectrally adapted patterns.

$$E_{P,Ref}[k,n] = E_{L,Ref}[k,n] \cdot PattCorr_{Ref}[k,n] \tag{52}$$

$$E_{P,Test}[k,n] = E_{L,Test}[k,n] \cdot PattCorr_{Test}[k,n] \tag{53}$$

3.2 Modulation

From the *unsmeared excitation patterns*, $E_z[k,n]$, a simplified loudness is calculated by raising the excitation to a power of 0.3. This value and the absolute value of its temporal derivation are smeared out over time.

$$\overline{E}_{der}[k,n] = a \cdot \overline{E}_{der}[k,n-1] + (1-a) \cdot \frac{48000}{StepSize} \cdot \left| E_2[k,n]^{0.3} - E_2[k,n-1]^{0.3} \right| \tag{54}$$

$$\overline{E}[k,n] = a \cdot \overline{E}[k,n-1] + (1-a) \cdot E_2[k,n]^{0.3} \tag{55}$$

The values for a are calculated as in (44) from the time constants given by:

$$\tau = \tau_0 + \frac{100\,Hz}{f_c} \cdot (\tau_{100} - \tau_0) \quad \left| \begin{array}{lll} \tau_{100} & = & 0.050\ s \\ \tau_0 & = & 0.008\ s \end{array} \right. \tag{56}$$

From the resulting values, $\overline{E}_{der}$ and $\overline{E}$, a measure for the modulation of the envelope at each filter output is calculated:

$$Mod[k,n] = \frac{\overline{E}_{der}[k,n]}{1 + \overline{E}[k,n] / 0.3} \tag{57}$$

The values $\overline{E}$ are also used later on in the calculation of the modulation difference.

3.3 Loudness

The specific loudness patterns of the Signal Under Test and the Reference Signal are calculated according to the formula

$$N[k,n] = const \cdot \left(\frac{1}{s[k]} \cdot \frac{E_{Thres}[k]}{10^4} \right)^{0.23} \cdot \left[\left(1 - s[k] + \frac{s[k] \cdot E[k,n]}{E_{Thres}[k]} \right)^{0.23} - 1 \right] \tag{58}$$

as given in [Zwicker and Feldtkeller, 1967]. The overall loudness of the Signal Under Test and the Reference Signal is calculated as the sum across all filter channels of all specific loudness values above zero.

$$N_{total}[n] = \frac{24}{Z} \cdot \sum_{k=0}^{Z-1} \max(N[k,n], 0) \tag{59}$$

The scaling constant is chosen as $const = 1.07664$ for the FFT-based peripheral ear model and $const = 1.26539$ for the filter bank-based peripheral ear model in order to give an overall loudness of one sone for a 40 dB_{SPL} sine tone at 1 kHz. Threshold index s and excitation at threshold E_{Thres} are calculated according to:

$$E_{Thres}[k] = 10^{0.364 \cdot \left(\frac{f}{1\ kHz} \right)^{-0.8}} \tag{60}$$

and

$$s[k] = 10^{\frac{1}{10}\left(-2 - 2.05 \cdot atn\left(\frac{f}{4\ kHz} \right) - 0.75 \cdot atn\left(\left(\frac{f}{1600\ Hz} \right)^2 \right) \right)} \tag{61}$$

respectively.

NOTE – Due to the different peripheral ear models, the loudness calculated here is not identical to the loudness as defined in ISO 532 “Acoustics – Method for calculating loudness levels”, 1995.

3.4 Calculation of the error signal

The error signal is only computed in the FFT-based model. It is calculated in the frequency domain by taking the difference between the outer and middle ear filtered power spectra of the reference and test signal (see § 2.1.4).

$$F_{noise}[k_f, n] = \left\| F_{eref}[k_f, n] \right| - \left| F_{etest}[k_f, n] \right\| \quad (62)$$

F_{noise} is mapped to the pitch domain using the algorithm described in § 2.1.5.

The output of this algorithm , $P_{noise}[n, k]$, are referred to as "*Noise patterns*".

4 Calculation of Model Output Variables

4.1 Overview

TABLE 10

Overview of the Model Output Variables used for the prediction of the basic audio quality

Model Output Variable (MOV)	Calculated in ... ear model		Used in ... version	
	FFT	filter bank	basic	advanced
$WinModDiff1_B$	yes	no	yes	no
$AvgModDiff1_B$	yes	no	yes	no
$AvgModDiff2_B$	yes	no	yes	no
$RmsModDiff_A$	no	yes	no	yes
$RmsNoiseLoud_B$	yes	no	yes	no
$RmsNoiseLoudAsym_A$	no	yes	no	yes
$AvgLinDist_A$	no	yes	no	yes
$BandwidthRef_B$	yes	no	yes	no
$BandwidthTest_B$	yes	no	yes	no
Total NMR_B	yes	no	yes	no
$RelDistFrames_B$	yes	no	yes	no
Segmental NMR_B	yes	no	no	yes
$MFPD_B$	yes	no	yes	no
ADB_B	yes	no	yes	no
EHS_B	yes	no	yes	yes

4.2 Modulation difference

Differences in the modulation of the temporal envelopes of the Signal Under Test and the Reference Signal are measured by computing a local modulation difference for each filter channel (equation 63), where Mod_{test} and Mod_{Ref} are derived from applying equation (57) to the reference R_{test} signal.

$$ModDiff[k,n] = w \cdot \frac{\left|Mod_{test}[k,n] - Mod_{Ref}[k,n]\right|}{offset + Mod_{Ref}[k,n]}$$

$$\begin{cases} w = 1.0 & \left|Mod_{test}[k,n] > Mod_{Ref}[k,n]\right| \\ w = negWt & \left|Mod_{test}[k,n] < Mod_{Ref}[k,n]\right| \end{cases} \quad (63)$$

A momentary modulation difference is calculated as the average of the local modulation differences over all filter channels (equation 64).

$$ModDiff[n] = \frac{100}{Z}\sum_{k=0}^{Z-1} ModDiff[k,n] \quad (64)$$

The threshold in quiet is taken into account by a level dependent weighting factor (equation 65) calculated from the modified excitation patterns for the reference signal as given in equation 55 and the internal noise function as defined in equation 36 for the filter bank-based ear model and equation 13 for the FFT-based ear model.

$$TempWt[n] = \sum_{k=0}^{Z-1} \frac{\overline{E}_{ref}[k,n]}{\overline{E}_{ref}[k,n] + levWt \cdot E_{Thres}[k]} \wedge 0.3 \quad (65)$$

The temporal averaging of the momentary modulation differences **`ModDiff[n]`** using the weighting factors **`TempWt[n]`** is described in § 5.2 (*Temporal averaging*). The values for the constants **`negWt`**, **`offset`** and **`levWt`** are given in Table 11.

TABLE 11

Model Output Variables estimating the overall modulation difference

MOV (Xxx=Win/Avg/Rms)	negWt	offset	levWt
$XxxModDiff1_B$	1	1	100
$XxxModDiff2_B$	0.1	0.01	100
$XxxModDiff_A$	1	1	1

4.2.1 RmsModDiff$_A$

The Model Output Variable $RmsModDiff_A$ is the squared average of the modulation difference calculated from the filter bank-based ear model. See § 5.2.2 for temporal averaging and Table 11 for constants.

4.2.2 WinModDiff1$_B$

The Model Output Variable *WinModDiff1*$_B$ is the windowed average of the modulation difference calculated from the FFT-based ear model. See § 5.2.3 for temporal averaging and Table 11 for constants. The temporal weighting factor given in equation 65 is not applied for this MOV.

4.2.3 AvgModDiff1$_B$ and AvgModDiff2$_B$

The Model Output Variables *AvgModDiff1*$_B$ and *AvgModDiff2*$_B$ are the linear average of the modulation difference calculated from the FFT-based ear model. The difference between *AvgModDiff2*$_B$ and *AvgModDiff1*$_B$ is that the constants are chosen differently. See § 5.2.1 for temporal averaging and Table 11 for constants.

4.3 Noise loudness

These Model Output Variables estimate the partial loudness of additive distortions in the presence of the masking Reference Signal. The formula for the partial loudness (equation 66) is designed to yield the specific loudness of the noise according to [Zwicker and Feldtkeller, 1967] if no masker is present and to yield something like the ratio between noise and mask if the noise is very small compared to the masker.

The partial noise loudness is calculated according to:

$$NL[k,n] = \left(\frac{1}{s_{test}} \cdot \frac{E_{Thres}}{E_0} \right)^{0.23} \cdot \left[\left(1 + \frac{\max\left(s_{test} \cdot E_{test} - s_{ref} \cdot E_{ref}, 0 \right)}{E_{Thres} + s_{ref} \cdot E_{ref} \cdot \beta} \right)^{0.23} - 1 \right] \tag{66}$$

where E_0 is always 1, E_{Thres} is the internal noise function $E_{Thres}[k]$ as defined in (36) and s is calculated according to:

$$s = ThresFac_0 \cdot Mod[k,n] + S_0 \tag{67}$$

When not described differently, the *spectrally adapted excitation patterns* (see § 3.1) are used as inputs: $E_{Test} = E_{P,Test}[k,n]$ and $E_{Ref} = E_{P,Ref}[k,n]$. The coefficient β, which determines the amount of masking, is calculated by:

$$\beta = \exp\left(-\alpha \cdot \frac{E_{test} - E_{ref}}{E_{ref}} \right) \tag{68}$$

The values of the momentary noise loudness are not taken into account until 50 ms after the overall loudness for either the left or the right audio channel has once exceeded a value of $N_{Thres} = 0.1$ sone for both test and Reference Signal (see § 5.2.4.2).

In the spectral averaging, the momentary values are normalized by the number of filter bands per critical band instead of the total number of filter bands, i.e. the result of the spectral averaging is multiplied by a factor of 24.

If the momentary noise loudness is below a threshold value NL_{min} it is set to zero.

TABLE 12

Model Output Variables estimating the overall noise loudness

MOV (Xxx=Win/Avg/Rms)	α	ThresFac$_0$	S$_0$	NL$_{min}$
XxxMissingComponents$_B$	1.5	0.15	1	0
XxxNoiseLoud$_B$	1.5	0.15	0.5	0
XxxMissingComponents$_A$	1.5	0.15	1	0
XxxNoiseLoud$_A$	2.5	0.3	1	0.1
XxxLinDist$_A$	1.5	0.15	1	0

4.3.1 RmsNoiseLoud$_A$

The Model Output Variable $RmsNoiseLoud_A$ is the squared average of the noise loudness calculated from the filter bank-based ear model. See § 5.2.2 for temporal averaging and Table 12 for constants.

4.3.2 RmsMissingComponents$_A$

The Model Output Variable $RmsMissingComponents_A$ is the squared average of the noise loudness calculated from the filter bank-based ear model. It is computed with the excitation patterns of test and Reference Signals interchanged in order to yield the loudness of components in the Reference Signal that are lost in the test signal. See § 5.2.2 for temporal averaging and Table 12 for constants.

4.3.3 RmsNoiseLoudAsym$_A$

The Model Output Variable $RmsNoiseLoudAsym_A$ is the weighted sum of the squared averages of the noise loudness (§ 4.3.1) and the loudness of lost signal components (§ 4.3.2), both calculated from the filter bank-based ear model.

$$RmsNoiseLoudAsym = RmsNoiseLoud + 0.5 \cdot RmsMissingComponents. \quad (69)$$

4.3.4 AvgLinDist$_A$

The Model Output Variable $AvgLinDist_A$ measures the loudness of the signal components lost during the spectral adaptation of the Signal Under Test and the Reference Signal. It uses the spectrally adapted excitation of the Reference Signal as the reference and the unadapted excitation of the reference as the test signal. This MOV is calculated from the filter bank-based ear model. See § 5.2.1 for temporal averaging and Table 12 for constants.

4.3.5 RmsNoiseLoud$_B$

The Model Output Variable $RmsNoiseLoud_B$ is the squared average of the noise loudness calculated from the FFT-based ear model. See § 5.2.2 for temporal averaging and Table 12 for constants.

4.4 Bandwidth

These Model Output Values estimate the mean bandwidth of the Signal Under Test and the Reference Signal in FFT lines.

For each frame the Local Bandwidth $Bw_{Ref}[n]$ and $Bw_{Test}[n]$ is calculated according to the pseudocode below.

4.4.1 Pseudocode

```
/* inputs */
FLevRef[], FlevelTest[]      :    level of FFT outputs in dB
/* outputs */
BwRef, BwTest                :    output patterns
/* intermediate values */
k                            :    index of FFT lines
ZeroThreshold                :    Bandwidth threshold
```

```
ZeroThreshold = -1.0E-10;
BwRef = BwTst = 0.0;
for(k=921;k<1024;k++)
{
```

```
  ZeroThreshold=max(ZeroThreshold,FLevelTst(k));
}


        for (k = 920; k>=0; k--)
        {
          if (FLevelRef[k] >= 10.0+ZeroThreshold)
         {
              BwRef = k+1;
         }
              break;
        }

        for (k = BwRef-1; k>=0; k--)
        {
          if(FLeveltest[k] >= 5.0+ZeroThreshold)
         {
              BwTest=k+1;
              break;
         }
        }
```

4.4.2 BandwidthRef$_B$ and BandwidthTest$_B$

BandwidthRef$_B$ is the linear average of BwRef, and BandwidthTest$_B$ is the linear average of BwTest. For averaging, only frames with BwRef > 346 are taken into account. Frames with low energy at the beginning and the end of the items are ignored (see § 5.2.4.4). See § 5.2.1 for temporal averaging.

4.5 Noise-to-Mask Ratio

The following model values are calculated from the noise and masking values.

The local NMR of the current frame ***n*** is defined as:

$$NMR_{local}[n] = 10 * \log_{10} \frac{1}{Z} \sum_{k=0}^{Z-1} \frac{P_{noise}[k,n]}{M[k,n]} \tag{70}$$

4.5.1 Total NMR$_B$

The Model Output Variable *Total NMR$_B$* is the linear average of the noise-to-mask ratio using

$$NMR_{tot} = 10 * \log_{10} \frac{1}{N} \sum_{n} \left(\frac{1}{Z} \sum_{k=0}^{Z-1} \frac{P_{noise}[k,n]}{M[k,n]} \right) \tag{71}$$

Frames with low energy at the beginning and the end of the items are ignored (see § 5.2.4.4).

4.5.2 Segmental NMR_A

The Model Output Variable *Segmental NMR_A* is the linear average of the local NMR. See § 5.2.1 for temporal averaging.

Frames with low energy at the beginning and the end of the items are ignored (see § 5.2.4.4).

4.6 Relative Disturbed $Frames_B$

The Model Output Variable *Relative Disturbed $Frames_B$* (abbreviation: $RelDistFrames_B$) represents the number of frames with:

$$\max_{\forall k}\left(10\cdot\log\left(\frac{P_{noise}[k,n]}{M[k,n]}\right)\right)\geq 1.5dB \qquad k\in[0,Z-1]$$

related to the total number of frames of the item.

Frames with low energy at the beginning and the end of the items are ignored (see § 5.2.4.4).

4.7 Detection probability

The MOVs defined in this section are based on $\tilde{E}[k, n]$ (***k*** band, ***n*** frame), which are the *excitation patterns* E[k, n] expressed in dB:

$$\tilde{E}[k,n]=10\cdot\log_{10}\left(E[k,n]\right) \tag{72}$$

For each frame ***n***:

The following steps are done independently for each channel ***c*** (values of ***c*** are left and right). The *logarithmic excitation patterns* are $\tilde{E}_{ref}[k,n]$ for the Reference Signal and $\tilde{E}_{test}[k,n]$ for the Signal Under Test respectively.

For each band ***k***:

– Calculate the asymmetric average excitation.

$$L[k,n] = 0.3\cdot\max(\tilde{E}_{ref}[k,n],\tilde{E}_{test}[k,n])+0.7\cdot\tilde{E}_{test}[k,n] \tag{73}$$

– Calculate the effective detection step size s. The following formula is an approximation to the just noticeable level difference as measured by [Zwicker and Fastl, 1990].

If L[k,n]>0:

$$s[k,n] = 5.95072\cdot((6.39468)/L[k,n])^{1.71332} + 9.01033\cdot 10^{-11}\cdot L[k,n]^4 + 5.05622\cdot 10^{-6}\cdot L[k,n]^3 - 0.00102438\cdot L[k,n]^2 + 0.0550197\cdot L[k,n] - 0.198719$$

else

$$s[k,n]=1.0 \qquad 10^{30} \tag{74}$$

– Calculate the signed error e

$$e[k,n]=\tilde{E}_{ref}[k,n]-\tilde{E}_{test}[k,n] \tag{75}$$

– If $\tilde{E}_{ref}[k,n]>\tilde{E}_{test}[k,n]$, then the steepness of slope ***b*** is set to 4.0 otherwise it is set to 6.0. This models the effect that an increase of the signal energy of the Signal Under Test in comparison to the Reference Signal is more striking than a decrease.

– Calculate the scale factor ***a***.

$$a[k,n] = \frac{10^{\frac{\log_{10}(\log_{10}(2.0))}{b}}}{s[k,n]} \tag{76}$$

– Calculate the probability of detection. Equation 76 sets the scale factor ***a*** such that if ***e[k,n]*** equals ***s[k,n]***, ***p_c[k,n]*** becomes ***0.5***.

$$p_c[k,n] = 1 - 10^{(-(a[k,n]\cdot e[k,n])^b)} \tag{77}$$

– Calculate the total number of steps above the threshold:

$$q_c[k,n] = \frac{|\mathrm{INT}(e[k,n])|}{s[k,n]} \tag{78}$$

– The binaural detection probability is:

$$p_{bin}[k, n] = \max(p_{left}[k, n], p_{right}[k, n]) \tag{79}$$

– The number of steps above threshold for the binaural channel is:

$$q_{bin}[k, n] = \max(q_{left}[k, n], q_{right}[k, n]) \tag{80}$$

The total probability of detection of channel ***c*** of frame ***n*** is:

$$P_c[n] = 1 - \prod_{\forall k}(1 - p_c[k,n]) \ , \tag{81}$$

where ***c*** can be either ***left***, ***right*** or ***bin***. The total number of steps above threshold for channel ***c*** of frame ***n*** is:

$$Q_c[n] = \sum_{\forall k} q_c[k,n] \tag{82}$$

4.7.1 Maximum Filtered Probability of Detection ($MFPD_B$)

A smoothed version of the detection probability for each channel ***c*** is calculated:

$$\tilde{P}_c[n] = (1 - c_0) \cdot P_c[n] + c_0 \cdot \tilde{P}_c[n-1] \tag{83}$$

where ***P_c[-1]=0***. The constant ***c_0*** depends on ***StepSize***:

$$c_0 = 0.9^{\mathrm{StepSize}/1024} \tag{84}$$

c_0 reduces the sensitivity to very short distortions.

The maximum filtered probability of detection (abbreviation: MFPD) is calculated:

$$PM_c[n] = \max\left(PM_c[n-1]\cdot c_1 , \tilde{P}_c[n]\right) \tag{85}$$

where ***PM_c[-1]*** is zero. The constant ***c_1*** depends on ***StepSize***:

$$c_1 = 0.99^{\mathrm{StepSize}/1024.} \tag{86}$$

c_1 models the effect that distortions at the beginning of a excerpt of audio are less severe than at the end of the excerpt due to forgetting. Note that this constant is useful for modelling listening tests where the subjects are not allowed to select shorter parts of the excerpt. For the present model, which is calibrated using data from listening tests according to Recommendation ITU-R BS.1116-1, c_1 should be 1.0.

The MOV *MFPD* is the value of ***PM_{bin}[n]*** for the last frame.

4.7.2 Average distorted block[6] (ADB_B)

The number of valid frames with a probability of detection of the binaural channel $P_{bin}[n]$ above 0.5 is counted ($n_{distorted}$).

For all valid frames the total number of steps above threshold of the binaural channel $Q_{bin}[n]$ is calculated:

$$Q_{sum} = \sum_{\forall n} Q_{bin}[n]$$

The distortion of the average distorted block ADB is calculated:

– if $n_{distorted}$ is zero then ADB = 0 (no distortion audible);

– if $n_{distorted}>0$ and $Q_{sum}>0$ then $ADB = \log_{10}((Q_{sum})/n_{distorted}))$;

– if $n_{distorted}>0$ and Q_{sum} is zero then ADB = –0.5.

4.8 Harmonic structure of error

A Reference Signal containing strong harmonics (e.g., bass clarinet, harpsichord) has a spectrum characterized by a number of regularly spaced peaks separated by deep valleys. Under some conditions, the error signal may inherit that structure. For example, noise mixed with such a signal is more likely to remain unmasked where the signal is low in the spectral valleys. The resulting error spectrum would then contain a structure similar to the original spectrum but offset in frequency to correspond to the locations of the valleys. This structure may result in a distortion with tonal qualities that could increase the salience of the error.

The error is defined as the difference in the log spectra of the reference and processed signals. The excitation pattern from the psycho-acoustic model is not used here because the non-linear frequency to Bark transformation would smear the harmonic structure.

4.8.1 EHS_B

Harmonic structure magnitude is obtained by identifying and measuring the largest peak in the spectrum of the autocorrelation function. Each correlation is calculated as the cosine of the angle between two vectors according to the following formula, where $\vec{F}_0$ is the error vector and $\vec{F}_t$ is the same vector lagged by a certain amount.

$$C = \frac{\vec{F}_0 \cdot \vec{F}_t}{\left|\vec{F}_0\right| \cdot \left|\vec{F}_t\right|} \tag{87}$$

The maximum lag for obtaining the autocorrelation function is the largest power of two that is smaller than half the FFT frequency component number corresponding to 18 kHz.

For example, at a sampling rate of 48 kHz and an FFT window size of 2 048 samples, the FFT component corresponding to 18 kHz is (18/24) x 1 024 = 768. Therefore, the maximum lag would be 384. The actual number of lags would be 256, which is the largest power of 2 less than 384. The first value of the correlation function would be obtained by aligning $F_t[0]$ with $F_0[0]$ and the last by aligning $F_t[0]$ with $F_0[255]$.

The resulting vector of correlations is windowed with a normalized Hann window and, after removing the DC component by subtracting the average value, a spectrum is computed with an FFT. The maximum peak in the spectrum identifies the dominant frequency in the autocorrelation function. The average value of this maximum over frames multiplied by 1000.0 is the Error Harmonic Structure (EHS) variable.

---

6 The term “block” is equivalent to “frame” in this context.

5 Averaging

5.1 Spectral averaging

If not indicated differently in the descriptions of the Model Output Variables (§ 4), the following algorithm is used when averaging the local values over frequency bands.

5.1.1 Linear average

The linear average value is calculated by:

$$AvgS = \frac{1}{Z} \cdot \sum_{k=0}^{Z-1} S[k] \tag{88}$$

where ***S*** stands for the name of the Model Output Variable and ***Z*** is the number of frequency bands.

5.2 Temporal averaging

If not indicated differently in the descriptions of the Model Output Variables (§ 4), one or several of the following algorithms are used when averaging the momentary values over time. The temporal weighting factor (if applied) is denoted by the symbol ***W***, and ***Z*** is the number of frequency bands.

5.2.1 Linear average

The linear average value (prefix "*Avg*") is calculated by:

$$AvgX = \frac{1}{N} \cdot \sum_{n=0}^{N-1} X[n] \tag{89}$$

where ***X*** stands for the name of the Model Output Variable and ***N*** is the number of time samples for which momentary values of ***X*** have been calculated.

In case a temporal weighting is applied (see § 4.2 *Modulation difference*), the linear average is calculated according to:

$$AvgX = \frac{\sum_{n=0}^{N-1} W[n] \cdot X[n]}{\sum_{n=0}^{N-1} W[n]} \tag{90}$$

instead.

5.2.2 Squared average

The squared average value (prefix "*Rms*") is calculated by:

$$RmsX = \sqrt{\frac{1}{N} \cdot \sum_{n=0}^{N-1} X[n]^2} \tag{91}$$

where ***X*** stands for the name of the Model Output Variable and ***N*** is the number of time samples for which momentary values of ***X*** have been calculated.

When a temporal weighting is applied (see § 4.2 *Modulation difference*), the squared average is calculated according to:

$$RmsX = \sqrt{Z} \cdot \sqrt{\frac{\sum_{n=0}^{N-1} W[n]^2 \cdot X[n]^2}{\sum_{n=0}^{N-1} W[n]^2}} \tag{92}$$

instead.

5.2.3 Windowed average

The windowed average value (prefix "*Win*") is calculated by:

$$WinX = \sqrt{\frac{1}{N-L+1} \cdot \sum_{n=L-1}^{N-1} \left(\frac{1}{L} \cdot \sum_{i=0}^{L-1} \sqrt{X[n-i]} \right)^4} \tag{93}$$

where X stands for the name of the Model Output Variable, N is the number of time samples for which momentary values of X have been calculated and L is the length of the sliding time window in time samples. The window length is approximately 100 ms, i.e. L is 4 for the FFT-based ear model and 25 for the filter bank-based ear model.

5.2.4 Frame selection

5.2.4.1 Delayed averaging

For the Model Output Variables using this criterion, the values calculated during the first 0.5 seconds of the measurement are not taken into account in the temporal averaging. *Delayed averaging* is used for the following Model output variables:

WinModDiff1, AvgModDiff1, AvgModDiff2, RmsNoiseLoudness, RmsNoiseLoudAsym, RmsModDiff, AvgLinDist

5.2.4.2 Loudness threshold

For the Model Output Variables using this criterion, all momentary values calculated until 50 ms after the overall loudness of one of the corresponding audio channels has once reached a value of N_{Thres} sone for both test and Reference Signal are not taken into account in the temporal averaging. The *Loudness Threshold* is used only for the Model Output Variables described in § 4.3.

5.2.4.3 Energy threshold

When the energy of the most recent half of a frame of 2 048 samples is less than 8 000*, in either the mono channel, or both, the left and right channels of the reference and test data, the frame is ignored. Frames have a 50 per cent overlap and only the half of the frame containing new data is evaluated. Application of this criterion prevents the processing of frames with very little energy.

This criterion is used only for the Model Output Variable described in § 4.8.

5.2.4.4 Data boundary

If the processed file contains noise before or after legitimate reference file data, the relative error can be very large since the reference level is zero. When this error is considered an artefact, it may be ignored by applying the data boundary rejection criterion.

When the files are first opened, the locations of the beginning and end of actual data in the reference file are identified. The beginning or end of data is defined as the first location, scanning from the start or end of the file, where the sum of the absolute values over five succeeding samples exceeds 200, in one of the corresponding audio channels. Frames which are fully outside of this range are subsequently ignored.

This criterion is used for the Model Output Variables described in § 4.8 and § 4.4 to 4.6.

---

* This number refers to input data with a 16-bit signed integer format with a range of -32 768 to 32 767 as used on compact disc.

5.3 Averaging over audio channels

When not indicated differently, in the case of stereo signals the MOVs of the left and right channel are averaged linearly after the temporal averaging.

6 Estimation of the perceived basic audio quality

The *perceived basic audio quality* is estimated by mapping several Model Output Variables to a single number using an artificial neural network structure with one hidden layer.

6.1 Artificial neural network

The activation function of the neural network is an asymmetric sigmoid:

$$sig(x) = \frac{1}{1+e^{-x}} \tag{94}$$

The network uses I inputs and J nodes in the hidden layer. The mapping is defined by a set of input scaling factors $a_{min}[i]$, $a_{max}[i]$, a set of input weights $w_x[i]$, a set of output weights $w_y[j]$ and a pair of output scaling factors b_{min} and b_{max}. The inputs are mapped to a *distortion index*

$$DI = w_y[J] + \sum_{j=0}^{J-1}\left(w_y[j] \cdot sig\left(w_x[I,j] + \sum_{i=0}^{I-1} w_x[i,j] \cdot \frac{x[i] - a_{min}[i]}{a_{max}[i] - a_{min}[i]} \right)\right) \tag{95}$$

which is directly related to the estimated *perceived basic audio quality* in terms of an *objective difference grade* (ODG). The relation between *distortion index* and *objective difference grade* is given by:

$$ODG = b_{min} + (b_{max} - b_{min}) \cdot sig(DI) \tag{96}$$

6.2 Basic Version

The Basic Version uses only the FFT-based ear model. It uses the following Model Output Variables: $BandwidthRef_B$, $BandwidthTest_B$, $Total\ NMR_B$, $WinModDiff1_B$, ADB_B, EHS_B, $AvgModDiff1_B$, $AvgModDiff2_B$, $RmsNoiseLoud_B$, $MFPD_B$ and $RelDistFrames_B$. These 11 Model Output Variables are mapped to a single quality index using a neural network as described in 6.1 (*Artificial neural network*) with three nodes in the hidden layer. The parameters of the mapping are given in Tables 13 to 17 below.

TABLE 13

Model Output Variables used in the Basic Version

Model Output Variable (MOV)	purpose
$WinModDiff1_B$	Changes in modulation (related to roughness)
$AvgModDiff1_B$	
$AvgModDiff2_B$	
$RmsNoiseLoud_B$	Loudness of the distortion
$BandwidthRef_B$	Linear distortions (frequency response etc.)
$BandwidthTest_B$	
$RelDistFrames_B$	Frequency of audible distortions
$Total\ NMR_B$	Noise-to-mask ratio
$MFPD_B$	Detection probability
ADB_B	
EHS_B	Harmonic structure of the error

TABLE 14

Scaling factors for the inputs of the Basic Version

index (i)	MOV (x[i])	$a_{min}[i]$	$a_{max}[i]$
0	$BandwidthRef_B$	393.916656	921
1	$BandwidthTest_B$	361.965332	881.131226
2	Total NMR_B	–24.045116	16.212030
3	$WinModDiff1_B$	1.110661	107.137772
4	ADB_B	–0.206623	2.886017
5	EHS_B	0.074318	13.933351
6	$AvgModDiff1_B$	1.113683	63.257874
7	$AvgModDiff2_B$	0.950345	1145.018555
8	$RmsNoiseLoud_B$	0.029985	14.819740
9	$MFPD_B$	0.000101	1
10	$RelDistFrames_B$	0	1

TABLE 15

Weights for the input nodes of the Basic Version

index (i)	MOV (x[i])	node 1 ($w_x[i,0]$)	node 2 ($w_x[i,1]$)	node 3 ($w_x[i,2]$)
0	$BandwidthRef_B$	–0.502657	0.436333	1.219602
1	$BandwidthTest_B$	4.307481	3.246017	1.123743
2	Total NMR_B	4.984241	–2.211189	–0.192096
3	$WinModDiff1_B$	0.051056	–1.762424	4.331315
4	ADB_B	2.321580	1.789971	–0.754560
5	EHS_B	–5.303901	–3.452257	–10.814982
6	$AvgModDiff1_B$	2.730991	–6.111805	1.519223
7	$AvgModDiff2_B$	0.624950	–1.331523	–5.955151
8	$RmsNoiseLoud_B$	3.102889	0.871260	–5.922878
9	$MFPD_B$	–1.051468	–0.939882	–0.142913
10	$RelDistFrames_B$	–1.804679	–0.503610	–0.620456
11	bias	–2.518254	0.654841	–2.207228

TABLE 16

Weights for the output node of the Basic Version

node 1 ($w_y[0]$)	node 2 ($w_y[1]$)	node 3 ($w_y[2]$)	bias ($w_y[3]$)
–3.817048	4.107138	4.629582	– 0.307594

TABLE 17

Scaling factors for the output of the Basic Version

	b_{min}	b_{max}
ODG	–3.98	0.22

6.3 Advanced Version

The *Advanced Version* uses both the filter bank-based ear model and the FFT-based ear model. It uses the Model Output Variables $RmsModDiff_A$, $RmsNoiseLoudAsym_A$, $AvgLinDist_A$, *Segmental* NMR_B and EHS_B. These 5 Model Output Variables are mapped to a single quality index using a neural network as described in 6.1 (*Artificial neural network*) with five nodes in the hidden layer. The parameters of the mapping are given in Tables 18 to 22 below.

TABLE 18

Model Output Variables used in the Advanced Version

Model Output Variable (MOV)	Purpose
$RmsNoiseLoudAsym_A$	Loudness of the distortion
$RmsModDiff_A$	Changes in modulation (related to roughness)
$AvgLinDist_A$	Linear distortions (frequency response etc.)
Segmental NMR_B	Noise-to-mask ratio
EHS_B	Harmonic structure of the error

TABLE 19

Scaling factors for the input nodes of the Advanced Version

index (i)	MOV (x[i])	$a_{min}[i]$	$a_{max}[i]$
0	$RmsModDiff_A$	13.299	2166.500
1	$RmsNoiseLoudAsym_A$	0.041	13.243
2	$AvgLinDist_A$	0.025	14.225
3	Segmental NMR_B	–25.019	13.467
4	EHS_B	0.062	10.227

TABLE 20

Weights for the inputs of the Advanced Version

index (i)	MOV (x[i])	node 1 (w_x[i,0])	node 2 (w_x[i,1])	node 3 (w_x[i,2])	node 4 (w_x[i,3])	node 5 (w_4[i,4])
0	$RmsModDiff_A$	21.212	–39.913	–1.383	–14.545	–0.321
1	$RmsNoiseLoudAsym_A$	–8.982	19.956	0.935	–1.687	–3.239
2	Segmental NMR_B	1.634	–2.878	–7.443	5.607	–1.783
3	EHS_B	6.104	19.587	–0.240	1.088	–0.511
4	$AvgLinDist_A$	11.556	3.892	9.720	–3.287	–11.031
2	bias	1.331	2.686	2.097	–1.328	3.087

TABLE 21

Weights for the output node of the Advanced Version

node 1 (w_x[i,0])	node 2 (w_x[i,1])	node 3 (w_x[i,2])	node 4 (w_x[i,3])	node 5 (w_4[i,4])	bias (w_y[4])
–4.697	–3.290	7.005	6.652	4.009	–1.360

TABLE 22

Scaling factors for the output of the Advanced Version

	b_{min}	b_{max}
ODG	–3.98	0.22

7 Conformance of implementations

7.1 General

This section presents a set of test items to verify the proper implementation of the method.

7.2 Selection

The test items were selected from Database 3 (DB3), which was used for the validation of the models. To simplify testing, a subset of the 84 items of DB3 was selected. This subset consists of 21 items. The major criterion for the selection was that the resulting MOVs and DI *(Distortion Index)* values cover a broad range.

7.3 Settings for the conformance test

The test items are available from the ITU as WAV-files (Microsoft RIFF format). All items were sampled at 48 kHz, 16 bit PCM. The reference and test signals as provided by the ITU are already time and level adapted to each other, so that no additional gain or delay compensation is required. The measurement algorithm must be adjusted to a listening level of 92 dB SPL.

7.4 Acceptable tolerance interval

In order to conform to the recommendation, the calculated DI values must reproduce the values given in Tables 23 and 24, with a tolerance of less than ±0.02 for all test items[7]. If an implementation does not produce results within this tolerance it does not conform to this Recommendation.

7.5 Test items

The following tables show the names of the reference and test items[8], and the resulting DI values. Table 23 is related to the Basic Version, and Table 24 contains the values for the Advanced Version.

TABLE 23

Test items and resulting DI values for the Basic Version

Item	DI	ODG
acodsna.wav	1.304	–0.676
bcodtri.wav	1.949	–0.304
ccodsax.wav	0.016	–1.863
dcodryc.wav	1.648	–0.458
ecodsmg.wav	1.731	–0.412
fcodsb1.wav	0.677	–1.195
fcodtr1.wav	1.419	–0.598

Item	DI	ODG
fcodtr2.wav	–0.045	–1.927
fcodtr3.wav	–0.715	–2.601
gcodcla.wav	1.781	–0.386
hcodryc.wav	2.291	–0.166
hcodstr.wav	2.403	–0.128
icodsna.wav	–3.029	–3.786
kcodsme.wav	3.093	0.038

Item	DI	ODG
lcodhrp.wav	1.041	–0.876
lcodpip.wav	1.973	–0.293
mcodcla.wav	–0.436	–2.331
ncodsfe.wav	3.135	0.045
scodclv.wav	1.689	–0.435

TABLE 24

Test items and resulting DI values for the Advanced Version

Item	DI	ODG
acodsna.wav	2.392	–0.132
bcodtri.wav	1.830	–0.361
ccodsax.wav	1.654	–0.455
dcodryc.wav	1.764	–0.394
ecodsmg.wav	1.490	–0.552
fcodsb1.wav	1.918	–0.318
fcodtr1.wav	1.333	–0.657
fcodtr2.wav	0.333	–1.533

Item	DI	ODG
fcodtr3.wav	–0.501	–2.395
gcodcla.wav	2.027	–0.269
hcodryc.wav	1.826	–0.363
hcodstr.wav	1.990	–0.285
icodsna.wav	–3.245	–3.823
kcodsme.wav	1.972	–0.293
lcodhrp.wav	1.337	–0.654
lcodpip.wav	2.093	–0.241

Item	DI	ODG
mcodcla.wav	1.364	–0.635
ncodsfe.wav	1.921	–0.316
scodclv.wav	1.893	–0.330

[7] To achieve this accuracy IEEE floating point arithmetic should be used.

[8] The names of the corresponding reference items are derived by replacing the substring “cod” in the names of the test items by “ref”, e.g. the reference item for “bcodtri.wav” is “breftri.wav”.

APPENDIX 1

(TO ANNEX 2)

Validation process

1 General

In 1994 ITU-R adopted Question ITU-R 210/10 “Objective Perceptual Quality Assessment Methods”and a Task Group was initiated. One of the first actions was to issue an open call for proposals, and responses from six model proponents were received.

A lot of efforts were spent to define the procedures for the validation process. It was found useful to compile a first database, referred to as DB1, consisting of material from listening tests which had already been performed. The main focus was on medium and high audio quality and consequently only results from listening tests in conformance with Recommendation ITU-R BS.1116 were considered. The material from these tests represented critical broadcast material for low bit rate codecs such as MPEG1 Layer II, MPEG1 Layer III, Dolby AC2, Mini Disc, NICAM and others. Database 1 was created to provide the model proponents with a common platform consisting of material that covered a large range of impairments, a variety of codecs and degradation from cascaded codecs. A detailed description of the tests compiled in Database 1 can be found in Appendix 2 to Annex 2.

Obviously, an objective measurement method of perceived audio quality that imitates human behaviour can only be validated on a database containing results from subjective tests. An appropriate validation requires a database which is based on unknown material. For this reason it was necessary to carry out new listening tests. As the measurement method should ideally target any type of artefact which could appear in broadcasting applications, not only coding artefacts should be included, but also more traditional artefacts like distortion and noise. Database 2 and Database 3 were created in 1996 and 1997 respectively to meet these requirements. In addition to the codecs included already in DB1 Dolby AC-3 and AAC were also included. Further details are found in Appendix 2 to Annex 2.

The validation should take into account the uncertainties, often presented as confidence intervals, inherent in subjective listening tests. The size of the confidence interval depends on a number of factors. The most important are the experience of the subjects, training procedures and the context in which the test items were presented, as well as the number of subjects.

The adaptation and the validation of the objective method given in this Recommendation is based on an “average expert listener”. The mean values from the subjective quality assessments together with the 95% confidence intervals are used to characterize the “average expert listener”.

Subjective listening tests are very sensitive to various factors that influence the results. The SDGs for both Database 2 and Database 3 were produced at three different test sites and a number of studies investigated whether the data really could be combined. Although not all of the studies came to identical conclusions it was found reasonable to merge the data and this merged database formed the basis for the validation.

The validation process was split into three phases:

– Phase 1: Competitive phase

– Phase 2: Collaborative phase

– Phase 3: Final selection

These phases will be described in detail in the following chapters.

2 Competitive phase

Six methods (DIX, NMR, PAQM, PERCEVAL, POM, TTA) were proposed for objective measurement of perceived audio quality and it was decided to compare the performance of these using Database 2 and a subset of Database 1. Database 2 was generated in the beginning of 1996. The selection of the final test material was a joint effort between SR (Sweden) and the BBC (United Kingdom). The listening tests were carried out at NRK in Norway, DR in Denmark and

at NHK in Japan. A statistical analysis of the data from the tests was prepared by Deutsche Telekom (Germany) and Teracom (Sweden). During Phase 1 the objective data were generated at a neutral site (Swisscom, Switzerland). The model proponents then received the first half of Database 2 for a final adaptation of the methods (Phase 2). Finally, new Objective Difference Grades were generated at Swisscom.

The analysis of the performance of the methods was carried out by Teracom (Sweden) as well as by the proponents themselves. Though the results of some of the proposed methods showed high correlation with the SDGs, there was a consensus that none of the methods fulfilled the requirements of the users. A separate study showed that none of the proposed methods was significantly better than the others. It was therefore decided to develop an improved measurement method as a joint effort between all the current proponents. The performance of the new method should be compared with one of the already established methods referred to as model B3.

3 Collaborative phase

The collaborative phase was based on the idea to combine the best elements of the different methods into one new method. To best fit the users' needs, it was decided to develop two versions of the method. One which is suitable for real-time implementations, and one that may require higher computational power to achieve higher accuracy.

The validation procedure for the new methods was designed in a similar way as the one for the competitive phase. A new database (DB3) had to be created. The items and conditions were finally defined in spring 1997 and compiled at SR, Swisscom and BBC. A full description of the database can be found in Appendix 2 to Annex 2. The subjective listening test were performed at three test sites, Deutsche Telekom, NHK and SR. All sites applied the "Triple stimulus hidden reference double blind method", described in Recommendation ITU-R BS.1116. The results from the listening tests were collected in Sweden. An extensive statistical analysis of the listening test results was performed at Teracom as well as by other parties. Due to this analysis some of the listeners were excluded from the further evaluation. The results from the test sites were combined to form the Database 3.

In Autumn of 1997 52 items out of the database were released to the proponents. The new methods were adapted to the new data. As there were several parameter settings that delivered similar results the decision for the final selection was delayed as long as possible. Finally in Switzerland the remaining 32 items were used to validate the new methods on an "unknown" data set.

In addition the results of a new listening test, carried out by CRC (Canada) were used to validate the new methods on "unknown" material. The selection and the verification process are described in the following sections.

4 Verification

Extensive tests of the 18 specified versions of the objective measurement method were performed. In this section the selection criteria are described, and the results from comparing the SDGs with the results obtained from 18 versions of the measurement method. The objective was to select and verify the optimal versions which will be recommended to the ITU.

Selection criteria

The correlation between subjective and objective results is the most obvious criterion to validate an objective method. In addition two further criteria that consider the reliability of the mean values were introduced for the validation – the Absolute Error Score (AES) and the Tolerance Scheme.

The Absolute Error Score (AES), introduced to relate the accuracy of the model to the accuracy of the listening test, is defined according to the following expression:

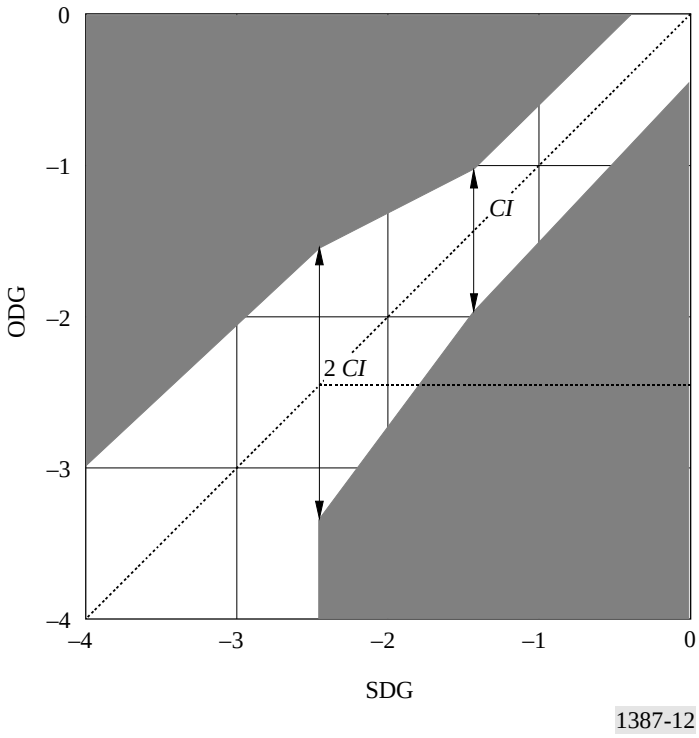
$$AES = 2 * \sqrt{\frac{\sum((ODG - SDG) / CI)^2}{N}}$$

if CI < 0.25 then CI = 0.25 where

CI is the confidence interval

The Tolerance Scheme was designed to allow different deviations of the ODGs from the SDGs at the upper and lower end of the impairment scale. The tolerated range is related to the confidence intervals of the listening tests. This range is limited to a minimum value of 0.25 grades. The distance of ODGs outside the Tolerance Scheme to the Tolerance Scheme were used to assess the quality of the measurement method.

FIGURE 12
Tolerance scheme, confidence interval $CI \geq 0.25$



4.1 Comparison of SDG and ODG values

The objective measurements were split into three different phases. During Phase 1 all of the 84 test items were unknown to everybody except for the selection panel. During Phase 2 information on 52 items were released. The information contained both the SDG values and the actual audio excerpts. In Phase 3 this knowledge was used to optimize the performance of the versions of the method. Please note that four additional versions were tested during Phase 3, compared to Phase 1. The presented SDG values were calculated from data generated by 75 qualified subjects.

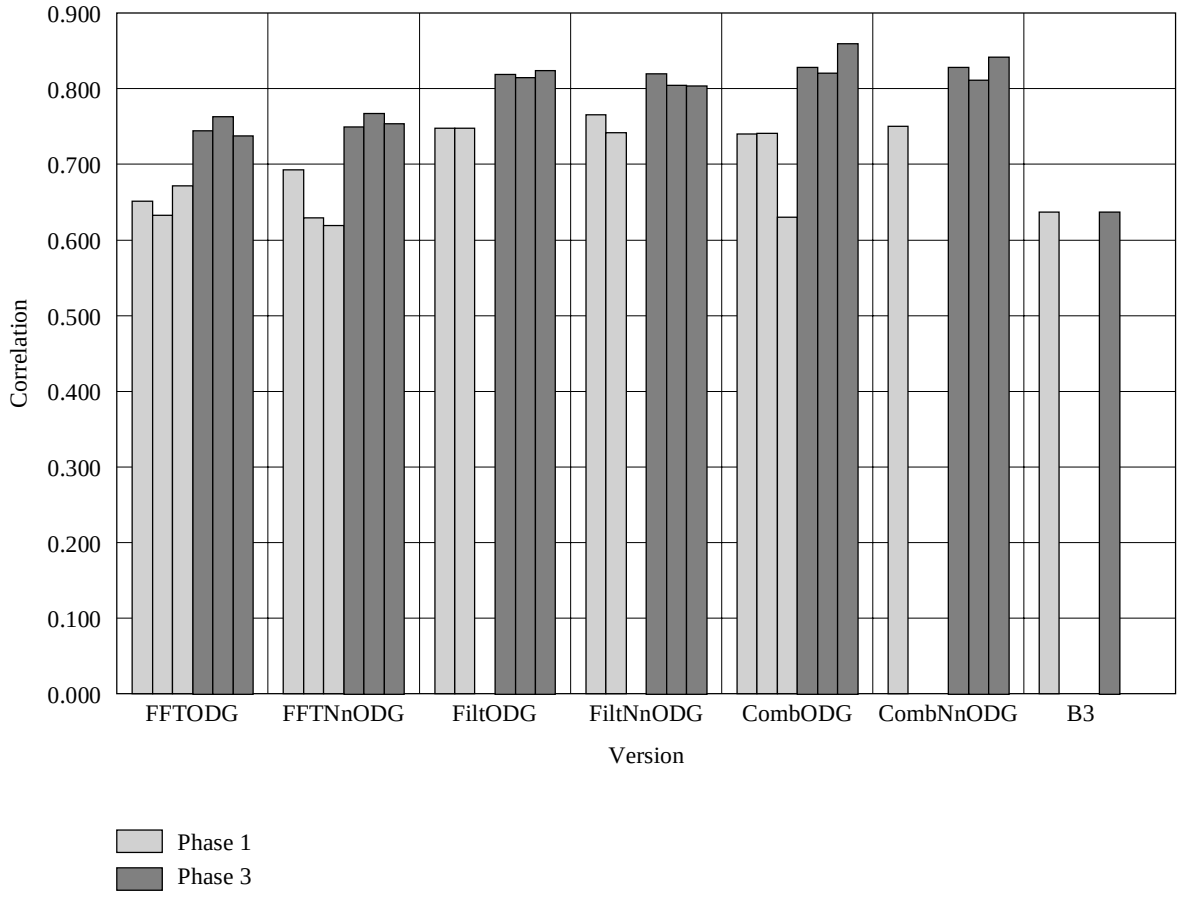
There are many different ways to evaluate how well the ODGs reflect the SDGs. Unfortunately no single value really illustrates the complete performance. Instead one has to look from a number of perspectives. The correlations are presented in § 4.2 and the Absolute Error Scores (AES) are presented in § 4.3. Model B3 is one of the models which were tested by ITU-R in 1996 and it had been decided that the various new versions should be compared with this older one.

4.2 Correlation

The correlation figures from both Phase 1 and Phase 3 have been plotted in Figure 13 (84 items) and Figure 14 (32 items).

FIGURE 13

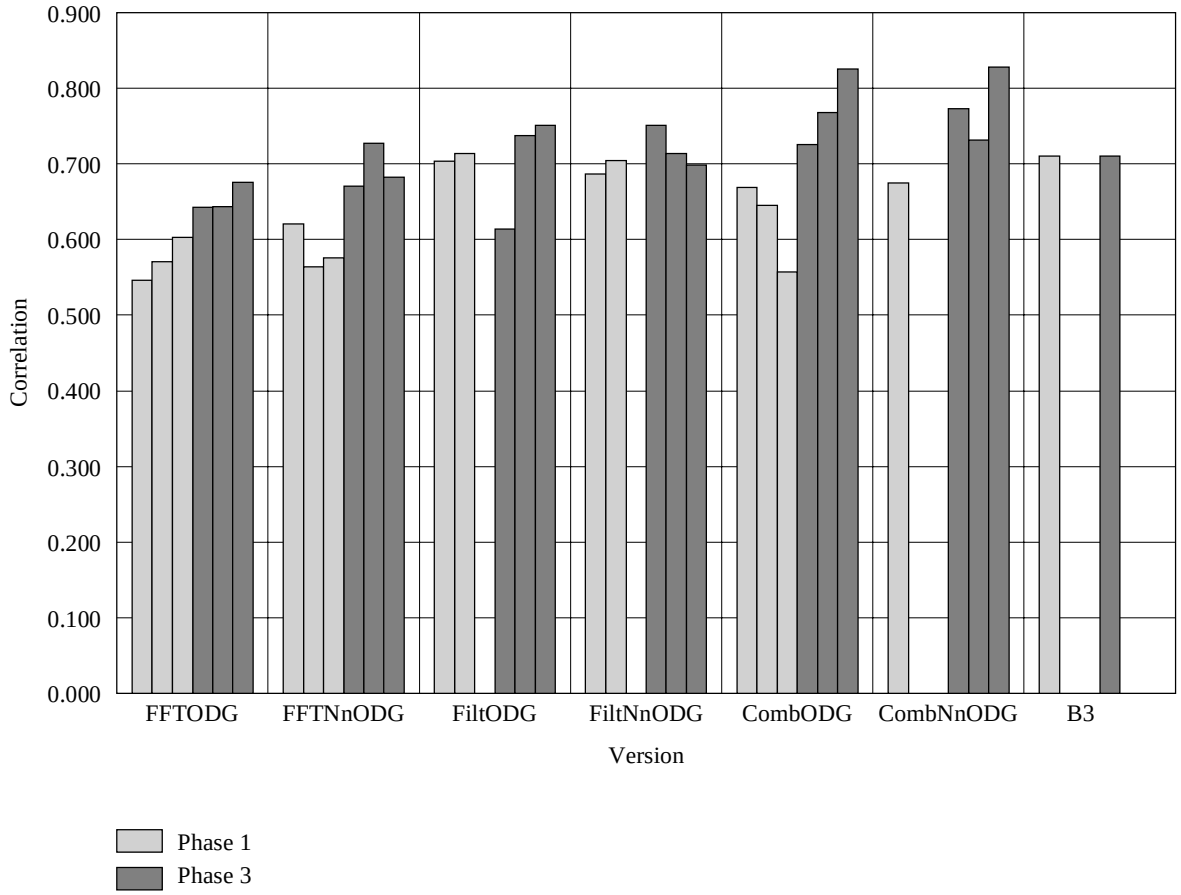
Correlation between SDG and ODG. All 84 items have been included



1387-13

FIGURE 14

Correlation between SDG and ODG. The 32 unreleased items have been included



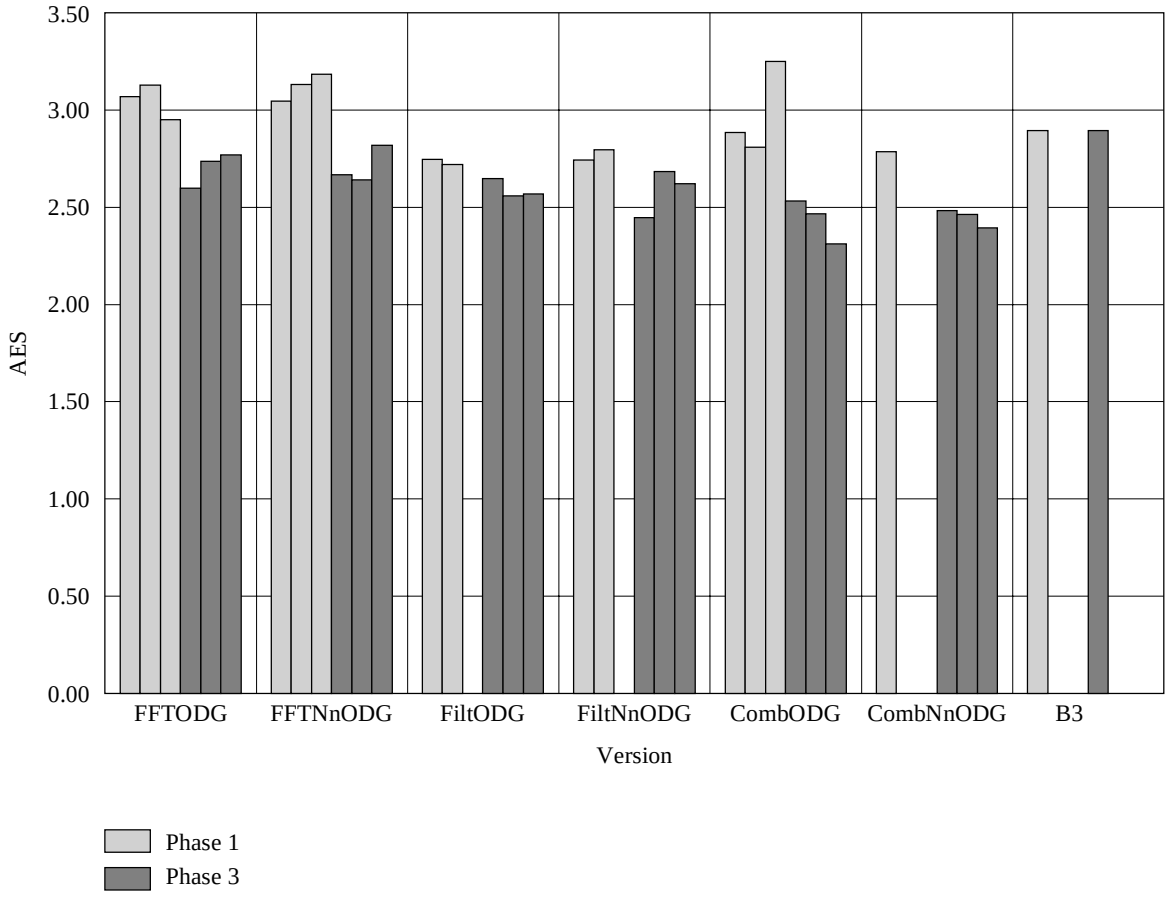
1387-14

4.3 Absolute Error Score (AES)

A model which on average produces ODG values within the SDG confidence interval will obtain an AES value close to 2. An overview of the AES values are given in Figure 15 to Figure 16.

FIGURE 15

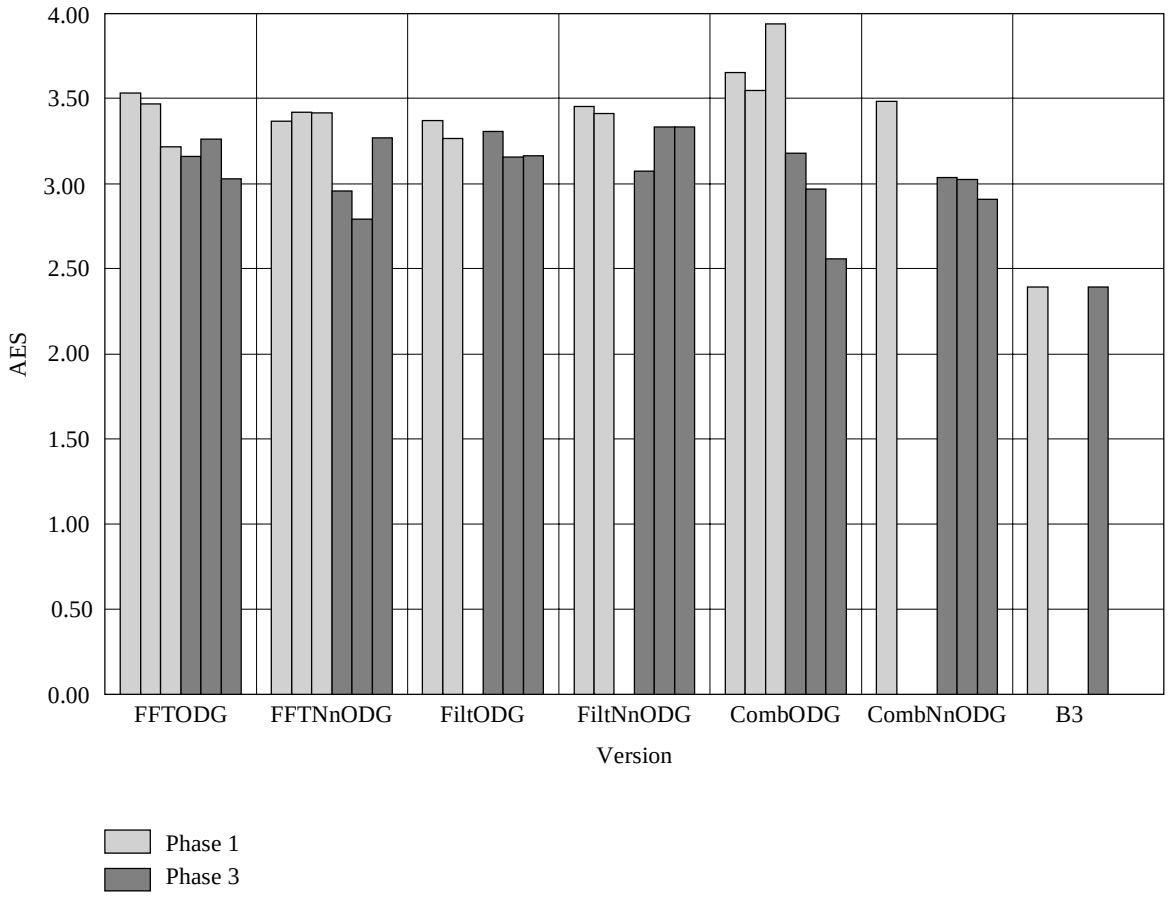
AES for the different versions. All 84 items have been included



1387-15

FIGURE 16

AES for the different versions. The 32 unreleased items have been included



1387-16

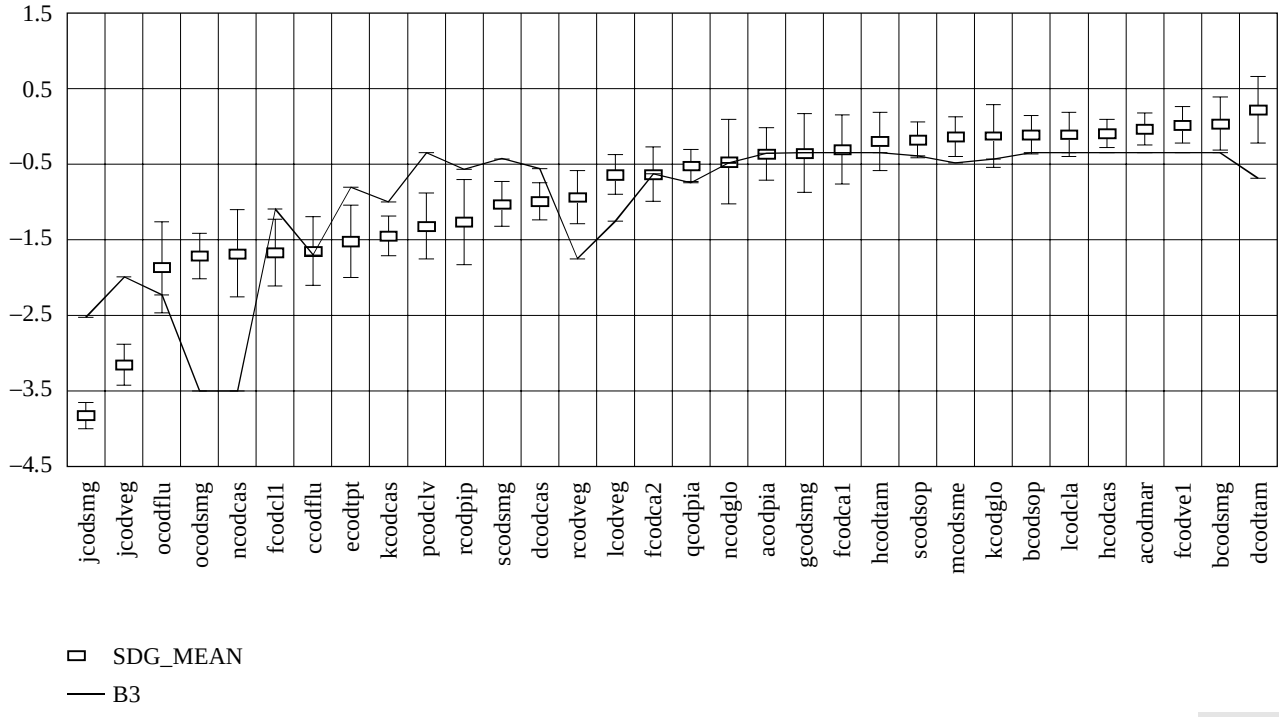
4.4 Comparison of ODG versus the confidence interval

Most versions performed similarly, as probably could be expected. A lot of plots were presented at the meeting but in this Recommendation the repertoire is limited. For much more details please refer to the full verification test report.

In Figure 17 through to Figure 22 the average SDG, confidence interval and ODG for the 32 unreleased items have been plotted for Model B3 and the model versions FftNnODG1 and CombNnODG3.

FIGURE 17

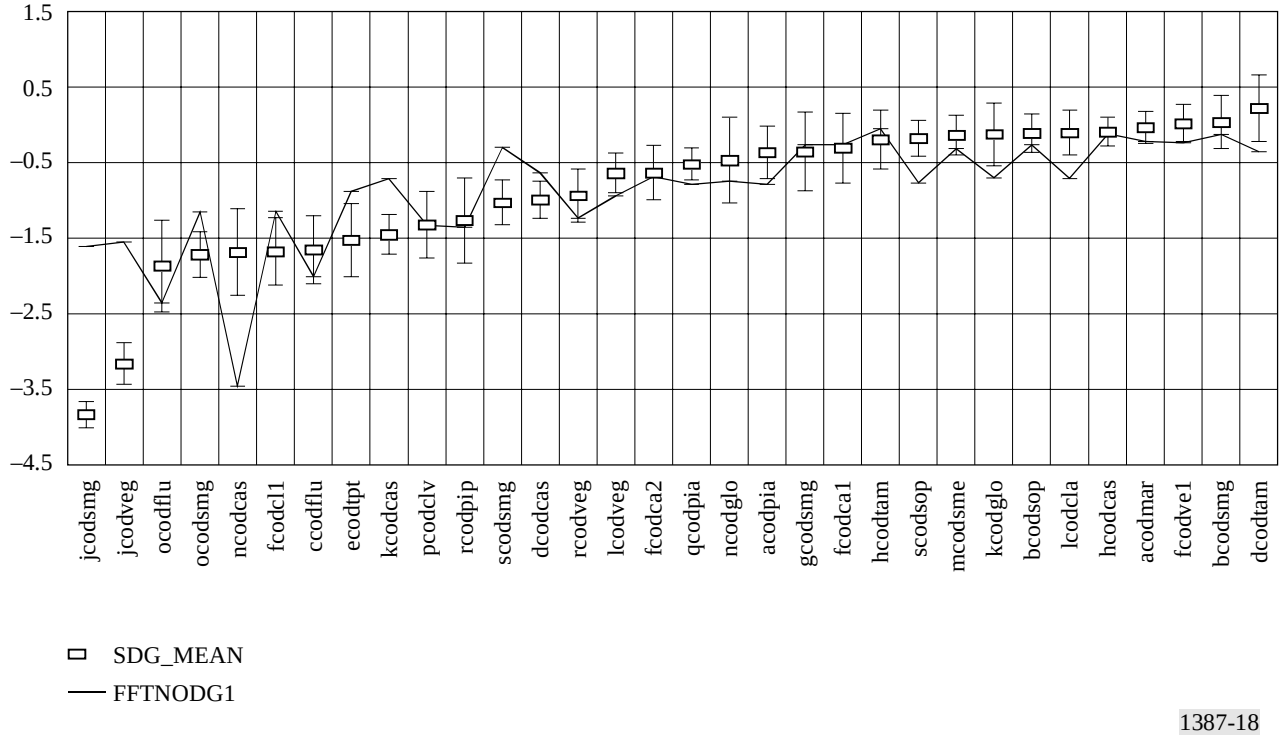
Model B3: plot of average SDG, confidence interval and ODG for the 32 unreleased items



1387-17

FIGURE 18

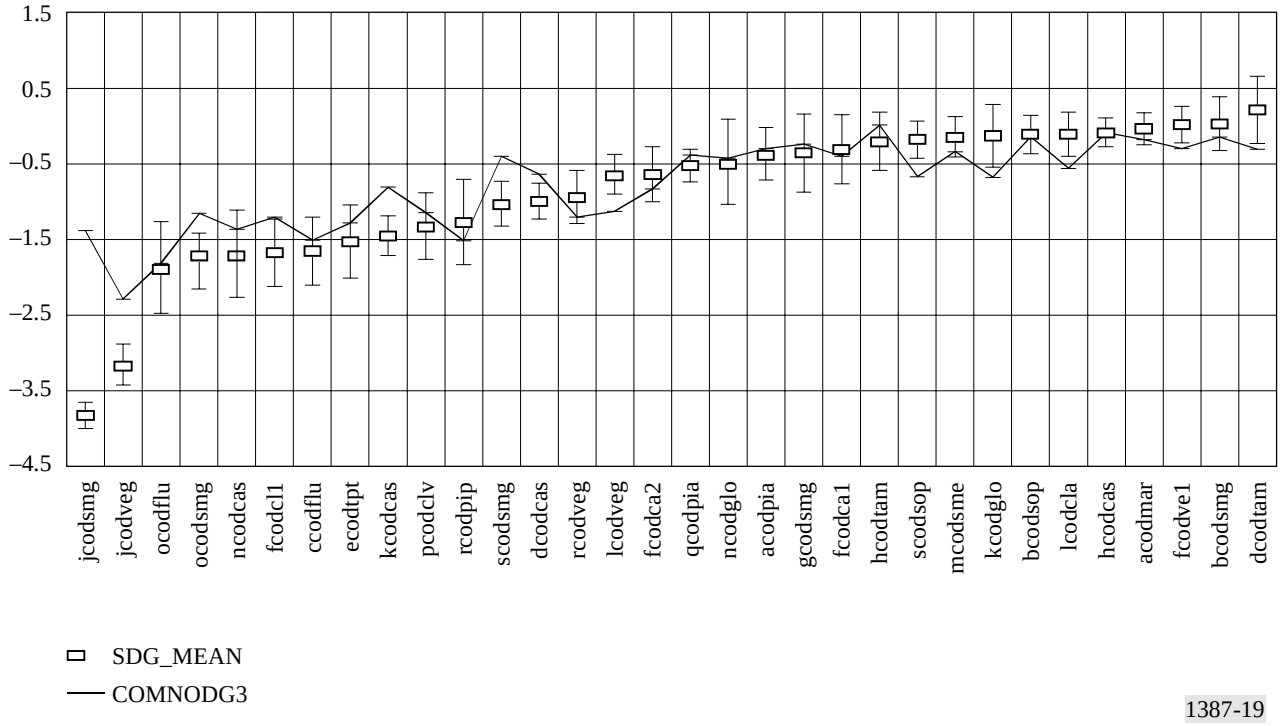
FFTNnODG1: plot of average SDG, confidence interval and ODG after the third phase for the 32 unreleased items



1387-18

FIGURE 19

CombNnODG3: plot of average SDG, confidence interval and ODG after the third phase for the 32 unreleased items



Similar plots, but in this case for all 84 items during Phase 3 are given in Figure 20 and Figure 21. In addition, Figure 22 illustrates the performance of the version CombNnODG3.

FIGURE 20

Model B3: plot of average SDG, confidence interval and ODG for all 84 items

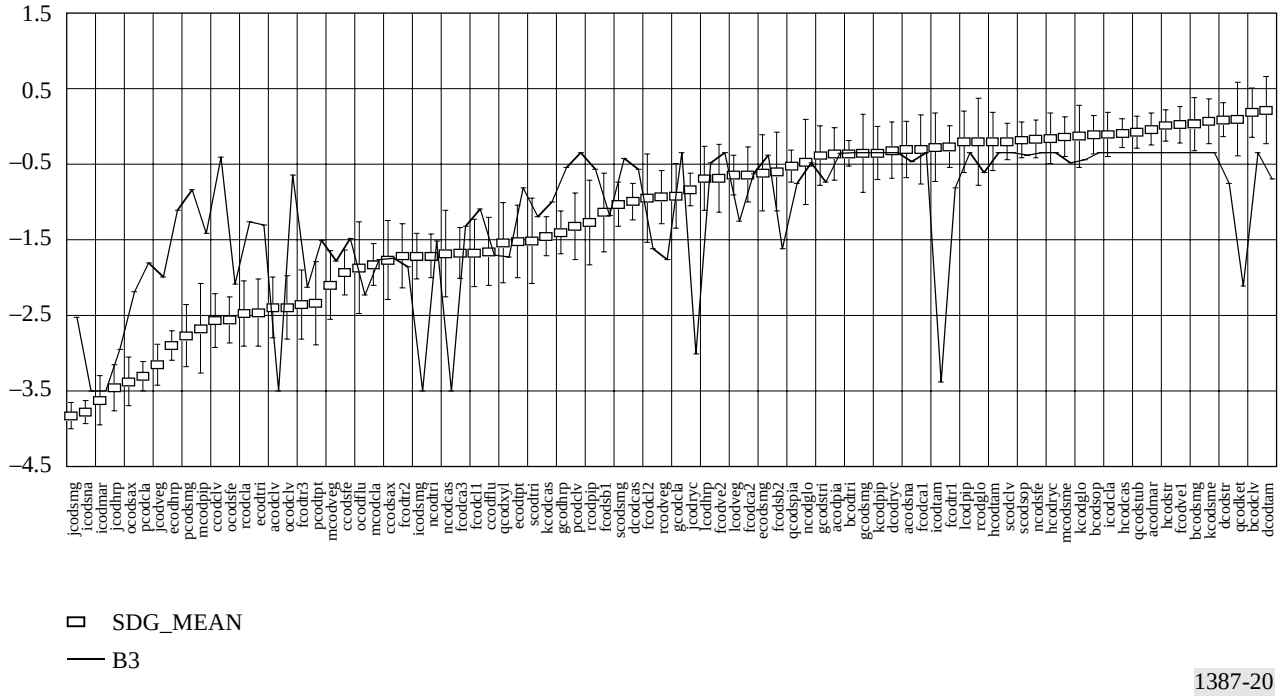


FIGURE 21

FFTNnODG1: plot of average SDG, confidence interval and ODG during Phase 3 for all 84 items

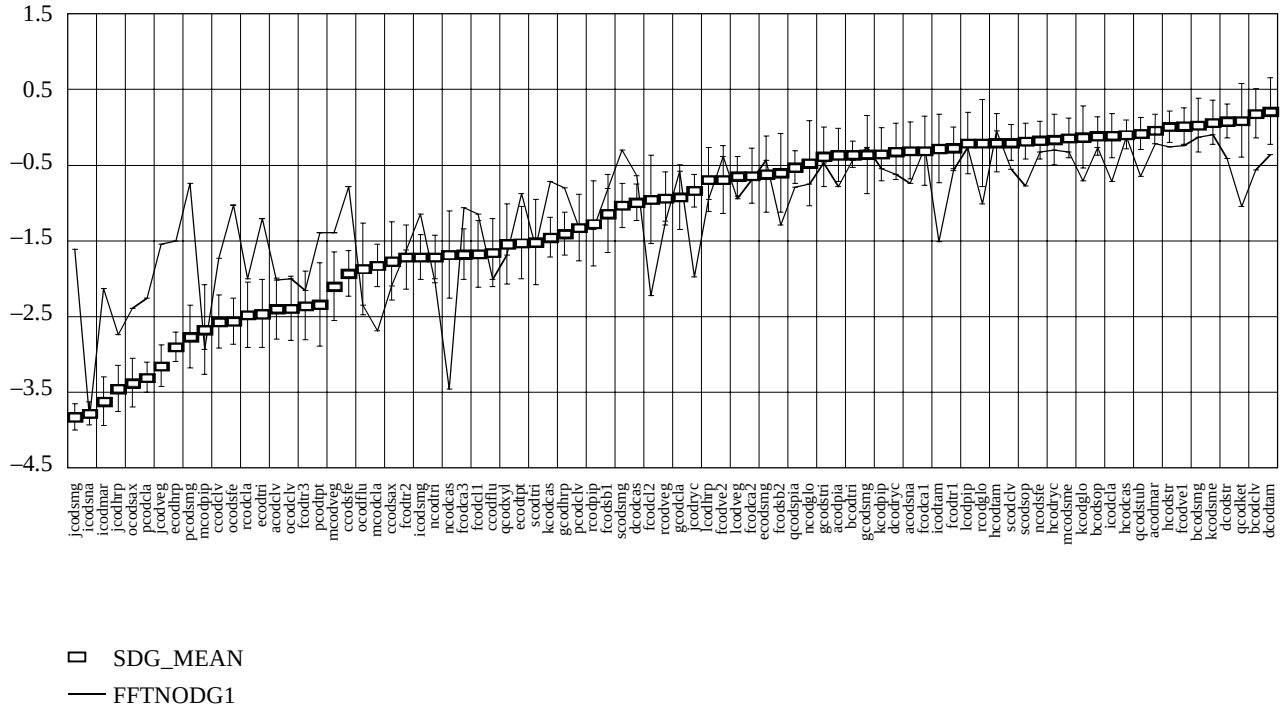
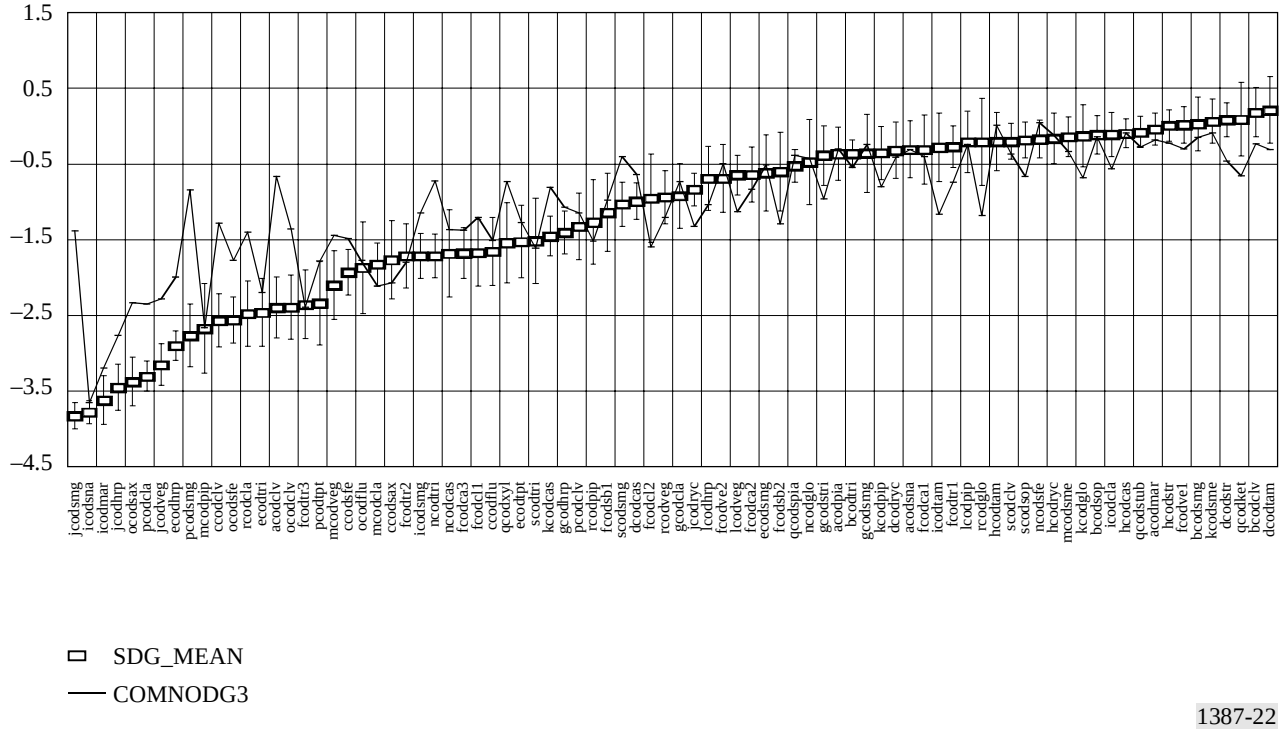


FIGURE 22

CombNnODG3: plot of average SDG, confidence interval and ODG during Phase 3 for all 84 items



4.5 Comparison of ODG versus the tolerance-interval

ITU-R has defined a target user requirement which can be mapped to a tolerance interval. The target requirements are more stringent for higher levels of audio quality and more relaxed for lower audio quality. The figures below illustrate the performance in this dimension of model B3 and the model versions FftNnODG1 and CombNnODG3 for all the 84 items during Phase 3.

FIGURE 23

B3: plot of average SDG, tolerance interval and ODG during Phase 3 for all 84 items

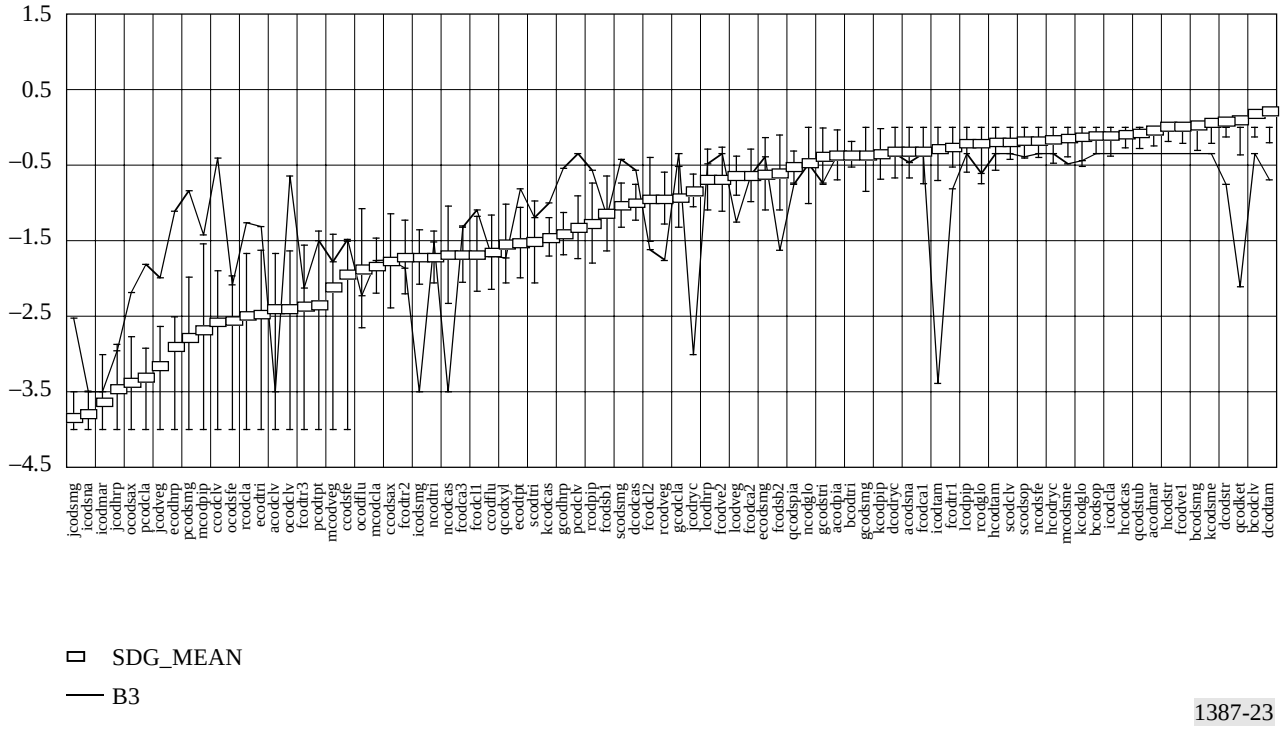
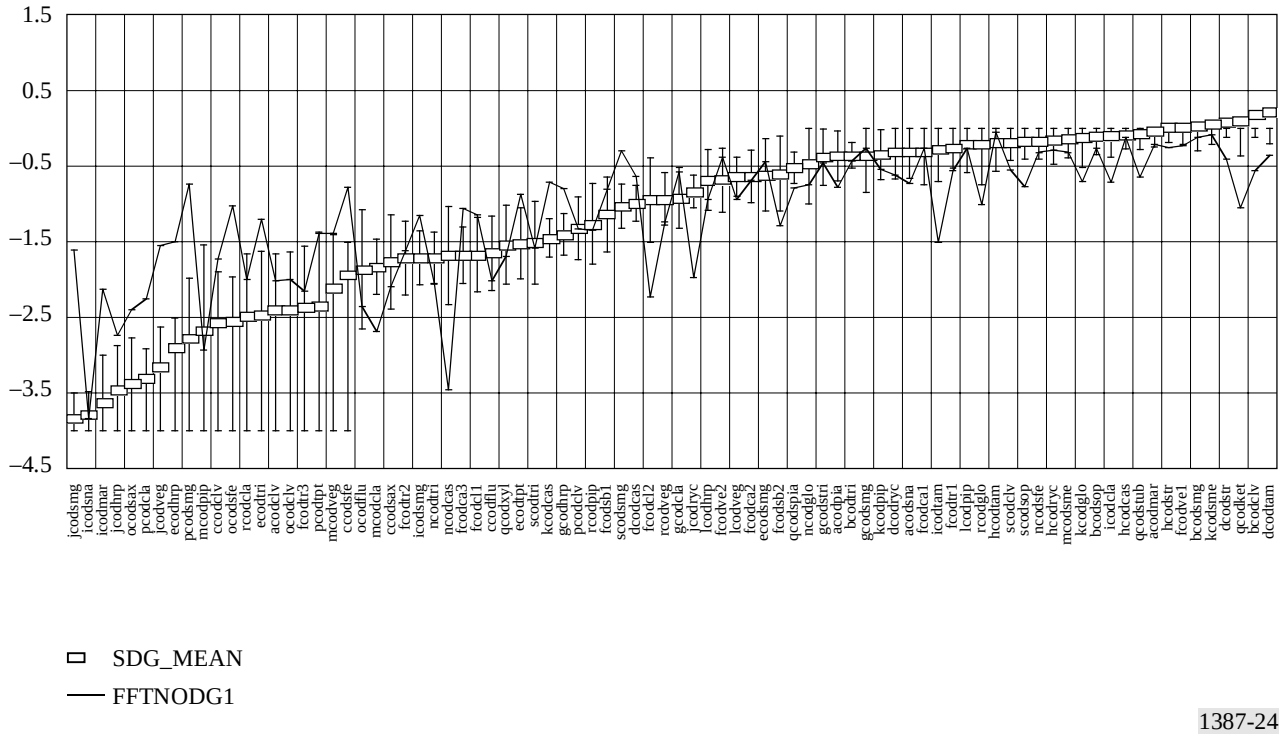


FIGURE 24

FFTNnODG1: plot of average SDG, tolerance interval and ODG during Phase 3 for all 84 items



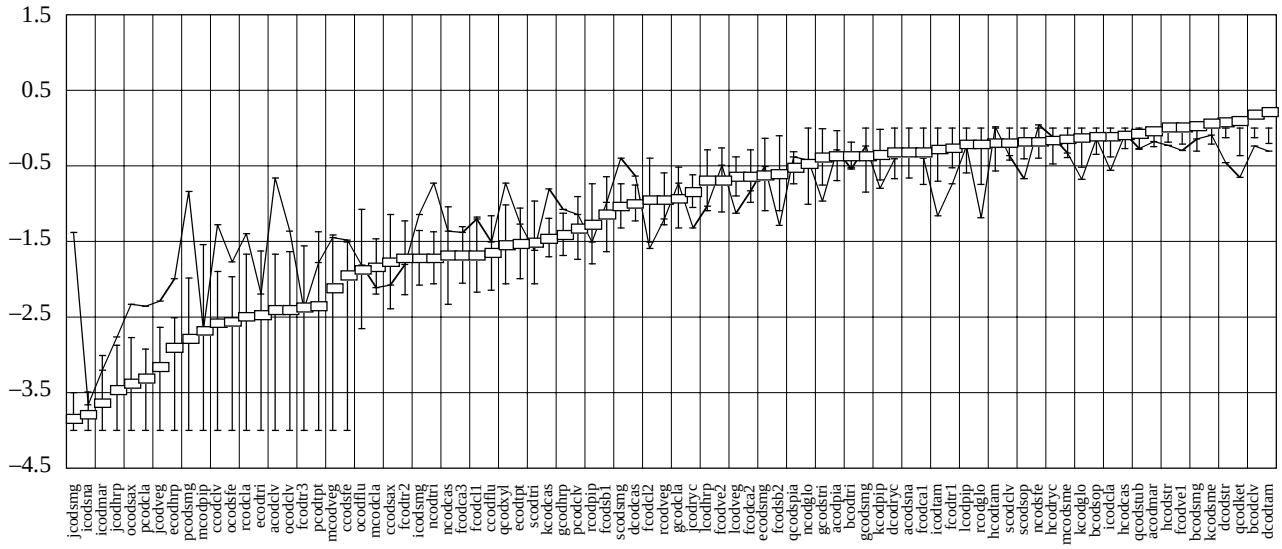
SDG_MEAN

FFTNODG1

1387-24

FIGURE 25

CombNnODG3: plot of average SDG, tolerance interval and ODG during Phase 3 for all 84 items



SDG_MEAN

COMNODG3

1387-25

5 Selection of the optimal model versions

Eighteen different model versions were submitted to the objective test site: 6 versions are based on the FFT, 6 versions use a perceptual adapted filter bank and 6 versions use a combination of both FFT and filter bank. The FFT group is targeted to be used in a measurement scheme capable of a real-time implementation, the remaining two groups need a higher computational complexity and are expected to achieve higher accuracy. The six model versions in each group can be divided into two subgroups: One using neural networks and one **not** using neural networks. The performance of the reference model from April 1996 (B3) on Database 3 is also included for comparison.

5.1 Pre-selection criteria based on correlation

– The performance of the 18 different model versions and reference model B3 were evaluated using two data sets which were mostly not used for the training of models (Phase 1 and Phase 3 second part and CRC have not been used for training).

– Database 3 was created especially for the evaluation of the perceptual models. About half of this database was used for the training of models. The correlation between the subjective results and the objective results on the remaining items (DB3_2nd) was used for the assessment of the objective models.

– The CRC database was created by the CRC for evaluation of perceptual audio coding schemes. This database was not used for training of the perceptual models. The correlation between the subjective results and the objective results for all items was used for the assessment of the objective models.

The results of Database 3 are given in Tables 25 and 26 (correlations and Absolute Error Scores).

In the pre-selection phase all inputs quantifying the performance of all model versions were taken into account with a high weight on DB3. Based on the overall comparison it was decided to make a further comparison of the two model versions within each group which appeared to be the best. Table 25 shows the correlation of these six (three times 2 model versions).

TABLE 25

Correlation between SDG and ODG

	FFTNnODG1	FFTNnODG2	FiltODG2	FiltODG3	CombNnODG3	CombODG3	B3
DB3_2nd	0.671	0.728	0.738	0.751	0.828	0.826	0.710
CRC	0.837	0.779	0.862	0.839	0.851	0.777	0.656

TABLE 26

Absolute Error Score

	FFTNnODG1	FFTNnODG2	FiltODG2	FiltODG3	CombNnODG3	CombODG3	B3
DB3_2nd	2.96	2.79	3.16	3.16	2.91	2.56	2.39
CRC	1.55	1.85	1.61	1.67	1.61	1.90	2.78

The two FFT-based model versions show different performance on the two databases. A decision to select one of these two needs further criteria. Taking into account that the CRC database was completely unknown, FFTNnODG1 seems to perform slightly better.

The combined model versions give better results on both databases in comparison to the FFT-based versions. However, they show worse performance for the CRC database in comparison to the filter bank versions. The filter bank versions can be regarded as a special case of the combined model versions where the weighting of the FFT-based version output variables is zero. Therefore the combined model versions are preferred.

5.2 Analysis of number of outliers

The accuracy of the subjective data coming from the listening tests is expressed via the 95% interval around the average of several listeners. The performance of model versions was evaluated also by looking at outliers. An item is considered as an outlier when the difference between the subjective data and the objective data is larger than two times the confidence interval.

Table 27 shows the number of outliers for the six model versions for DB3. Sensitive means that a model version indicates a lower audio quality than the subjective assessment and insensitive is the opposite.

TABLE 27

Outliers

	FFTNnODG1	FFTNnODG2	FiltODG2	FiltODG3	CombNnODG3	CombODG3
Sensitive	10	4	4	4	3	5
Insensitive	13	13	11	13	12	14
Total	23	17	15	17	15	19

Looking at the number of outliers, model version FFTNnODG2 seems to perform better than model version FFTNnODG1. Model version CombNnODG3 shows the best performance among the more accurate versions.

5.3 Analysis of severeness of outliers

TABLE 28

Items with deviations of more than 1.0 difference grades between prediction and SDG

Number of outliers	14	4	4	12	12	9	6
Model version	B3	CombODG3	CombNnODG3	FftNnODG1	FftNnODG2	FiltODG2	FiltODG3
item	jcodsmg	jcodsmg	jcodsmg	jcodsmg	jcodsmg	jcodsmg	jcodsmg
	qcodket	qcodket	pcodsmg	qcodket	qcodket	rcodpip	rcodpip
	pcodsmg	pcodsmg	ccodclv	pcodsmg	pcodsmg	qcodket	qcodket
	pcodcla	icodmar	acodclv	ocodsfe	pcodcla	pcodsmg	pcodsmg
	ocodclv			ncodcas	ocodsfe	ocodsax	ocodsax
	ncodcas			mcodcla	ncodglo	kcodpip	fcodsb2
	mcodpip			jcodveg	kcodcas	jcodveg	
	jcodryc			jcodryc	jcodveg	fcodsb2	
	icodtam			icodtam	jcodryc	fcodcl2	
	icodsmg			icodmar	icodtam		
	fcodsb2			fcodcl2	ecodhrp		
	ecodhrp			ecodhrp	ccodclv		
	ccodclv						
	acodclv						

TABLE 29

Items with deviations of more than 1.5 difference grades between prediction and SDG

number of outliers	8	2	2	3	2	2	2
model version	B3	CombODG3	CombNnODG3	FftNnODG1	FftNnODG2	FiltODG2	FiltODG3
item		jcodsmg	jcodsmg	jcodsmg	jcodsmg	jcodsmg	jcodsmg
	qcodket		pcodsmg				
	pcodsmg	pcodsmg		pcodsmg	pcodsmg		
						pcodsmg	pcodsmg
				ncodcas			
	ncodcas						
	jcodryc						
	icodtam						
	icodsmg						
	ccodclv						
	acodclv						

6 Conclusion

Following the decisions made during the development, two model versions out of the proposed 18 versions have been selected in order to fulfil the requirements for the defined applications of objective measurement methods. A low-complexity version, designed to allow for a cost-efficient real-time implementation, and a higher accuracy version, not necessarily operating in real-time were defined. For the selection process, the above listed criteria were applied and analysed.

As the real-time version an FFT-based model, referred to as “FFTNnODG1” has been selected for the following reasons:

None of the FFT-based versions showed a significant advantage over the others concerning one of the above listed criteria. Regarding the correlation coefficients between ODGs and SDGs, the number and severeness of outliers of each of the verified versions had its pros and cons. Nevertheless, it should be noted that “FFTNnODG1” achieved the best correlation on the CRC’97 database which was completely unknown (0.837).

For the higher accuracy version, preference was given to the combined version, incorporating an FFT and a filter bank, because such an approach also incorporates the subset of a pure filter bank model and thus should have a better performance potential. Altogether six different versions of a combined model were available for the selection process. The selected version “CombNnODG3”, showed less outliers as well as a higher correlation (r=0.851 for CRC’97) than the other versions. This version’s correlation for the complete Database 3 had the same order of magnitude as the one of the version “CombODG3”, but showed a higher correlation compared to the other versions.

APPENDIX 2

(TO ANNEX 2)

Descriptions of the reference databases

1 Introduction

During the development of the method for objective measurement of perceived audio quality, a number of databases were used for the training and validation.

Some of the databases listed contain both headphone and loudspeaker data, and some have headphone data only. For databases with separate sets of data available for loudspeaker and headphone presentation, only headphone data were used.

An item is defined as an audio fragment used in the subjective assessment. A condition refers to a single degradation condition. All the items were used for all experimental conditions except in the DB2 and DB3 studies. DB3 was partly used for training and partly for validation (52 of the 84 items were used for training in the second phase of the validation).

Training

- MPEG90
  - The mean SDG per item quite uniformly covered the range from 0.0 to –4.0.
  - ISO/IEC/JTC 1/SC 2/WG11 MPEG/Audio test report, Document MPEG90/N0030, October 1990.
- MPEG91
  - At least 88 per cent of the mean SDG per item were above –2.0, and the range was 0.1 to –3.8.
  - ISO/IEC/JTC 1/SC 2/WG 11 MPEG/Audio test report, Document MPEG91/N0010, June 1991.
- ITU92DI
  - 80% per cent of the mean SDG per item were above –2.0, and the range was 0.1 to –3.4.
  - See also Rec. ITU-R BS.1115.
- ITU92CO
  - At least 96 per cent of the mean SDG per item were above –2.0, and the range was 0.2 to –2.4.
  - See also Rec. ITU-R BS.1115.
- ITU93
- Most of the mean SDG per item were above –2.0, and the range was –0.1 to –2.3. There was no significant difference between the data from the two labs.
  - Grusec et al, 1997 and see also Rec. ITU-R BS. 1115.
- MPEG95
  - At least 63 per cent of the mean SDG per item were above –2.0, and the range was –0.2 to –3.8.
  - Meares and Kim, 1995.
- EIA95
  - At least 93 per cent of the mean SDG per item were above –2.0, and the range was 0.1 to –3.7.
  - Grusec et al, 1997.
- DB2
  - Not all the items were used for all conditions.

Validation

- DB3
  - Not all the items were used for all conditions.
- CRC97
  - The mean SDG per item quite uniformly covered the range from 0.1 to –3.6.
  - Soulodre et al, 1998.

The following sections describe the items that were included in the different databases and the conditions that were applied.

2 Items per database

Item	MPEG90	MPEG91	ITU92DI	ITU92CO	ITU93	MPEG95	EIA95	DB2	DB3	CRC97
Accordian/Triangel		*								
Åsa Jinder			*	*	*					
Bag Pipe						*		*	*	
Bag Pipe-2								*		
Bass Clarinet								*		
Bass Guitar	*		*	*						
Bass Synth	*									
Carmen		*								
Castanets	*		*	*	*	*		*	*	
Clarinet					*		*	*	*	*
Clarinet-mono								*		
Clarinet2								*		
Claves									*	
Dalarnas Spelmansförbund "Trettondagsmarschen"			*	*						
Dire Straits "Ride Across the River"			*				*			*
Double Bass										*
Drum								*		
Fireworks	*									
Flute									*	
George Duke		*								
Glockenspiel	*	*				*	*	*	*	
Harpsicord			*	*	*	*	*	*	*	*
Horn								*		
Kettle drums									*	
Marimba								*	*	
MPE mono = Speech male engl. mono								*		
Music and rain							*			*
Muted trumpet							*			
Ornette Coleman	*	*	*	*						
Pearl Jam							*			
Percussion		*						*		
Piano Schubert									*	
Pitch Pipe						*			*	*
Ravel "Feria"			*							
Ry Cooder								*	*	
Ry Cooder (mono)										
Saxophon									*	
Snare drum									*	
Soprano Mozart									*	
Speech female engl						*			*	
Speech female germ									*	
Speech male engl	*	*							*	
Speech male germ			*	*	*				*	
Stravinsky "Wind Octet"			*	*						
Strings								*	*	
Strings mono								*		
Suzanne Vega "Toms Diner"	*	*	*	*				*	*	*
Suzanne Vega with breaking glass							*			
Tambourine		*						*	*	
Tracy Chapman	*									
Triangle			*	*				*	*	
Trumpet								*	*	*
Trumpet (Haydn)	*									
Tuba								*	*	
De sålde sina hemman (violin solo)					*					
Water Sound							*			
Wind Ensemble								*		
Xylophone									*	

3 Experimental conditions

For all bit rates with the indication kbit/s stereo, the total bit rate is given, e.g. 256 kbit/s stereo means that 256 kbit/s is allocated in total for both channels of a stereo signal. If nothing else is indicated, stereo refers to independent channel coding.

3.1 MPEG90

Three bit rates: 64 kbit/s mono, 192 kbit/s and 256 kbit/s stereo, not all material was available for this database.

– Musicam.

– SB-ADPCM.

3.2 MPEG91

Three bit rates: 64 kbit/s mono, 192 kbit/s, and 256 kbit/s stereo.

– MPEG1 Layer I.

– MPEG1 Layer II.

– MPEG1 Layer III.

– MUSICAM.

– ASPEC.

– NICAM.

3.3 ITU92DI

Five distribution codecs: 240 kbits/s stereo.

Each item was processed by the same codec three times in tandem with a 0.1 dB drop in level before each pass.

– MPEG1 Layer II.

– MPEG1 Layer III.

– Dolby AC-2.

– Aware.

– NHK.

3.4 ITU92CO

Six contribution codecs: 360 kbits/s stereo. Each item was processed by the same codec three times in tandem with a 0.1 dB drop in level before each pass.

– MPEG1 Layer II.

– MPEG1 Layer III.

– Dolby AC-2.

– Dolby Low-Delay.

– Aware.

3.5 ITU93

MPEG1 Layer II tandem codec configurations:

– Emission codec alone at 256 kbit/s stereo.

– Emission codec alone at 192 kbit/s stereo (joint stereo coding).

– Eight contribution codecs at 360 kbit/s followed by one emission codec at 256 kbit/s, all in stereo.

– Eight contribution codecs at 360 kbit/s followed by one emission codec at 192 kbit/s, all in stereo.

– Five contribution codecs at 360 kbit/s followed by three distribution codecs at 240 kbit/s and one emission codec at 256 kbit/s, all in stereo.

– Five contribution codecs at 360 kbit/s followed by three distribution codecs at 240 kbit/s and one emission codec at 192 kbit/s, all in stereo.

3.6 MPEG95

Codec implementations (64 kbit/s):

– Twenty-two encoding variations were selected from a larger set of encoding methods available from 6 codecs implementing a subset of 4 low resolution and 17 high resolution time/frequency models.

– Participating organizations were AT&T, Fraunhofer, Sony, GCL, RAI/Alcatel, and Philips.

– All items were monaural recordings presented binaurally.

3.7 EIA95

– Eureka 147/MPEG1 Layer II #1	224 kbit/s stereo (joint stereo coding)
– Eureka 147/MPEG1 Layer II #2	192 kbit/s stereo (joint stereo coding)
– AT&T/Lucent	160 kbit/s stereo
– AT&T/Lucent/Amati #1	128 kbit/s stereo
– AT&T/Lucent/Amati #2	160 kbit/s stereo
– VOA/JPL	160 kbit/s stereo
– USADR-FM #1	128-256 kbit/s stereo (variable bit rate)
– USADR-FM #2	128-256 kbit/s stereo (variable bit rate)
– USADR-AM	96 kbit/s stereo

3.8 DB2

– MPEG1 Layer II	256 kbit/s stereo, 1, 3, 5, 7, and 9 stages
– Dolby AC2	256 kbit/s stereo, 1, 3, 5, 7, and 9 stages
– MPEG1 Layer II	192 kbit/s stereo (joint stereo coding)
– MPEG1 Layer II	64 kbit/s mono
– MPEG2 Layer II	64 kbit/s mono
– MPEG1 Layer II	384 kbit/s stereo
– MPEG1 Layer III	128, 160, 192 kbit/s, all stereo.
– APT-X	256 and 384 kbit/s both stereo.
– Quantization distortion	
– Analogue recording 1, 2, 3 stages	
– Clipping	

3.9 DB3

– NICAM	
– MiniDisc and MiniDisc + Layer II	192 kbit/s, stereo (joint stereo coding)
– Dolby AC2	256 kbit/s stereo, 1, 3, 5, 7, and 9 stages
– MPEG1 Layer II	selection from Swisscom database, >192 kbit/s stereo
– MPEG1 Layer III	128 and 160 kbit/s both stereo (joint stereo coding)
– MPEG AAC	128 kbit/s stereo (joint stereo coding)
– MPEG Layer III	128 + Layer II, 384 + Layer II, 224 kbit/s, all stereo
– Dolby AC3	256 kbit/s stereo
– Dolby AC3	256 + MPEG Layer II, 224 kbit/s, both stereo
– Quantization distortion	
– THD	
– Noise	

3.10 CRC97

–	AT&T PAC	64, 96, 128, and 160 kbit/s, all stereo
–	Dolby AC3	128, 160, and 192 kbit/s, all stereo
–	MPEG1 Layer II software	128, 160, and 192 kbit/s, all stereo
–	MPEG1 Layer II hardware (ITIS)	96, 128, 160, 192 kbit/s, all stereo
–	MPEG4 AAC	96 and 128 kbit/s, both stereo
–	MPEG1 Layer III	128 kbit/s stereo

4 Items per condition for DB2 and DB3

4.1 DB2

		Condition No.	Items
Test site I, NHK Japan			
Layer II, 256 kbit/s	1 stage	CO13	CLA,RYC,SB1,STR
	3 stages	CO11	CLA,RYC,SB1,STR
	5 stages	CO19	CLA,RYC,SB1,STR
	7 stages	CO18	CLA,RYC,SB1,STR
	9 stages	CO15	CLA,RYC,SB1,STR
NBC (Dolby AC2)	1 stage	CO1A	CAS,RYC,STR,WIN
	3 stages	CO12	CAS,RYC,STR,WIN
	5 stages	CO17	CAS,RYC,STR,WIN
	7 stages	CO16	CAS,RYC,STR,WIN
	9 stages	CO14	CAS,RYC,STR,WIN
Test site II, DR Denmark			
Layer II, 256 kbit/s	1 stage	CO2B	CLA,RYC,SB1,STR
Layer II, 192 kbit/s js		CO25	CLA,RYC,SB1,STR
Layer II, 64 kbit/s mono		CO27	MLA,MPE,MTR,MYC
NBC (Dolby AC2)	5 stages	CO29	CAS,RYC,STR,WIN
MPEG2/L2 LSF		CO22	MLA,MPE,MTR,MYC
Analogue 1		CO23	PER
Analogue 2		CO2A	PER
Analogue 3		CO28	PER
Errors 1		CO24	GLO,HRN,TRI
Errors 2		CO21	GLO,HRN,TRI
Clipping		CO26	BAS,CL2,TUB
Test site III, NRK Norway			
Layer II, 384 kbit/s		CO34	CLA,RYC,SB1,STR
Layer II, 256 kbit/s	1 stage	CO31	CLA,RYC,SB1,STR
NBC (Dolby AC2)	5 stages	CO3B	CAS,RYC,STR,WIN
Layer III (ASPEC3), 192 kbit/s		CO32	CLA,STR,TAM,VEG
Layer III (ASPEC3), 128 kbit/s		CO39	CLA,STR,TAM,VEG
Layer III (ASPEC3), 160 kbit/s		**CO3A**	**CLA,STR,TAM,VEG**
APT-X, 256 kbit/s		CO33	HAR,SB2,STR,TPT
APT-X, 384 kbit/s		CO36	HAR,SB2,STR,TPT
Quantizing dist. 1		CO35	DRU
Quantizing dist. 2		CO37	DRU
Quantizing dist. 3		CO38	DRU

Test items

STR Swedish folk music, SR recording, previously used
SB1 Bagpipes, SR recording
SB2 Bagpipes, SR recording
CLA Clarinet, SQUAM 16/2
TAM Tambourine, SR recording, previously used
WIN Stravinski, Wind ensemble, previously used
TPT Trumpet, SQUAM 21/2
HAR Harlequin ensemble, BBC recording G 49/17
VEG Suzanne Vega, old master, previously used
CAS Castanettes, SQUAM 27
SPE German speech, SQUAM 54
RYC Ry Cooder, CD: JAZZ tr 11 (0.25 – 0.47)
PER Percussion, Japanese Bass Marimba, CD: Sony/CBS 32DC 5027
HRN Horn, SQUAM 23/2
GLO Glockenspiel, SQUAM 35/1, previously used
TRI Triangle, SQUAM 32/2
DRU Drums, SQUAM 28
CL2 Clarinet, SQUAM 16/2
BAS Bass Clarinet, SQUAM 17
TUB Tuba, SQUAM 24
MPE Mono mix of SPE
MTR Mono mix of STR
MLA Mono mix of CLA
MYC Mono mix of RYC

Test/Item	Clarinet	Clarinet mono	Ry Cooder	Ry Cooder mono	Bag Pipes 1	Strings	Strings mono	Castanettes	Wind Ensemble	MPE mono	Tambourine	Suzanne Vega	Harpsichord	Bag Pipes 2	Trumpet	Drum	Percussion	Glockenspiel	Horn	Triangle	BAS	Clarinet 2	Tuba
MPEG1 Layer 2, 256 kbit/s, 1 stage	X		X		X	X																	
MPEG1 Layer 2, 256 kbit/s, 3 stages	X		X		X	X																	
MPEG1 Layer 2, 256 kbit/s, 5 stages	X		X		X	X																	
MPEG1 Layer 2, 256 kbit/s, 7 stages	X		X		X	X																	
MPEG1 Layer 2, 256 kbit/s, 9 stages	X		X		X	X																	
Dolby AC2, 256 kbit/s, 1 stage			X			X		X	X														
Dolby AC2, 256 kbit/s, 3 stages			X			X		X	X														
Dolby AC2, 256 kbit/s, 5 stages			X			X		X	X														
Dolby AC2, 256 kbit/s, 7 stages			X			X		X	X														
Dolby AC2, 256 kbit/s, 9 stages			X			X		X	X														
MPEG1 Layer 2, 192 kbit/s joint stereo	X		X		X	X																	
MPEG1 Layer 2, 64 kbit/s mono		X		X			X			X													
MPEG2 Layer 2, 64 kbit/s mono		X		X			X			X													
MPEG1 Layer 2, 384 kbit/s	X		X		X	X																	
MPEG1 Layer 3 (ASPEC 3), 192 kbit/s	X					X					X	X											
MPEG1 Layer 3 (ASPEC 3), 128 kbit/s	X					X					X	X											
MPEG1 Layer 3 (ASPEC 3), 160 kbit/s	X					X					X	X											
APT-X, 256 kbit/s						X							X	X	X								
APT-X, 384 kbit/s						X							X	X	X								
Quantizing distorsion 1																X							
Quantizing distorsion 2																X							
Quantizing distorsion 3																X							
Analogue recording 1 stage																	X						
Analogue recording 2 stages																	X						
Analogue recording 3 stages																	X						
Bit errors 1																		X	X	X			
Bit errors 2																		X	X	X			
Clipping																					X	X	X

4.2 DB3

Item/Test	Name	1	2MD	2MDL2	3 1Step	3 3Step	3 5Step	3 7Step	3 9Step	4	5	6 low	6 high	7	8	9	10	11	12	13
13(1) Flute	flu							X	X											
16(2) Clarinet	cla				X		X			X		X		X	X					
20(1) Saxophon	sax							X	X											
21(2) Trumpet	tpt					X	X													
24(2) Tuba	tub																	X		
26(1) Claves	clv		X	X			X	X	X											X
27 Castanets	cas									X			X			X	X		X	
28 Snare drum	sna	X																		X
30 Kettle drums	ket																	X		
32(1/2) Triangle	tri		X	X	X	X				X									X	
35(1/2) Glockenspiel	glo											X	X						X	
36(1) Xylophone	xyl																	X		
40(1) Harpsicord	hrp				X	X					X			X						
49 Speech female engl	sfe							X	X										X	
54 Speech male germ	smg	X	X	X	X	X	X				X									
60 Piano Schubert	pia																	X		X
61 Soprano Mozart	sop		X	X																
53 Speech female germ	sfg																			
50 Speech male engl	sme												X		X					
Ref_tam	tam	X														X	X			
Ref_str	str															X	X			
Ref_har	har																			
Ry Cooder	ryc										X					X	X			
Susanne Vega	veg									X	X	X		X	X					
Pitch Pipe	pip											X	X	X	X					
Marimba	mar	X																		X
Bag Pipe	sb1									X										
Name		i	b	s	g	e	p	c	o	f	j	r	k	l	m	h	d	q	n	a

1.) NICAM
2.) MD and MD + L2 (192 kbit/s)
3.) AC2 (256 kbit/s), 1, 3, 5, 7,9
4.) L2 (ST d-b, ³192 kbit/s)
5.) Layer2 (256 kbit/s), 8 stages
6.) Layer3, 128 and 160
7.) AAC, 128 kbit/s
8.) L3 (128) + L2 (384) + L2
9.) AC3 (256)
10.) AC3 (256) + L2 (224)
11.) Quantizing distortion
12.) THD
13.) Noise

Name Examples:

Reference: irefflu
Test: icodflu

Glossary

Absolute Error Score (AES)

The Absolute Error Score is derived from a formula developed especially for evaluating the quality of the results obtained from an objective perceptual measurement method. It takes the confidence intervals of the average values of subjective listening tests into account.

Basic Audio Quality

The Basic Audio Quality is defined as a global subjective attribute which includes any and all detected differences between the Reference Signal and a processed version of it.

Coding Margin

The Coding Margin is a quality parameter which measures the headroom of inaudible coding artefacts to the threshold when these artefacts become audible.

Model Output Variables (MOV)

The Model Output Variables are intermediate output values of the perceptual measurement method. These variables are based on basic psycho-acoustical findings and may therefore be used to characterize the coding artefacts further.

Objective Difference Grade (ODG)

The Objective Difference Grade is the main output parameter of the perceptual measurement method. It corresponds to the SDG and is the measurement parameter giving the global Basic Audio Quality. The ODG has a range between 0 .. – 4.

Off-line measurement

Measurement procedure which does not interact with the ongoing programme transmission.

On-line measurement

Measurement procedure which relies on the ongoing programme transmission, or parts thereof.

Subjective Difference Grade (SDG)

In a listening test according to Recommendation ITU-R BS.1116 the Basic Audio Quality of the hidden reference and the processed version of the reference are graded on the five-grade impairment scale. The difference grade is defined as the grade given to the Signal Under Test minus the grade given to the Reference Signal. The SDG should ideally have a range between 0 .. – 4. If the reference was not identified correctly, the SDG is positive.

Abbreviations

ADB	Average Distorted Block
AES	Absolute Error Score
ASD	Auditory Spectral Difference
Avg	Average (linear)
BAQ	Basic Audio Quality
Bw	Bandwidth
CI	Confidence Interval
CM	Coding Margin
DBn	Database n (1, 2 or 3)
DC	Direct Current
DFT	Discrete Fourier Transform
DIX	Disturbance Index
DUT	Device Under Test
EHS	Error Harmonic Structure
ERB	Equivalent Rectangular Bandwidth
fac	factor
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
IIR	Infinite Impulse Response
ISO	International Standards Organization
JNLD	Just Noticeable Level Difference
MFPD	Maximum Filtered Probability of Detection
MOV	Model Output Variable
MPEG	Moving Picture Expert Group
NL	Noise Loudness
NMR	Noise-To-Mask Ratio
OASE	Objective Audio Signal Evaluation
OCM	Objective Coding Margin
ODG	Objective Difference Grade
PAQM	Perceptual Audio Quality Measure

PEAQ	Objective Measurements of Perceived Audio Quality
PERCEVAL	Perceptual Evaluation
POM	Perceptual Objective Measure
REF	Reference Signal
res	Resolution
r.m.s.	Root Mean Squared
ROEX	Rounded Exponential
ROV	Rate of Output Values
SCM	Subjective Coding Margin
SDG	Subjective Difference Grade
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level
SUT	Signal under Test
THD	Total Harmonic Distortion
Win	Windowed Average

REFERENCES

Aures W., [September 1984] *Berechnungsverfahren für den Wohlklang beliebiger Schallsignale, ein Beitrag zur gehörbezogenen Schallanalyse*. Dissertation an der Fakultät für Elektrotechnik der Technischen Universität München.

Beerends J.G. and Stemerdink J.A., [December 1992] *A perceptual audio quality measure based on a psychoacoustic sound representation*, J. Audio Eng. Soc., Vol. 40, pp 963-978.

Beerends J.G. and Stemerdink J.A., [February 1994] *Modeling a cognitive aspect in the measurement of the quality of music codecs*, Contribution to the 96th AES Convention, Amsterdam, preprint 3800.

Beerends J.G. and Stemerdink J.A., [March, 1994] *A perceptual speech quality measure based on a psycho-acoustic sound representation*, J. Audio Eng. Soc., Vol. 42, pp 115-123.

Beerends J.G., van den Brink W.A.C. and Rodger B., [May 1996] *The role of informational masking and perceptual streaming in the measurement of music codec quality*, Contribution to the 100th AES Convention, Copenhague, preprint 4176.

Brandenburg K., [1987] *Evaluation of quality for audio encoding at low bit rates*, Contribution to the 82nd AES Convention, London, preprint 2433.

Bregman A.S., [1990] *Auditory scene analysis: The perceptual organisation of sound,* MIT Press, Cambridge MA.

Cohen E. A. and Fielder L.D., [May 1992] *Determining noise criteria for recording environments*, J. Audio Eng. Soc. Vol. 40, pp 384-402.

Colomes C., Lever M., Rault J.B. and Dehery Y.F., [April, 1995] *A perceptual model applied to audio bit-rate reduction*, J.Audio Eng. Soc., Vol. 43, pp 233-240.

Feiten B., [March 1997] *Measuring the Coding Margin of Perceptual Codecs with the Difference Signal*. 102nd AES-Convention München, preprint 4417.

Grusec, T., Thibault L. and Soulodre, G. [September 1997] *EIA/NRSC DAR systems subjective tests*. *Part 1*: *Audio codec quality*, IEEE Transactions on Broadcasting, Vol. 43, No. 3.

Karjalainen J., [March 1985] *A new auditory model for the evaluation of sound quality of audio system*, Proceedings of the ICASSP, Tampa, Florida, pp 608-611.

Leek M.R. and Watson C.S., [1984] *Learning to detect auditory pattern components*, J. Acoust. Soc. Am. Vol. 76, pp 1037-1044.

Meares D.J., Kim, [July 1995] S-W, "NBC time/frequency module subjective tests: overall results", ISO/IEC JTC1/SC29/WG11 N0973 MPEG95/208.

Moore B.C., [1986] *Frequency Selectivity in Hearing,* Academic Press, London.

Moore B.C., [1989] *An introduction to the psychology of hearing*, Academic Press, London.

Paillard B., Mabilleau P. Morisette S. and Soumagne J., [1992.] Perceval: *Perceptual evaluation of the quality of audio signals*, J. Audio Eng. Soc., Vol. 40, pp 21-31.

Schroeder M.R., Atal B.S. and Hall J.L., [December 1979] *Optimizing digital speech coders by exploiting masking properties of the human ear*, J. Acoust. Soc. Am., Vol. 66, pp 1647-1652.

Soulodre G., Grusec T., Lavoie M. and Thibault L., [March 1998] *Subjective evaluation of state-of-the-art 2-channel audio codecs*, Journal of the Audio Engineering Society.

Sporer T., [October 1997] *Objective audio signal evaluation — applied psychoacoustics for modeling the perceived quality of digital audio*, 103rd AES-Convention, New York, preprint 4512.

Terhardt E., [1979] *Calculating Virtual Pitch, Hearing Research*, Vol. 1, pp 155-182.

Thiede T. and Kabot E., [1996] *A New Perceptual Quality Measure for Bit Rate Reduced Audio*, Contribution to the 100th AES Convention, Copenhague, preprint 4280.

Treurniet W.C. [1996] *Simulation of individual listeners with an auditory model*. Proceedings of the Audio Engineering Society, Copenhague, Danemark, Reprint Number 4154.

von Bismarck G., [1974] *Sharpness as an attribute of the timbre of steady sounds*. Acustica 30, pp 159-172.

Zwicker E. and Feldtkeller R., [1967] *Das Ohr als Nachrichtenempfänger*. Stuttgart: Hirzel Verlag.

Zwicker E. and Fastl H., [1990] *Psycho-acoustics, Facts and Models*. Berlin; Heidelberg: Springer Verlag.

BIBLIOGRAPHY

Grusec T., Thibault L., and Soulodre G. [1995] *Subjective evaluation of high quality audio coding systems: methods and results in the two-channel case*, preprint 4065 (F-5), Proceedings of the AES, New York.