UNIVERSITY OF NOVI SAD
FACULTY OF TECHNICAL SCIENCIES

ANALYSIS OF METHODS FOR OBJECTIVE EVALUATION OF QUALITY OF AUDIO SIGNALS AND APPLICATION IN IMPLEMENTATION OF AN ENCODER ON A CLASS OF DIGITAL SIGNAL PROCESSORS

- Master's Thesis –

Mentor:
Ph.D. prof. Vladimir Kovačević

Candidate:
B.Sc. Goran Marković

Novi Sad, July 2006.

This thesis was originally written in Serbo-Croatian language. Original title is “Analiza metoda za objektivno merenje kvaliteta audio signala i njihova primena pri realizaciji kodera u jednoj klasi digitalnih signalnih procesora” and the original translation of the title was “Analyses of methods for objective evaluation of quality of audio signals and application in implementation of a coder on a class of digital signal processors”[1]. The original can be obtained in The Library of the Faculty of Technical Sciences at University of Novi Sad in Serbia or from the author.

Scientific field:	Electrical Engineering
Scientific discipline:	Computer engineering
Subject/Key words:	Computer engineering / audio, objective evaluation of audio quality, psychoacoustics, artificial neural network, digital signal processor
Abstract:	Methods for objective evaluation of audio quality and their use in implementation of an encoder on a digital signal processor are analyzed in the thesis. Changes of the model described in ITU-R BS.1387 are proposed.
Defended on:	03. 11. 2006.

Thesis defends board:

President:	Branko Kovačević, Ph D
Member:	Vladimir Kovačević, Ph D
Member:	Miroslav Popović, Ph D
Member:	Vlado Delić, Ph D
Member:	Nikola Teslić, Ph D

[1] Original translation had mistake that “koder” was translated as “coder” instead of “encoder"

Abbreviations

DVD	–	*Digital Versatile Disc*
GSM	–	*Global System for Mobile Communications*
VoIP	–	*Voice over Internet Protocol*
DAB	–	*Digital Audio Broadcasting*
DVB	–	*Digital Video Broadcasting*
PEAQ	–	*Perceptual Evaluation of Audio Quality*
SDG	–	*Subjective Difference Grade*
MOV	–	*Model Output Variable*
ODG	–	*Objective Difference Grade*
PCM	–	*Pulse Code Modulation*
FFT	–	*Fast Fourier Transformation*
SPL	–	*Sound Pressure Level* ($L_{SPL}=20\log_{10}(p/p_{ref})$ [dB])
ATH	–	*Absolute Threshold of Hearing*
DC	–	*An offsetting of a signal from zero*
IIR	–	*Infinite Impulse Response*
FIR	–	*Finite Impulse Response*
DI	–	*Distortion Index*
FS	–	*Full Scale*
ITD	–	*Interaural Time Difference*
ILD	–	*Interaural Level Difference*
HRTF	–	*Head Related Transfer Function*
AES	–	*Absolute Error Score, ODG error measure proposed in ITU-R BS.1387*

Symbols

N_F	–	Length of a Hann window, i.e. length of an FFT
F_S	–	Sampling frequency of input signal
Z	–	Number of frequency subbands
Δz	–	Distance between the central frequencies of subbands

1. Introduction

The quality of audio signal is one of the key factors in the making of sound reproduction devices and systems for broadcasting such as: MP3 player, DVD, GSM, VoIP, DAB, DVB etc. Traditional objective methods of testing, which include Signal to Noise Ratio – SNR or Total Harmonic Distortion – THD, usually do not produce representative evaluation of the quality of an audio signal. The cause of this lies in the non-linear distortions produced by a codec as well as in disregarding the distinctive features of the human hearing ability. Codecs can rather efficiently compress simple signals which are used in the traditional testing methods, and the distortion occurs when they are applied on extremely complex signals.

The listener is the only relevant factor in the evaluation process in which he or she should be selected to represent the final consumer of the product being tested.

There are two groups of tests: the first is the one in which an original signal is unavailable for detection and the second one in which it is available. For evaluation of voice quality, e.g. in GSM or VoIP, an original signal is not present for detection and a test signal has lots of distortions. In this Master's thesis my aim is to present and describe the tests with available original signal and compressed signal with approximately transparent quality.

Modern audio encoders are State-of-the-Art products of current technology: although a high level of compression is achieved, the sound remains almost identical to the original. Most listeners are not able to hear the difference between the compressed and the original audio signal. It is the consumer, or at least the majority of them who confirm the quality of a product by expressing their satisfaction with it. When an expert listener confirms the high quality of a product, it is expected that average listeners will also be satisfied with its quality, i.e. they will not detect the difference between the compressed and the original audio signal.

These requirements lead to the proposal of the method for subjective evaluation of small impairments in audio systems ITU-R BS.1116, in 1994, and its revision in 1997 [ITU97]. This proposal gives a definition of an expert listener, whose evaluation of quality is the relevant one.

Since it is not always practical, or even possible, to carry out the methodology proposed in ITU-R BS.1116, or any other relevant subjective evaluation of quality, redefinition of the objective testing methods is required. One particular requirement that stands out is that an objective method must provide as similar results to those from the subjective tests as possible.

The first attempts in changing the objective methods, in a way that they meet new requirements, were made back in 1979. [SCHR79]. They were inspired at first by the development of voice codecs. Parallel development of digital music coding at the beginning of the 1980s required objective methods which would be applied on it, too. K. Brandenburg[2] proposed the first method for evaluation of quality of a compressed music signal [BRA87] in 1987. Its development was made possible thanks to the previous

[2] Ph.D.prof. Karlheinz Brandenburg, born in 1954., Erlangen, Germany, since 2000 full professor at Institute for media technologies within Technical University of Ilmenau and director of Fraunhofer institute.

psychoacoustic experiments, like the ones performed by Zwicker[3] [ZWI67]. The objective method, which would be able to provide equally good results for both voice and music, is at this moment not possible to create [KEY99].

International Telecommunications Union - ITU formed an action group 10/4. Their intention was to propose an objective model based on human perception for measuring the quality of wide band audio codecs. Certain already familiar ideas were taken under consideration: NMR, OASE, PAQM, PERCEVAL, POM, DIX and Toolbox approach. The group proposed the model in 1998 called Perceptual Evaluation of Audio Quality – PEAQ and published the document ITU-R BS.1387, with audited revision ITU-R BS.1387-1 [ITU01]. A very important fact is that PEAQ was based on the modeling of the subjective test proposed in ITU-R BS.1116 by algorithmic approach. Therefore, it is essential to understand the frame of subjective experiments through the interpretation of the results of objective tests, e.g., as it is the case with subjective tests, evaluation of artificial, simple signals is not relevant, so that is why music signals, voice signals and the like are relevant for evaluation process.

Testing of implementation of a codec on digital signal processors inevitably requires the existence of a reliable and accurate objective measuring of degraded audio quality. In practice PEAQ turned out to be insufficient enough in terms of reliability and accuracy. In creating of the proposal ITU-R BS.1387 a crucial factor was a limited power of the existing processors, therefore some models and methods were selected so that the tests could be carried out in an acceptable time frame, which affected and diminished their precision.

This Master's Thesis will present analysis of the existing models and methods, and select the most suitable ones for the development of a new implementation, taking the proposal ITU-R BS.1387 as a frame of reference. Method and model suitability is determined by their precision and complexity, i.e. the execution time. Complexity is important to enable possible implementation, which would run in real time, on a DSP with limited resources.

After the implementation of the selected models and methods, neural network was trained, whose structure was taken from ITU-R BS.1387, so that the previously defined set of test vectors would give results as similar as possible to those of the subjective tests.

Finally, the reference implementation of Opera and two implementations which came out of this Master's Thesis were compared - APEAQ, the implementations being based on the proposal ITU-R BS.1387, and modified APEAQ, in which the selected methods were incorporated and a new neural network was trained.

[3] Ph.D.prof. Eberhard Zwicker, (1924.-1990.), Oringen, Germany, director of Institute of electroacoustics within Technical University of Munich from 1967. to 1990.

2. Aim and application of objective evaluation

This chapter will present the basic methodology in contemporary evaluation of audio quality through subjective and objective tests, as well as the application of objective evaluation with the emphasis on its use for codec implementation on a digital signal processor.

2.1. Subjective evaluation of audio quality

The subjective evaluation of small impairments between audio signals was described in [ITU97] and proposed as a standard ITU-R BS.1116. This standard is used for evaluation of the systems which produce impairments that cannot be detected without rigid control of experimental conditions and appropriate statistical analysis. It is recommended not to be used for measuring the systems which cause profound quality impairments.

It is necessary for the listeners that take part in the test to be expert listeners, i.e. to have above average hearing. This requirement would enable the detection of problems which could be discovered during long-term use of the tested systems. In order to determine the expertise of a listener, the criterion-referenced test is set before actual testing. It leads to the elimination and to disregarding of evaluation results of those listeners who were unable to tell the difference between the references and the tested audio signals.

It is also equally important for the tested audio material to be relevant, so that, based on its evaluation, one could make an assessment of a system. It might be useful to provide *low anchor* signals which expert listeners, unlike non-expert ones, can easily detect. The sequences last from 10 to 25 seconds. Hearing devices, including headphones or speakers, must be of high quality, and also there are certain standards required regarding the room used for the testing.

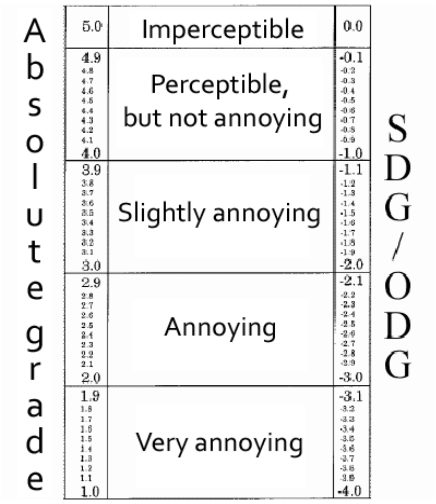
The application of a subjective test is called “double-blind triple-stimulus with hidden reference”. A listener is presented with 3 signals: A – known reference signal, B and C – hidden reference signal and test signal. The listener does not know which of the presented signals B and C is the reference signal and which is the test one. The signals can be presented and listen to randomly, infinite number of times. A so-called basic audio quality of the signals B and C is evaluated, which encompasses all detected degradations compared to the signal A. The grade scale ranges from 1 to 5, and it is continuous, which means possible grades can be 3.4, 2.8 etc. The grades are presented on the scale in the picture 2.1.

Subjective Difference Grade - SDG is the difference between absolute grades for the basic audio quality of the tested signal and the hidden reference:

$$SDG = Grade_{test} - Grade_{hidden\_reference} \quad (2.1)$$

It has the range between -4 and +4. -4 represents annoying impairment of the tested signal, and SDG higher than 0 shows listener’s inability to differentiate the tested signal from the reference one. So far there were subjective tests which pointed out the cases in which SDG was up to +0.25. Considerably higher grades exist only in theory and if they do occur they probably imply an error made during result processing.

Finally, the average SDG of all participants for each of the tested signals is determined.



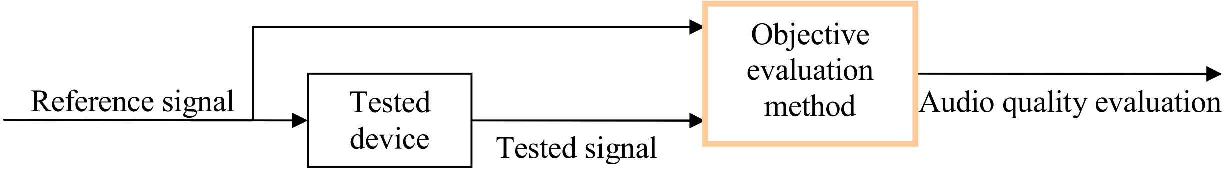
2.1 – Subjective grade scale

2.2. Objective evaluation of audio quality and its application

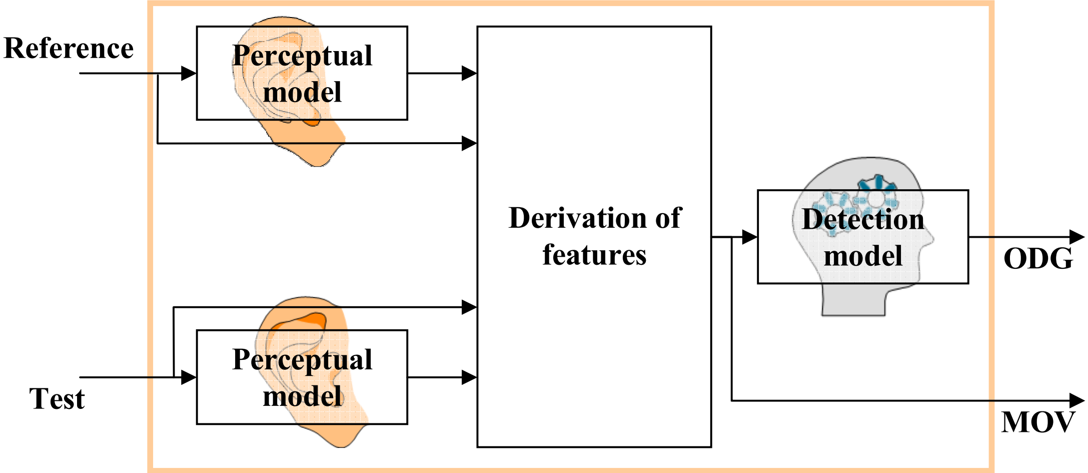
As it was already pointed out, the use of subjective tests is in many cases complicated and impractical, and sometimes even impossible. In such cases objective tests can be used, which provide us with similar evaluations as the subjective ones.

2.2.1. Concept of objective evaluation

All of the well-known methods for evaluation of audio quality, also including ITU-R BS.1387, are very similar. Their structure is shown in the block-diagram:



2.2 – Basic concept of objective evaluation



2.3 – Standard structure of an objective evaluation model

At the input there are reference and tested signals. The tested signal is usually the output of a codec for the reference signal. The first step should be modeling of the processes in a human ear for each signal. Comparing the outputs from a perceptual model results in distortions of the tested signal. That way we get Model Output Variables – MOV, which are reduced by some algorithms for simulation of cognitive processes to one parameter that presents the average value of the tested audio quality. In ITU-R BS.1387 artificial neural network is used for the simulation of cognitive processes, and the final average grade is called Objective Difference Grade – ODG. ODG corresponds to SDG from the subjective tests, the scale and its value is the same (picture 2.1).

2.2.2. Application of objective model

Objective evaluation of audio quality is first of all used on audio codecs. There are already known applications, as well as the proposals for those that have not been used yet.

Application	Description
Implementation assessment	Assessment procedure for different implementations of sound processing equipment, mainly audio codecs
Final evaluation of functional quality	For final testing of equipment before launching it on the market or using it
On-line monitoring	Observation of transmission quality of an audio signal
Equipment and connections status	Elaborate analysis of part of the equipment or connections
Codec identification	For determination of type or implementation of a codec
Codec development	Through implementation of the existing codecs or the development of new ones
Network and system planning	Optimization of prices and features of networks and systems
Aid for subjective tests	For selection of material that will be included in subjective tests

Table 2.1 – Application of objective evaluation

Implementation assessment. When buying sound processing equipment (e.g. an audio codec), customers need to try out different products in order to buy suitable one or the one that meets their needs. This requires great precision, especially for ranking different products. An example for quality evaluation of audio codecs can be found in [SAL04], [HYD06] and [PRO01].

Final evaluation of functional quality. Before a certain piece of equipment, electrical circuit or the entire equipment is put in use, a quick checking reduces the possibility of malfunction. For this final testing the speed is more important than its precision.

On-line monitoring. During the broadcast of a radio or television audio signal it is possible to observe its quality. This requires working in real time and consequently a quick enough algorithm.

Equipment and connections status. In order to guarantee the working state of audio connections or equipment, thorough testing of their quality is necessary from time to time.

Unlike on-line monitoring, real time observation is unnecessary. Great precision and elaborate testing is required.

Codec identification. In order to identify which codec is used for compressing of a tested signal, measuring system should compare patterns of codecs' characteristics. Data base with characteristics' patterns of known codecs is required. However, the issue in question is this application's feasibility, since there is no measure for determination of similar patterns.

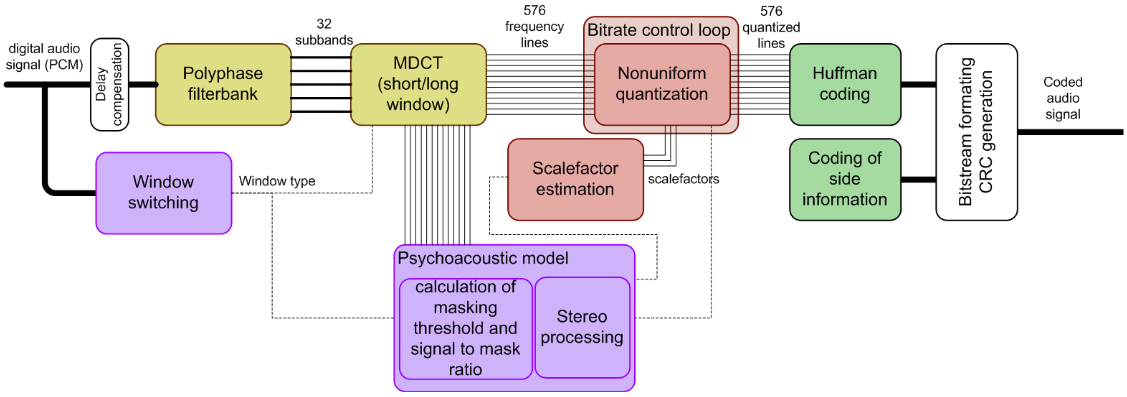
Codec development. Objective evaluation can be applied in the implementation of an encoder or decoder on digital signal processors. The evaluation of the audio signal, which is compressed by the implemented encoder, must not be considerably worse than that of the reference encoder on a PC platform. Also, it can be used in the development of a new codec – through selection of parameters which have impact on the quality or when checking for possible bugs that may occur during algorithm implementation. This application requires great precision in the measuring process. An example of successful error detection during algorithm implementation can be found in [HYD04].

Network and system planning. Computer networks are also used for music, voice and video transmissions in real time. The quality of network affects the transmission of such data. For network planning, beside traditional methods, perceptual evaluation can also be used. An example is given in [CON02]. Also, an example of the objective evaluation method in a system planning for finding music information can be found in [REI04].

Aid for subjective tests. Selection of the audio material for subjective tests is of great importance for their relevance. Continuous and extensive listening may produce inaccurate results due to listeners' fatigue. Objective evaluation can be used for the selection of such samples which would contribute to achieving more accurate results of the subjective tests.

2.2.3. Application in implementation of encoders on digital signal processors

During an encoder implementation on a digital signal processor, the starting point is the existing implementation on a PC platform. If there are more available implementations, objective evaluation can determine which of them could achieve the best quality. It is expected that the difference in quality during implementation on a digital signal processor and a PC platform is minimal, and therefore, the ranking of implementation on a PC platform is identical as it is on a DSP platform. During the selection of implementation, a price, complexity of its realization and available processor resources also play an important role. If it turns out that based on these parameters only one implementation is appropriate for realization on a DSP platform, objective evaluation is unnecessary.



2.3 – Implemented MP3 encoder

The bitstream of MPEG-1 layers 1, 2, and 3 is standardized in ISO/IES11172-3 [ISO92]. MPEG-2 standard is described in ISO/IES 13818-3 [ISO98]. MPEG-2.5 is not a part of an official standard and represents additional sampling rates compared to MPEG-2. Algorithm for MPEG encoder is not standardized. Any kind of implementation is allowed, provided that an encoder produces appropriate functional output. Different encoder implementations can profoundly differ in quality, and objective evaluation can therefore contribute to their improvement. Only the format of a compressed audio signal (bitstream) is standardized. On the other hand, some parts of the algorithm for decoder are standardized and a decoder must have required precision.

As an example there is the implementation of MPEG-1/2/2.5 Layer 3 (known as MP3) encoder made in Micronas[4]. The starting point is implementation of Fraunhofer institute[5]. APX processor also has floating point support, but with less accuracy compared to a PC, which inevitably leads to differences in the outputs.

The first step was to customize a code for arithmetics of APX processor, which lead to different results from the original ones. Apart from these differences, some bugs were found in the referent implementation, and some parts of algorithms were optimized, too [GOR04]. That is why it was necessary to compare outputs of the original code and the one adjusted for APX.

Before checking the quality, it is important to check whether the output bitstream is properly formed. Several decoders were used (mpg123, mad, 13dec, LAME, Winamp and Cooledit). A few outputs of original FhG encoder could not be decoded. All decoders successfully decoded the outputs of the customized encoder. This confirms that some errors in the original encoder were properly corrected. Since it is impossible to decode the outputs for which Fraunhofer implementation causes an error, these are not used for comparative evaluation of quality.

Comparative output quality evaluation of the original implementation and the customized one is carried out in a few steps:

- Original PCM signals are encoded by the original and the customized encoder
- Encoder outputs are decoded by a proven decoder
- Decoder outputs are aligned in time with the original signal

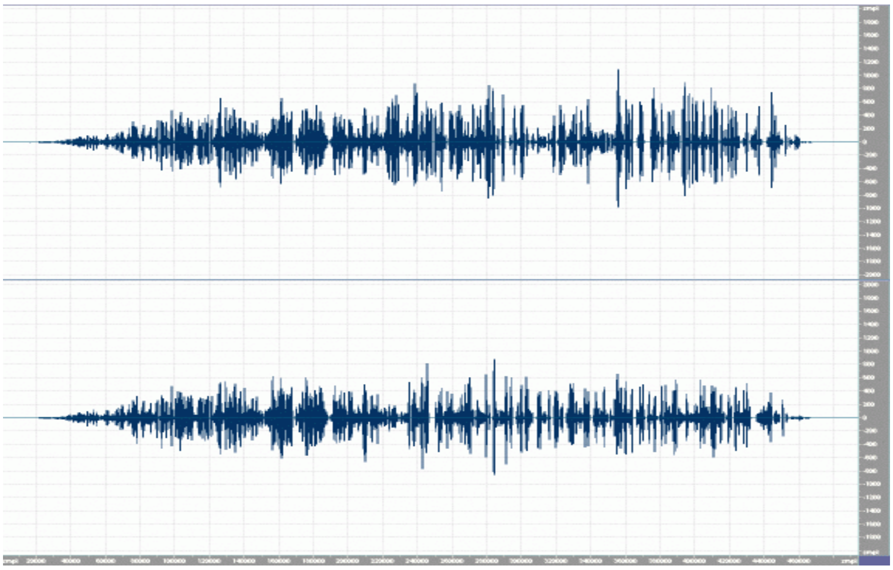
[4] Micronas, http://www.micronas.com

[5] Fraunhofer-Gesellschaft, abbrevation FhG, http://www.fraunhofer.de

- If signal's sampling rate is 8, 12, 24 or 32 kHz, then it is changed to 48 kHz. Signal with sampling rate 11025 and 22050 Hz are resampled to 44.1 kHz.
- ODG_{ref} for reference encoder output is calculated through comparison with the original PCM signal
- ODG_{test} for customized encoder output is calculated through comparison with the original PCM signal
- ODG_{ref} and ODG_{test} are compared

On the available set of tests, the changed encoder had a better ODG. The difference between ODG_{ref} and ODG_{test} was below the accuracy of the objective evaluation method. All in all, this proves that the changes were properly made.

Customized encoder is compiled by a complier and is started in the APX simulator. Due to APX architecture and optimizations done by the complier, outputs which are produced by the MP3 encoder, being started on the simulator, are not always identical to the encoder outputs on a PC platform. The outputs on some signals differ from each other up to 12 bits.

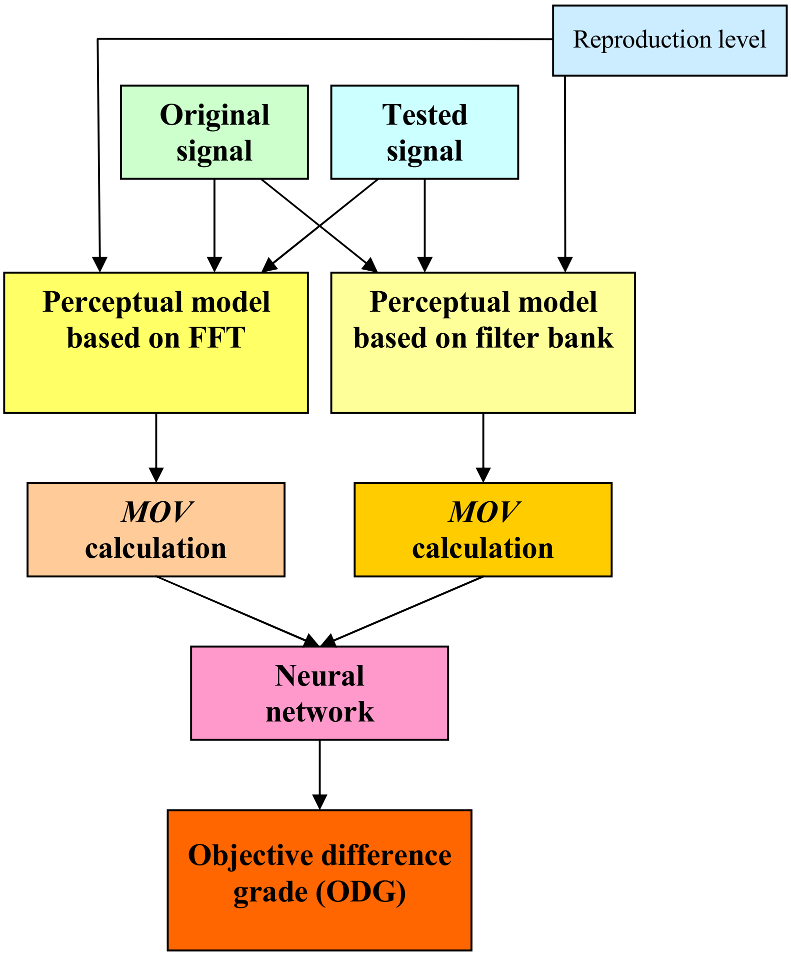


2.4 – Difference of the outputs on PC and APX platforms

The picture shows the signal which produced the greatest differences. These differences are within the 12-bit range. For this, and a few other signals with great differences, the same method is applied as the one used during the comparison of the Fraunhofer and the customized source codes. The difference in ODG is 0.037, which is less than objective evaluation accuracy (less than 0.1). On other signals the difference is a lot smaller, below 0.01. Therefore, it is expected that the listeners would not be able to hear this difference in signals.

This application of objective evaluation is not only limited to the testing of implementation of encoders. The same method can be applied on the implementation of decoders, when it is impossible to achieve required precision according to the standard or when it has to be sacrificed due to limited resources.

3. Proposal ITU-R BS.1387 for objective evaluation of audio quality


Reproduction level
Original signal
Tested signal
Perceptual model based on FFT
Perceptual model based on filter bank
MOV calculation
MOV calculation
Neural network
Objective difference grade (ODG)

3.1 General block-diagram of PEAQ - Advanced version

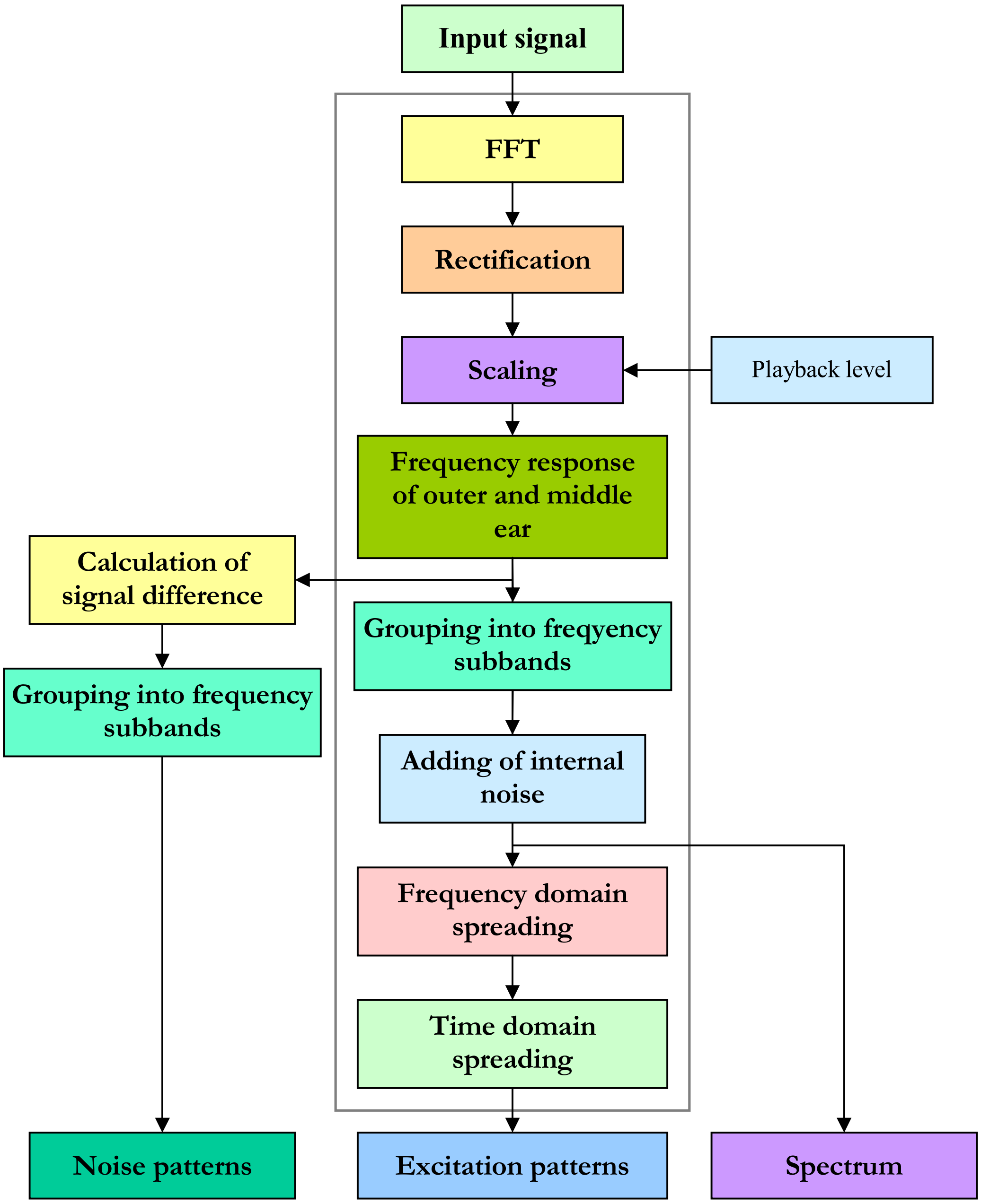
The method for objective evaluation of audio quality described in ITU-R BS.1387 [ITU01] includes two perceptual models (one based on the filter bank and the one based on FFT), MOV calculation and mapping from MOVs to the objective evaluation of audio quality – ODG. Two versions are described: the basic one, which uses only FFT and the advanced one, which uses both the filter bank and FFT.

The details from the basic version, which are not used in the advanced one, will not be described. There are at least two publicly accessible implementations of the basic version ([LER02] and [KAB04]), as well as its detailed description in [GOT03]. Since the most part of the basic version is included in the advanced one, the parameters and MOVs, which are used only in the basic version, will be left out, consequently simplifying further description.

In the advanced version, three MOVs are determined by perceptual model output based on the filter bank and two MOVs are defined by perceptual model output based on FFT.

Playback level is expressed in decibels and has influence on the excitation patterns spreading in the frequency domain. The original and the tested signals must be aligned in time. Sampling rate F_S has to be 48 kHz.

3.1. Perceptual model based on FFT


Input signal
FFT
Rectification
Scaling
Playback level
Frequency response of outer and middle ear
Calculation of signal difference
Grouping into freqyency subbands
Grouping into frequency subbands
Adding of internal noise
Frequency domain spreading
Time domain spreading
Noise patterns
Excitation patterns
Spectrum

3.2 Perceptual model based on FFT

Excitation patterns are calculated only for the original signal. *Spectra* are calculated for both the original and the tested signals, and *noise patterns* represent their difference.

Input signal is processed in frames of $N_F = 2048$ samples with 50% overlap:

$$t_n[k] = t[n \cdot \frac{N_F}{2} + k], \qquad 0 \le k < N_F, n = 0,1,2,... \quad (3.1)$$

3.1.1. Transformation into frequency domain and scaling

The frame is windowed with a scaled Hann window:

$$h_w[k] = \frac{1}{2}\sqrt{\frac{8}{3}}\left[1 - \cos(2\pi \frac{k}{N_F - 1})\right] \quad (3.2)$$

$$t_w[k] = h_w[k] \cdot t_n[k],^6 \qquad 0 \le k < N_F \quad (3.3)$$

The frame of windowed data is transformed from time to frequency domain by using discrete Fourier[7] transformation (DFT):

$$x[k_f] = \frac{1}{N_F}\sum_{k_t=0}^{N_F-1} t_w[k_t] e^{-\frac{j2\pi k_f k_t}{N_F}}, \qquad 0 \le k_f < N_F \quad (3.4)$$

DFT is realized by FFT algorithm.

Rectification is achieved by transformation of a complex spectrum to amplitudes, which are then scaled, so that a certain playback level of signal is simulated. Some parts of a perceptual model depend on the playback level. Scaling factor is calculated with the following formula:

$$fac = \frac{10^{\frac{L_p}{20}}}{\gamma(f_c)\sqrt{\frac{8}{3}}\frac{A_{\max}}{4}\frac{N_F - 1}{N_F}} \quad (3.5)$$

Where L_p is the assumed playback level of the full scale sine signal, which, if there are no other data, is set to already determined value of 92 dB SPL. $A_{\max}$ is the maximum amplitude of the sine signal (32768 for 16bit input), $\gamma(f_c)$ is the factor which depends on the sine signal's frequency, whose maximum absolute value is observed. $\gamma(f_c)$ = 0.8497 for frequency of 1019.5 Hz proposed in [ITU01]. Scaling factor calculation is more elaborately analyzed in [KAB06].

3.1.2. Absolute threshold of hearing and frequency response

Absolute threshold of hearing is modeled by frequency response of the outer and middle ear, and by adding of internal noise (after grouping into frequency subbands). Formula that is used is given in [TER79] with the change of the last exponent from 4.0 to 3.6:

$$threshold / dB = 3.64(f / kHz)^{-0.8} - 6.5e^{-0.6(f/kHz - 3.3)^2} + 0.001(f / kHz)^{3.6} \quad (3.6)$$

Coefficients which are used in this step are given with:

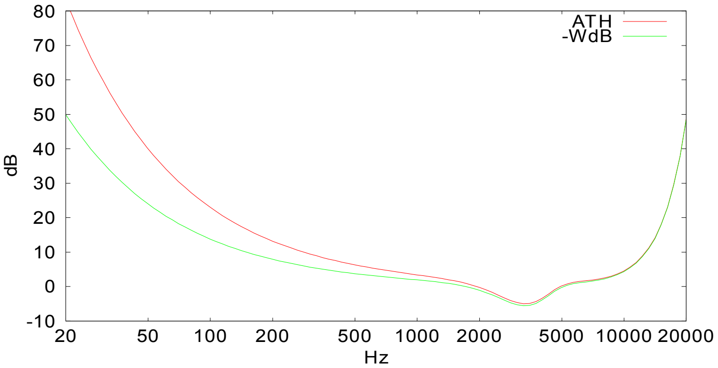
[6] Formulas in chapter 3 are taken from [ITU01] and [KAB06] unless stated otherwise. Also, they were changed to some extent due to consistency and clarity.

[7] Jean Baptiste Joseph Fourier, (1768.-1830.), France.

$$W_{dB}[k]/dB = -0.6 \cdot 3.64 \cdot (\frac{f[k]}{kHz})^{-0.8} + 6.5 \cdot e^{-0.6(\frac{f[k]}{kHz}-3.3)^2} - 0.001 \cdot (\frac{f[k]}{kHz})^{3.6} \qquad (3.7)$$

$$f[k] = k \cdot \frac{F_S}{N_F}, \qquad 0 \le k_f < \frac{N_F}{2} \qquad (3.8)$$

$$W[k] = 10^{\frac{W_{dB}[k]}{20}} \qquad (3.9)$$



3.3 Absolute threshold of hearing (ATH) and frequency response (W_{dB})

Absolute threshold of hearing (3.6) and the inverse frequency response of outer and middle ear (3.7) are shown on the graph. Their difference represents the internal noise (3.16).

The scaled outputs from DFT are multiplied by coefficients (3.9):

$$X[k] = |fac \cdot x[k] \cdot W[k]|^2, \qquad 0 \le k_f < \frac{N_F}{2} \qquad (3.10)$$

The result is the energy of each of the spectrum lines with the applied frequency response of the outer ear.

3.1.3. Calculation of signal difference

Signal difference is calculated with the following formula:

$$X_{diff}[k] = X_{ref}[k] - 2 \cdot \sqrt{X_{ref}[k] \cdot X_{test}[k]} + X_{test}[k], \qquad 0 \le k < \frac{N_F}{2} \qquad (3.11)$$

where X_{ref} and X_{test} are spectrum lines of the original and the tested signal, in that order.

3.1.4. Grouping into frequency subbands[8]

The energies of single spectrum lines are then grouped, changing frequency scale, using the following formula:

[8] The term frequency subband is used, not critical band, because the bands which are used do not correspond to the bands as defined by Zwicker.

$$z/Bark = 7\cdot\sinh^{-1}(\frac{f/Hz}{650}) \qquad (3.12)$$

This formula is an approximation, given in [SCHR79], of the scale which was defined in [ZWI67]. Measuring unit on this scale is Bark[9]. Border and central frequencies in the subbands ($f_l[i]$, $f_c[i]$ and $f_u[i]$) are given in the table 6 in [ITU01]. There are $Z = 55$ of them and the width of each one, except the last subband, is approximately $\Delta z = ¼$ Barks. The grouping is carried out by summation of the energies of spectrum lines in each subband. Central frequency of each line is given with (3.8), and its width is F_s/N_F. If a spectrum line is on a border, then it is distributed between subbands on whose border it lies:

$$E_b[i] = \sum_{k=k_l[i]}^{k_u[i]} U[i,k]\cdot X[k], \qquad 0\le i\le Z \qquad (3.13)$$

$$U[i,k] = \max[0,\min(f_u[i],\frac{2k+1}{2}\frac{F_s}{N_F}) - \max(f_l[i],\frac{2k-1}{2}\frac{F_s}{N_F})]\cdot\frac{N_F}{F_s}, \qquad (3.14)$$

where $k_l[i]$ and $k_u[i]$ are indices of spectrum lines on the lower and upper border of a frequency subband i. If energy in one of the subbands is 0, 10^{-12} is added. The same procedure is used for grouping the spectrum lines of signal difference (3.11):

$$E_{diff}[i] = \sum_{k=k_l[i]}^{k_u[i]} U[i,k]\cdot X_{diff}[k], \qquad 0\le i\le Z \qquad (3.15)$$

This results in ***noise patterns $E_{diff}[i]$*** used later for calculation of *MOV $SNMR_B$*.

3.1.5. Internal noise

Then the internal noise is added, which is the second part of the absolute threshold of hearing modeling process (3.6):

$$E_{INdB}[i]/dB = 0.4\cdot 3.64\cdot\left(\frac{f_c[i]}{kHz}\right)^{-0.8} \qquad (3.16)$$

$$E_{IN}[i] = 10^{\frac{E_{INdB}[i]}{10}} \qquad (3.17)$$

$$E[i] = E_b[i] + E_{IN}[i], \qquad 0\le i\le Z \qquad (3.18)$$

3.1.6. Frequency masking

The energies from each of the subbands are then smeared in order to model frequency masking. Slopes of spreading function are given with the following:

$$\frac{S_l}{dB/Bark} = 27 \qquad (3.19)$$

$$\frac{S_u(i,E)}{dB/Bark} = -24 - \frac{230Hz}{f_c[i]} + 0.2\cdot 10\cdot\log_{10}(E)/dB \qquad (3.20)$$

[9] The unit was named after Heinrich Georg Barkhausen, (1881.-1956.), Dresden, Germany.

They determine how the energies will be smeared over the whole spectrum:

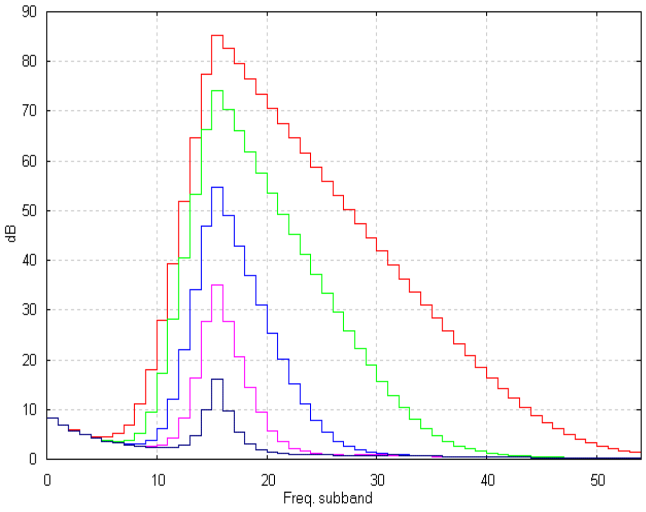
$$S_{dB}(i,l,E)=\begin{cases} S_l\cdot(i-l)\cdot\Delta z, & i\le l\\ S_u(i,E)\cdot(i-l)\cdot\Delta z, & i\ge l\end{cases}\qquad(3.21)$$

$$S(i,l,E)=\frac{1}{A(l,E)}10^{S_{dB}(i,l,E)/10}\qquad(3.22)$$

$$A(l,E)=\sum_{i=0}^{Z-1}10^{S_{dB}(i,l,E)/10}\qquad(3.23)$$

$$E_S[i]=\frac{1}{B_S[i]}\left(\sum_{l=0}^{Z-1}\left(E[l]\cdot S(i,l,E[l])\right)^{0.4}\right)^{\frac{1}{0.4}},\qquad 0\le i\le Z\qquad(3.24)$$

$$B_S[i]=\left(\sum_{l=0}^{Z-1}\left(S(i,l,1)\right)^{0.4}\right)^{\frac{1}{0.4}},\qquad 0\le i\le Z\qquad(3.25)$$



3.4 – Excitation patterns for 1 kHz sine signal

E*$_S$*[i] are ***unsmeared excitation patterns***, which means they were not smeared in time. Picture 3.4 shows excitation patterns for the sine frequency signal of 1 kHz and it clearly demonstrates the upper curve which depends on the signal intensity.

3.1.7. Forward masking

Forward masking is modeled by first order low-pass IIR filter, which smears excitation patterns in time:

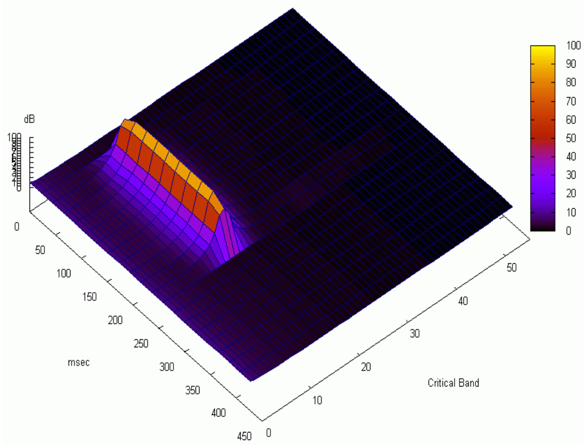
$$\tau[i]=0.008+\frac{100Hz}{f_c[i]}(0.030-0.008),\qquad 0\le i\le Z\qquad(3.26)$$

$$\alpha[i] = e^{-\frac{1}{F_{ss}\cdot\tau[i]}}, \qquad F_{SS} = \frac{F_S}{N_F/2} \qquad (3.27)$$

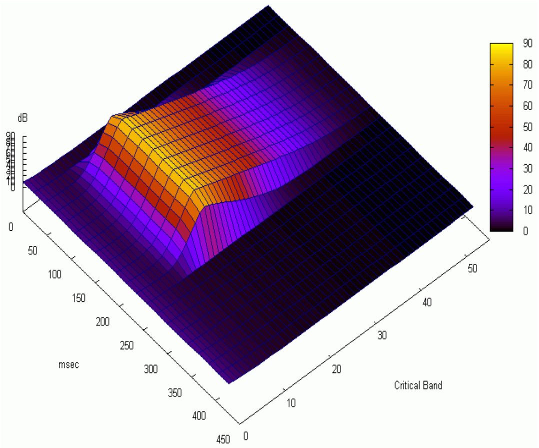
$$E_f[i,n] = \alpha[i]\cdot E_f[i,n-1] + (1-\alpha[i])\cdot E_S[i], \qquad E_f[i,-1] = 0 \qquad (3.28)$$

$$\tilde{E}_S[i,n] = \max(E_f[i,n], E_S[i]), \qquad (3.29)$$

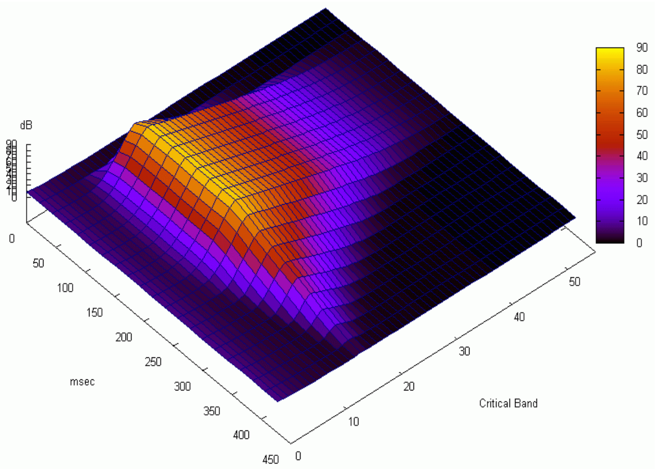
***Excitation patterns* $\tilde{E}_S$[*i*]** are the final products of the perceptual model based on FFT.



Before smearing over frequencies



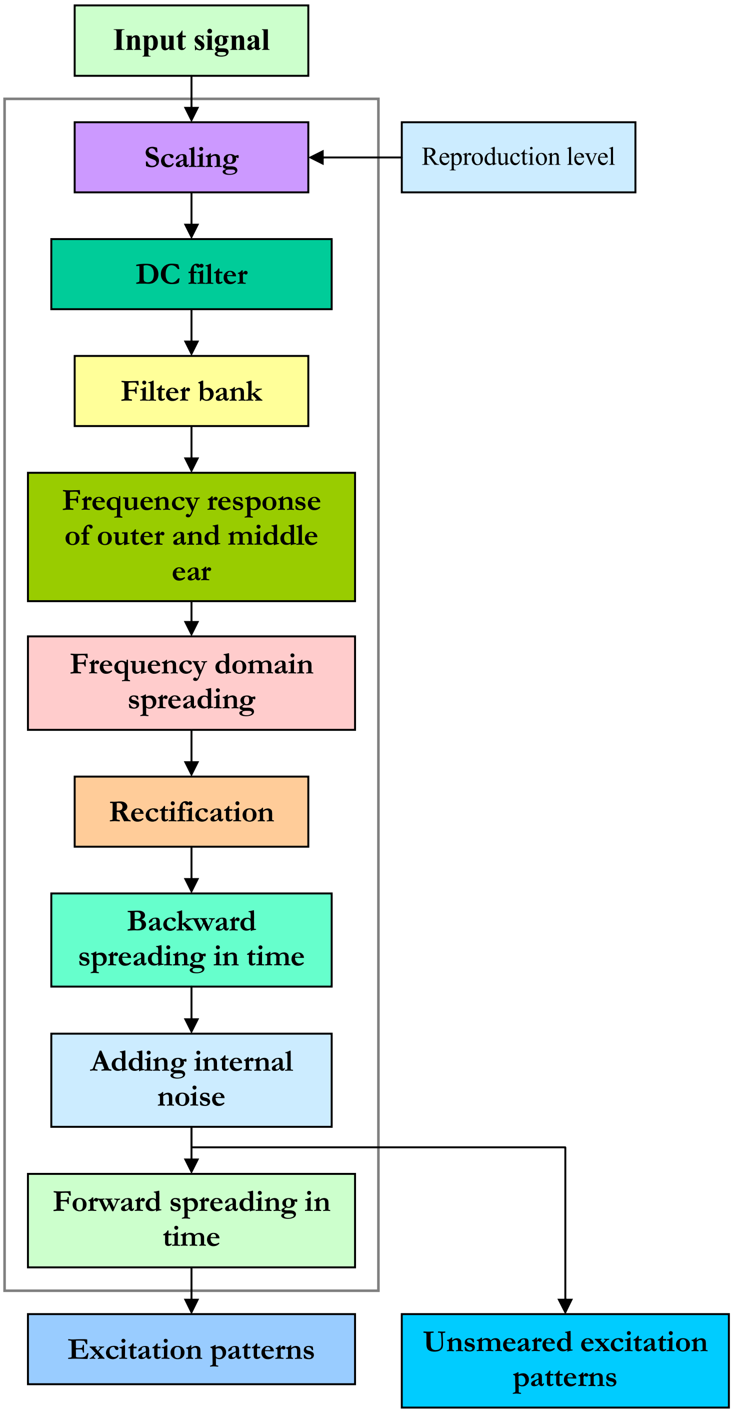
After smearing over frequencies



After smearing in time

3.5 – Excitation patterns in different phases

3.2. Perceptual model based on a filter bank


Input signal
Scaling
Reproduction level
DC filter
Filter bank
Frequency response of outer and middle ear
Frequency domain spreading
Rectification
Backward spreading in time
Adding internal noise
Forward spreading in time
Excitation patterns
Unsmeared excitation patterns

3.6 – Perceptual model based on a filter bank

The basis for this model is the filter bank which consists of Z = 40 pairs of filters, which are equally spaced on the Bark scale, i.e. in approximation of the scale which is used. Thanks to this even arrangement on the Bark scale, there is no explicit grouping into frequency subbands. This model is based on DIX model, first presented in [THI96], and more elaborately presented in [THI99].

3.2.1. Scaling

Scaling factor is, similarly as with the FFT model, calculated with the following formula:

$$fac = \frac{10^{\frac{L_p}{20}}}{A_{\max}} \tag{3.30}$$

$$t_s[n] = fac \cdot t[n] \tag{3.31}$$

Where L_p is the expected playback level of the full scale sine signal, which, if there are no other data, is set on already determined value of 92 dB SPL. $A_{\max}$ is the maximum amplitude of the sine signal (32767 for 16bit input).

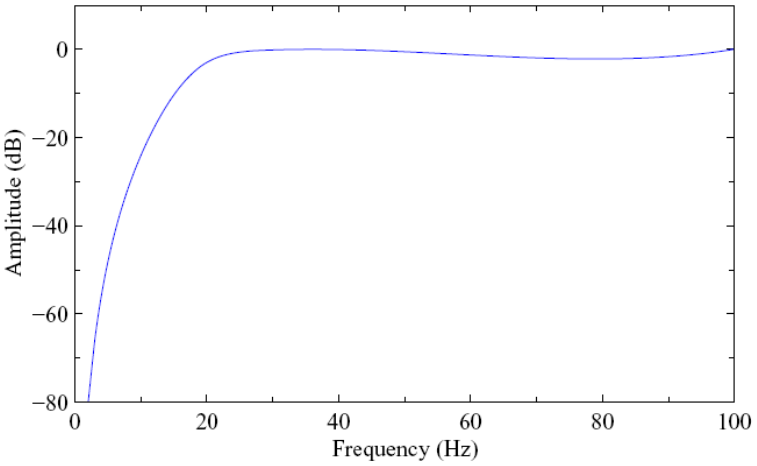
3.2.2. Removal of DC component

Sensitivity of the filter bank to subsonic components requires use of a DC rejection filter. Butterworth fourth order high-pass filter is used, with cutoff frequency of 20 Hz. DC filter is realized by the cascade of 2 second order IIR filters:

$$x_a[n] = a_{01}x_a[n-1] + a_{02}x_a[n-2] + t_s[n] - 2t_s[n-1] + t_s[n-2] \tag{3.32}$$

$$x_{hp}[n] = a_{11}x_{hp}[n-1] + a_{12}x_{hp}[n-2] + x_a[n] - 2x_a[n-1] + x_a[n-2] \tag{3.33}$$

$$a_{01} = 1.99517, a_{02} = -0.995174$$
$$a_{11} = 1.99799, a_{12} = -0.997998$$



3.7 Frequency response of DC filter [KAB06]

3.2.3. Filter bank

Signal then passes through the filter bank:

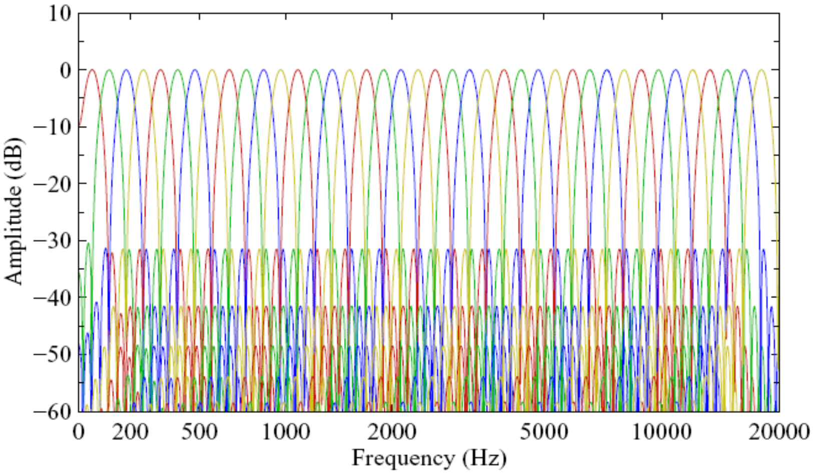
$$h_{re}[i,n] = \frac{4}{N[i]}\sin^2\left(\pi\frac{n}{N[i]}\right)\cos(2\pi\frac{f_c[i]}{F_S}(n-\frac{N[i]}{2})), \quad 0 \le i < Z$$

$$h_{im}[i,n] = \frac{4}{N[i]}\sin^2\left(\pi\frac{n}{N[i]}\right)\sin(2\pi\frac{f_c[i]}{F_S}(n-\frac{N[i]}{2})), \quad 0 \le n < N[i] \quad (3.34)$$

$$\begin{aligned} x_{re}[i,n] &= \sum_{m=0}^{N[i]-1} h_{re}[i,n]x_{hp}[I_S \cdot n - m - D[i]], \\ x_{im}[i,n] &= \sum_{m=0}^{N[i]-1} h_{im}[i,n]x_{hp}[I_S \cdot n - m - D[i]], \end{aligned} \quad \begin{aligned} 0 \le i < Z \\ 0 \le n < N[i] \end{aligned} \quad (3.35)$$

$$D[i] = 1 + \frac{N[0]-N[i]}{2}$$

Where I_S = 32 is a factor of subsampling on the output. $N[i]$ are the lengths of filters, and $f_c[i]$ are central frequencies and are shown in the table 8 in [ITU01]. The filters are complex and have linear phase. In order to have all filters to be of the same phase, the delay of $D[i]$ is applied to each filter.



3.8 Frequency response of the filter bank [KAB06]
Frequency axis is linear on the Bark scale

3.2.4. Absolute threshold of hearing and frequency response

Absolute threshold of hearing is, as in the FFT model, modeled in two phases: the frequency response of the outer and middle ear and by adding the internal noise. The formula for coefficients is the same as in the FFT model (3.9), with one difference being that for *f [k]* central frequencies are used from the table 8 in [ITU01]:

$$x_{wre}[i,n] = W[i] \cdot x_{re}[i,n],$$

$$x_{wim}[i,n] = W[i] \cdot x_{im}[i,n], \qquad 0 \le i < Z \qquad (3.36)$$

3.2.5. Frequency masking

The outputs from the filter bank are then smeared to model the frequency masking. The slopes of the spreading function are given with the following formula:

$$\frac{S_l}{dB/Bark} = 31 \qquad (3.37)$$

$$\frac{S_u(i,E)}{dB/Bark} = \min(-4,-24 - \frac{230Hz}{f_c[i]} + 0.2 \cdot 10 \cdot \log_{10}(E)/dB) \qquad (3.38)$$

They determine how amplitudes will be smeared over the whole spectrum:

$$S_{dB}(i,l,E) = \begin{cases} S_l \cdot (i-l) \cdot \Delta z, & i \le l \\ S_u(i,E) \cdot (i-l) \cdot \Delta z, & i \ge l \end{cases} \qquad \Delta z = 0.70678 \qquad (3.39)$$

$$S(i,l,E) = 10^{S_{dB}(i,l,E)/20} \qquad (3.40)$$

$$\widetilde{S}[i,l,n] = \alpha \cdot \widetilde{S}[i,l,n-1] + (1-\alpha) \cdot S(i,l,E[l,n]), \qquad \alpha = e^{-\frac{I_S}{F_s \cdot 0.1}} \qquad (3.41)$$

$$\begin{aligned} \widetilde{x}_{wre}[i,n] &= \sum_{l=0}^{Z-1} x_{wre}[l,n] \cdot \widetilde{S}[i,l,n], \\ \widetilde{x}_{wim}[i,n] &= \sum_{l=0}^{Z-1} x_{wim}[l,n] \cdot \widetilde{S}[i,l,n], \end{aligned} \qquad 0 \le i < Z \qquad (3.42)$$

where $E[l,n]$ = $(x_{wre}[l,n])^2 + (x_{wim}[l,n])^2$. Since it is almost impossible for the slopes of a spreading function to change very fast based on the intensity, the slopes are therefore smoothed by first order low-pass filter (3.41) with temporal constant of 0.1 s. Considering the fact that smearing in the spectrum is carried out before any kind of non-linear operation (e.g. rectification), the relation between spectrum and temporal characteristics of the filters is preserved. Consequently, the outputs of the filter bank after smearing are identical to outputs that would be directly produced by filters that model exponential slopes that correspond to human hearing.

The energies are then calculated in each bank (rectification):

$$E_0[i,n] = (\widetilde{x}_{wim}[i,n])^2 + (\widetilde{x}_{wre}[i,n])^2, \qquad 0 \le i < Z \qquad (3.43)$$

3.2.6. Backward masking

Thus calculated energies are smeared in time by a low-pass FIR filter, in order to enable modeling of backward masking:

$$E_1[i,n] = \frac{0.9761}{I_{S2}} \sum_{m=0}^{2 \cdot I_{S2}-1} E_0[i, I_{S2} \cdot n - m] \cdot \cos^2\left(\pi \frac{(m - I_{S2} + 1)}{2 \cdot I_{S2}}\right), \qquad 0 \le i < Z \qquad (3.44)$$

Outputs are calculated at every I_{S2} = 6 inputs, so that sampling frequency $E_1[i,n]$ equals $F_S/192$. The length of this filter is 8 ms, which corresponds to the length of backward masking of about 2 ms.

3.2.7. Internal noise

Next, the internal noise is added (the same equations are used as in the FFT model (3.16) with the central frequencies correspondent to the filter bank):

$$E_S[i,n] = E_1[i,n] + E_{IN}[i], \qquad 0 \le i < Z \qquad (3.45)$$

resulting in ***unsmeared excitation patterns E*** ***E_S***.

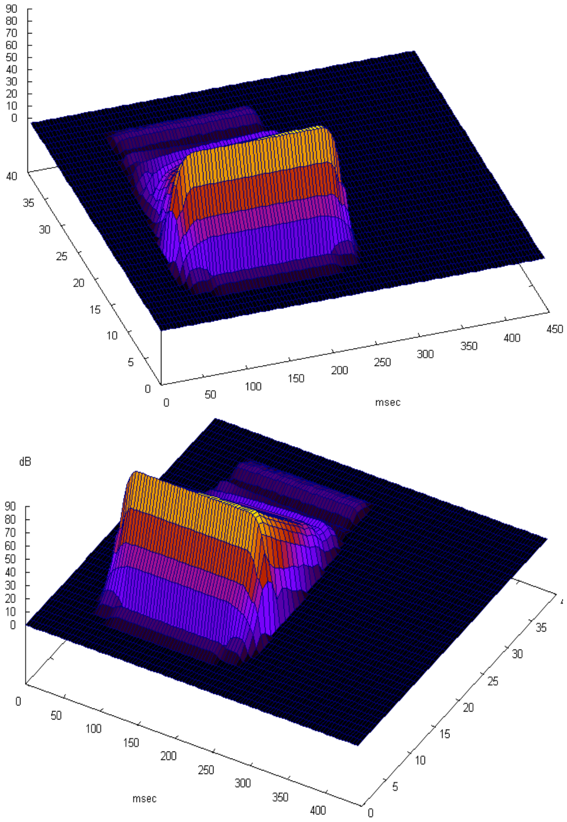
3.2.8. Forward masking

Forward masking is modeled by a first order low-pass IIR filter, which smears *excitation patterns* in time:

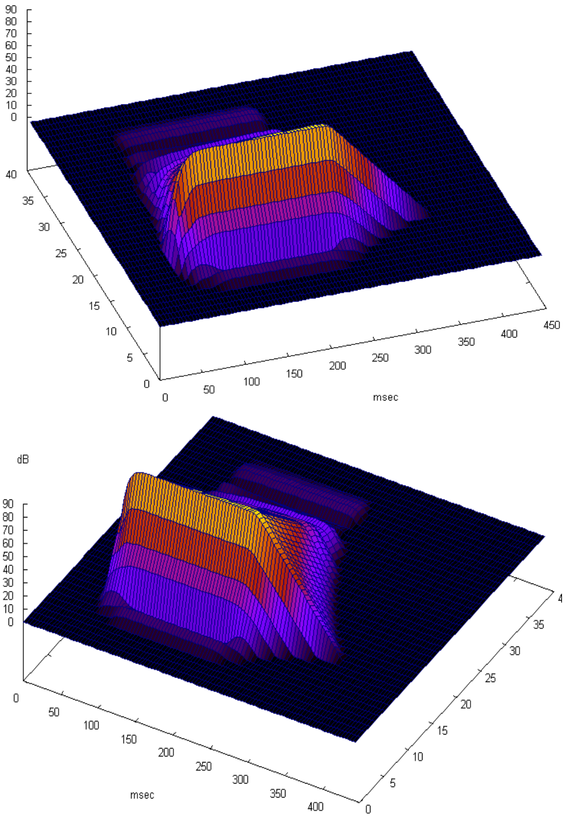
$$\tau[i] = 0.004 + \frac{100Hz}{f_c[i]}(0.020 - 0.004), \qquad 0 \le i < Z \qquad (3.46)$$

$$\alpha[i] = e^{-\frac{1}{F_{ss}\cdot\tau[i]}}, \qquad F_{SS} = \frac{F_S}{192} \qquad (3.47)$$

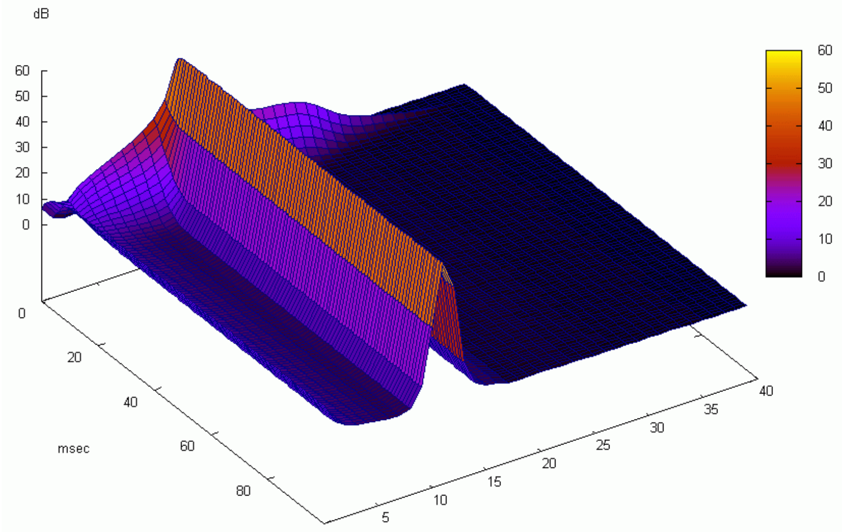
$$\tilde{E}_S[i,n] = \alpha[i]\cdot\tilde{E}_S[i,n-1] + (1-\alpha[i])\cdot E_S[i,n], \qquad \tilde{E}_S[i,-1] = 0 \qquad (3.48)$$



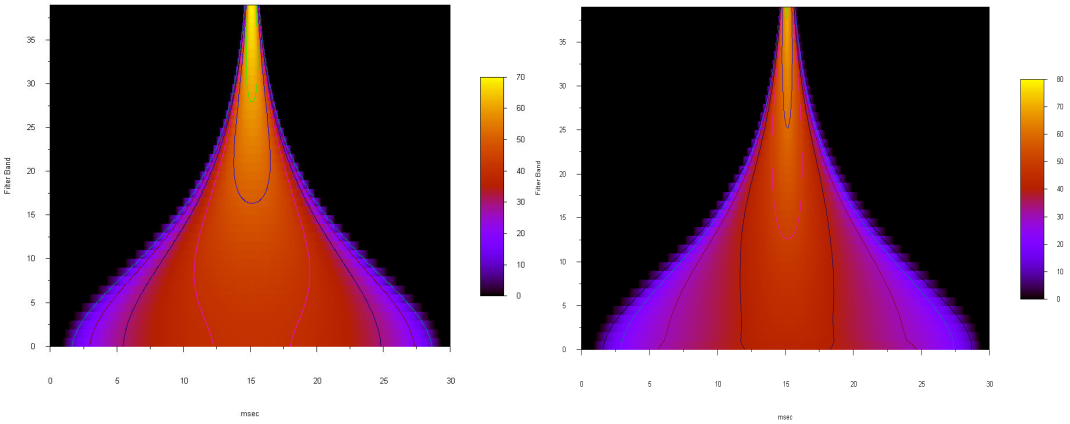
3.9 – Excitation patterns before temporal smearing



3.10 – Excitation patterns after temporal smearing



3.11 Excitation patterns for 1 kHz 60 dB sine signal



3.12 Impulse response of the filter bank before and after frequency smearing

Excitation patterns $\tilde{E}_S$ are the final products of the perceptual model based on the filter bank. Graphs 3.9 to 3.12 demonstrate *excitation patterns* of the sine signal and the impulse response of the filter bank, that were produced through the implementation of this model and which are correspondent to the graphs shown in [THI99].

3.3. Preprocessing of excitation patterns

Preprocessing is, in the advanced version, necessary only for *excitation patterns* from the perceptual model based on the filter bank. It includes adjusting the volume of the original and the tested signals, calculation of temporal envelope modulation as well as calculation of the signal loudness. If not stated otherwise, zero starting conditions are used.

3.3.1. Level adaptation for the reference and the tested signals

First *excitation patterns* ($\tilde{E}_{Sref}$ and $\tilde{E}_{Stest}$) are smeared in time, by applying a first order low-pass filter:

$$\tau[i] = 0.008 + \frac{100Hz}{f_c[i]}(0.050 - 0.008), \qquad 0 \le i < Z \tag{3.49}$$

$$\alpha[i] = e^{-\frac{1}{F_{ss} \cdot \tau[i]}}, \qquad F_{SS} = \frac{F_S}{192} \tag{3.50}$$

$$\begin{aligned} P_{ref}[i,n] &= \alpha[i] \cdot P_{ref}[i,n-1] + (1-\alpha[i]) \cdot \tilde{E}_{S,ref}[i,n] \\ P_{test}[i,n] &= \alpha[i] \cdot P_{test}[i,n-1] + (1-\alpha[i]) \cdot \tilde{E}_{S,test}[i,n] \end{aligned} \tag{3.51}$$

Momentary correction factor is then calculated with the following formula:

$$C_L[n] = \left(\frac{\sum_{i=0}^{Z-1} \sqrt{P_{test}[i,n] \cdot P_{ref}[i,n]}}{\sum_{i=0}^{Z-1} P_{test}[i,n]} \right)^2 \tag{3.52}$$

which is used for level adaptation of the signals:

$$\begin{aligned} E_{L,ref}[i,n] &= \begin{cases} \tilde{E}_{S,ref}[i,n] / C_L[n] & C_L[n] > 1 \\ \tilde{E}_{S,ref}[i,n] & C_L[n] \le 1 \end{cases} \\ E_{L,test}[i,n] &= \begin{cases} \tilde{E}_{S,test}[i,n] & C_L[n] > 1 \\ \tilde{E}_{S,test}[i,n] \cdot C_L[n] & C_L[n] \le 1 \end{cases} \end{aligned} \tag{3.53}$$

3.3.2. Frequency adaptation of patterns

After that, correlation, smoothed over time, between the patterns of the original and the tested signals is calculated ($\alpha[i]$ are used which are defined in (3.50)):

$$\begin{aligned} R_n[i,n] &= \alpha[i] \cdot R_n[i,n-1] + E_{L,test}[i,n] \cdot E_{L,ref}[i,n] \\ R_d[i,n] &= \alpha[i] \cdot R_d[i,n-1] + E_{L,ref}[i,n] \cdot E_{L,ref}[i,n] \end{aligned} \tag{3.54}$$

being further used in calculation of correction factors:

$$\begin{aligned} R_{ref}[i,n] &= \begin{cases} 1 & R_n[i,n] \ge R_d[i,n] \\ R_n[i,n] / R_d[i,n] & R_n[i,n] < R_d[i,n] \end{cases} \\ R_{test}[i,n] &= \begin{cases} R_d[i,n] / R_n[i,n] & R_n[i,n] \ge R_d[i,n] \\ 1 & R_n[i,n] < R_d[i,n] \end{cases} \end{aligned} \tag{3.55}$$

If $R_d[i,n] = 0$ and $R_n[i,n] > 0$, then it is $R_{ref}[i,n] = 1$ and $R_{test}[i,n] = 0$. If $R_n[i,n] = 0$, then it is $R_{ref}[i,n] = R_{ref}[i\text{-}1,n]$ and $R_{test}[i,n] = R_{test}[i\text{-}1,n]$. If $i = 0$, then $R_{ref}[i,n] = R_{test}[i,n] = 1$.

Correction factors are averaged between three adjacent subbands and are smoothed over time (*α*[*i*] defined in (3.50) are used):

$$\begin{aligned} PC_{ref}[i,n] &= \alpha[i]\cdot PC_{ref}[i,n-1]+(1-\alpha[i])\cdot\frac{1}{M_1+M_2+1}\sum_{k=-M_1[i]}^{M_2[i]}R_{ref}[i+k,n] \\ PC_{test}[i,n] &= \alpha[i]\cdot PC_{test}[i,n-1]+(1-\alpha[i])\cdot\frac{1}{M_1+M_2+1}\sum_{k=-M_1[i]}^{M_2[i]}R_{test}[i+k,n] \end{aligned} \qquad (3.56)$$

$$M_1[i]=\min(1,i),\quad M_2[i]=\min(1,Z-i-1) \qquad (3.57)$$

In ITU-R BS.1387 it is not specified how is *PC_{ref}* supposed to be initialized, which can be interpreted as though they should be initialized with 0, but [KAB06] and [BAU01] express certain doubts about the issue. Initial variables do not have significant influence on the final MOVs, and logical initialization with 1, as proposed in [KAB06], gives accurate values if identical signals are compared, opposed to the initialization with 0.

Correction coefficients are then applied for spectrum adaptation of patterns:

$$\begin{aligned} E_{P,ref}[i,n] &= PC_{ref}[i,n]\cdot E_{L,ref}[i,n] \\ E_{P,test}[i,n] &= PC_{test}[i,n]\cdot E_{L,test}[i,n] \end{aligned} \qquad 0\le i<Z \qquad (3.58)$$

3.3.3. Calculation of modulation

Average loudness is calculated based on the unsmeared excitation patterns (3.45) with the following formula:

$$\begin{aligned} \overline{E}_{ref}[i,n] &= \alpha[i]\cdot \overline{E}_{ref}[i,n-1]+(1-\alpha[i])\cdot\left(E_{S,ref}[i,n]\right)^{0.3} \\ \overline{E}_{test}[i,n] &= \alpha[i]\cdot \overline{E}_{test}[i,n-1]+(1-\alpha[i])\cdot\left(E_{S,test}[i,n]\right)^{0.3} \end{aligned} \qquad 0\le i<Z \qquad (3.59)$$

Average loudness difference is (F_{SS} and *α*[*i*] are used, which are defined in (3.50)):

$$\begin{aligned} \overline{D}_{ref}[i,n] &= \alpha[i]\cdot \overline{D}_{ref}[i,n-1]+(1-\alpha[i])\cdot F_{SS}\left|\left(E_{S,ref}[i,n]\right)^{0.3}-\left(E_{S,ref}[i,n-1]\right)^{0.3}\right| \\ \overline{D}_{test}[i,n] &= \alpha[i]\cdot \overline{D}_{test}[i,n-1]+(1-\alpha[i])\cdot F_{SS}\left|\left(E_{S,test}[i,n]\right)^{0.3}-\left(E_{S,test}[i,n-1]\right)^{0.3}\right| \end{aligned} \qquad (3.60)$$

Based on these values *M*[*i,n*] - modulation of temporal envelopes is calculated:

$$\begin{aligned} M_{ref}[i,n] &= \frac{\overline{D}_{ref}[i,n]}{1+\overline{E}_{ref}[i,n]/0.3} \\ M_{test}[i,n] &= \frac{\overline{D}_{test}[i,n]}{1+\overline{E}_{test}[i,n]/0.3} \end{aligned} \qquad 0\le i<Z \qquad (3.61)$$

3.3.4. Calculation of total loudness

Specific loudness patterns are determined with the formula proposed in [ZWI67]:

$$N[i,n]=1.26539\cdot\left(\frac{E_{Thres}[i]}{s[i]\cdot 10^4}\right)^{0.23}\left[\left(1-s[i]+\frac{s[i]\cdot\tilde{E}_S[i,n]}{E_{Thres}[i]}\right)^{0.23}-1\right] \qquad (3.62)$$

where $E_{Thres}[i]$ and *s*[*k*] are defined with:

$$E_{Thres}[i] = 10^{\frac{3.64}{10}\left(\frac{f_c[i]}{kHz}\right)^{-0.8}} \tag{3.63}$$

$$s[i] = 10^{\left(-2-2.05\cdot\tan^{-1}(f/4kHz)-0.75\tan^{-1}\left(f^2/2.56kHz\right)\right)/10} \tag{3.64}$$

Total loudness is a sum of the specific loudness patterns:

$$N_{tot}[n] = \frac{24}{Z}\sum_{i=0}^{Z-1}\max(N[i,n],0) \tag{3.65}$$

3.4. Calculation of Model Output Variables

In the advanced version described in ITU-R BS.1387 five *MOVs* are used: *RmsModDiff*$_A$, *RmsNoiseLoudAsym*$_A$, *AvgLinDist*$_A$, *SNMR*$_B$ and *EHS*$_B$. For averaging in time, of *MOV* values in single frames, the values from all the frames, that fulfill the required conditions described in chapter 3.5, are used. Variable *n* represents the ordinal number of a frame. Frame number 0 represents the first frame, which fulfills the required conditions for the selection of frames, and frame *N*-1 is the last one.

After averaging in time, the final value of each MOV is arithmetic mean between the channels (left and right or mono):

$$MOV = \frac{1}{Num\_of\_Channels}\sum_{iChn=0}^{Num\_of\_Channels} MOV_{chn}[iChn] \tag{3.66}$$

3.4.1. RmsModDiff$_A$

RmsModDiff$_A$ represents the difference of modulation patterns of the original and the tested signal (3.61) Modulation difference value in one temporal frame is a sum of modulation differences in all the frequency subbands:

$$ModDiff[i,n] = \frac{\left|M_{test}[i,n] - M_{ref}[i,n]\right|}{1 + M_{ref}[i,n]} \tag{3.67}$$

$$ModDiff_{tot}[n] = \frac{100}{Z}\sum_{i=0}^{Z-1} ModDiff[i,n] \tag{3.68}$$

For this variable, weight coefficients are used for averaging in time ($E_{IN}[i]$ are the same as in (3.45)):

$$Wt[n] = \sum_{i=0}^{Z-1}\frac{\overline{E}_{ref}[i,n]}{\overline{E}_{ref}[i,n] + \left(E_{IN}[i]\right)^{0.3}} \tag{3.69}$$

$$RmsModDiff_A = \sqrt{Z\cdot\frac{\sum_{n=0}^{N-1}\left(Wt[n]\cdot ModDiff_{tot}[n]\right)^2}{\sum_{n=0}^{N-1}\left(Wt[n]\right)^2}} \tag{3.70}$$

3.4.2. Loudness of distortions

Partial loudness of distortions is calculated with the following formula ($E_{IN}[i]$ are the same as in (3.45)):

$$N_L[i,n] = \left(\frac{E_{IN}[i]}{s_{test}[i,n]}\right)^{0.23}\left[\left(1-\frac{\max\left(s_{test}[i,n]\cdot E_{test}[i,n]-s_{ref}[i,n]\cdot E_{ref}[i,n],0\right)}{E_{IN}[i]+\beta[i,n]\cdot s_{ref}[i,n]\cdot E_{ref}[i,n]}\right)^{0.23}-1\right] \quad (3.71)$$

$$s[i,n] = T_0 M[i,n]+1 \quad (3.72)$$

$$\beta[i,n] = e^{\alpha\left(E_{ref}[i,n]-E_{test}[i,n]\right)/E_{ref}[i,n]} \quad (3.73)$$

where $M[i,n]$ are given with the formula (3.61). The value in a single time frame is defined by:

$$N_{LM}[n] = \max\left(N_{L,\min}, \frac{24}{Z}\sum_{i=0}^{Z-1} N_L[i,n]\right) \quad (3.74)$$

MOV	α	T_0	$N_{L,min}$
$RmsNoiseLoud_A$	2.5	0.3	0.1
$RmsMissingComponents_A$	1.5	0.15	0
$AvgLinDist_A$	1.5	0.15	0

Table 3.1 – Constants for loudness of distortions

3.4.3. RmsNoiseLoudAsym$_A$

RmsNoiseLoudAsym$_A$ represents linear combination *RmsNoiseLoud$_A$* and *RmsMissingComponents$_A$*:

$$RmsNoiseLoudAsym_A = RmsNoiseLoud_A + 0.5 RmsMissingComponents_A \quad (3.75)$$

RmsNoiseLoud$_A$ and *RmsMissingComponents$_A$* are root mean square averages of $N_{LM}[n]$ in time:

$$N_{LRMS} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}\left(N_{LM}[n]\right)^2} \quad (3.76)$$

$$N_{LRMS} \equiv RmsNoiseLoud_A \mid RmsMissingComponents_A$$

For calculating *RmsNoiseLoud$_A$* spectrum adapted patterns are used (3.58), $E_{P,ref}$ in place of E_{ref} and $E_{P,test}$ in place of E_{test}. *RmsNoiseLoud$_A$* represents loudness of distortions in the tested signal, which do not exist in the original signal. Constants that are used in the formulas (3.71-3.74) are given in the table 3.1.

RmsMissingComponents$_A$ represents loudness of components in the original signal, which do not exist in the tested one. Reference and tested signals change their places in the formulas (3.71-3.74), so $E_{P,ref}$ is used in place of E_{test} and $E_{P,test}$ in place of E_{ref}.

3.4.4. AvgLinDist$_A$

AvgLinDist$_A$ defines loudness of signal components that disappeared during spectrum adaptation of the *excitation patterns*. *AvgLinDist$_A$* represents the measure for linear distortions [THI99]. $E_{P,ref}$ is used in place of E_{ref} and $\tilde{E}_{S,ref}$ (3.48) in place of E_{test}. Linear averaging in time is used:

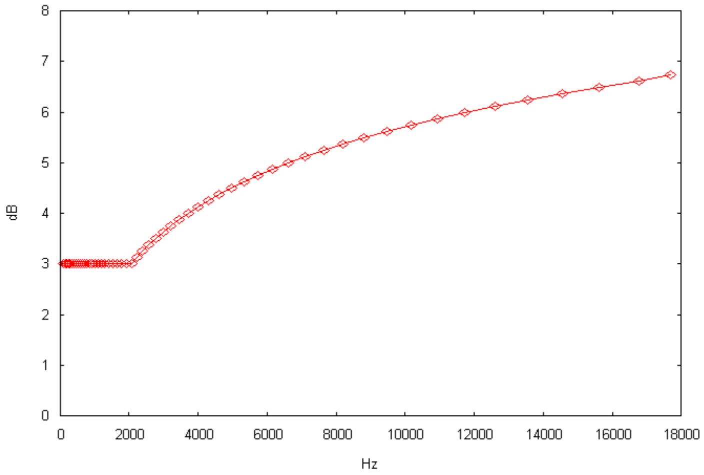
$$AvgLinDist_A = \frac{1}{N}\sum_{n=0}^{N-1} N_{LM}[n] \qquad (3.77)$$

3.4.5. SNMR$_B$

SNMR$_B$ - Segmental Noise-to-Mask Ratio is described in [BRA87] and represents relation between noise and masking threshold. Masking threshold M_T is calculated based on the *smeared excitation patterns* of the reference signal from the FFT model (3.48):

$$M_T[i,n] = \widetilde{E}_{S,ref}[i,n] \cdot 10^{-m_{dB}[i]/10}, \qquad 0 \le i < Z \qquad (3.78)$$

$$m_{dB}[i] = \begin{cases} 3 & i \le \dfrac{12}{\Delta z} \\ i \cdot \dfrac{\Delta z}{4} & i > \dfrac{12}{\Delta z} \end{cases} \qquad 0 \le i < Z \qquad (3.79)$$



3.11 Masking offset $m_{dB}[i]$ (3.79),
Markers define central frequencies of the frequency subbands

NMR in a time frame n is given by:

$$NMR_{loc}[n] = 10 \cdot \log_{10}\left(\frac{1}{Z}\sum_{i=0}^{Z-1}\frac{E_{diff}[i,n]}{M_T[i,n]}\right) \qquad (3.80)$$

E_{diff} are *noise patterns* defined in (3.15).
For the final *SNMR$_B$* linear averaging in time is used:

$$SNMR_B = \frac{1}{N}\sum_{n=0}^{N-1} NMR_{loc}[n] \tag{3.81}$$

3.4.6. EHS$_B$

EHS$_B$ - Error Harmonic Structure comes from [PAI92] and represents harmonic structure of difference between the original and the tested signal. In [ITU01] *EHS$_B$* is not clearly defined compared to the description in, [KAB06]. First, the difference between logarithm of spectra, with applied frequency response of the outer ear, is defined (3.10):

$$D[i] = \log\left(X_{test}[i]\right) - \log\left(X_{ref}[i]\right) = \log\left(\frac{X_{test}[i]}{X_{ref}[i]}\right) \tag{3.82}$$

Based on $D[i]$ normalized autocorrelation is defined:

$$C[i] = \frac{D_i \cdot D_0^T}{\sqrt{|D_i|^2 \cdot |D_0|^2}}, \qquad 1 \le i \le L_{\max}, \quad L_{\max} = 256 \tag{3.83}$$

$$D_i = [D[i], \ldots, D[i + L_{\max} - 1]], \qquad D_i \cdot D_0^T = \sum_{k=0}^{L_{\max}-1} D[k] \cdot D[i+k] \tag{3.84}$$

$|D_i|^2$ can be efficiently calculated using recursion [KAB06]:

$$|D_i|^2 = \begin{cases} \sum_{k=0}^{L-1} (D[k])^2, & i = 0 \\ |D_{i-1}|^2 + (D[i + L_{\max} - 1])^2 - (D[i-1])^2, & 1 \le i \le L_{\max} \end{cases} \tag{3.85}$$

Resulting correlation vector $C[i]$ after removal of a DC component, is windowed by a normalized Hann window and its spectrum $EH[k]$ is calculated using FFT:

$$\overline{C} = \frac{1}{L_{\max}} \sum_{i=1}^{L_{\max}} C[i] \tag{3.86}$$

$$h_w[i] = \frac{1}{2}\sqrt{\frac{8}{3}}\left[1 - \cos(2\pi \frac{i}{L_{\max} - 1})\right], \qquad 0 \le i < L_{\max} \tag{3.87}$$

$$C_w[i] = h_w[i]\left(C[i+1] - \overline{C}\right) \tag{3.88}$$

$$EH[k] = \left| \frac{1}{L_{\max}} \sum_{i=0}^{L_{\max}-1} C_w[i] e^{-j \cdot 2 \cdot \pi \cdot k \cdot i / L_{\max}} \right|^2, \qquad 0 \le k < L_{\max} \tag{3.89}$$

The highest peak value *EH$_{max}$* in spectrum after the first valley determines the dominant frequency. Linear averaging in time *EH$_{max}$* gives *EHS$_B$*:

$$EHS_B = \frac{1000}{N} \sum_{n=0}^{N-1} EH_{\max}[n] \tag{3. 90}$$

3.5. Selection of frames

ITU-R BS.1387 is not very clear about the selection of frames which play a role in averaging *MOV* in time. Additional explanations found in [THI99], together with the analysis in [KAB06] and the experimenting with various possibilities, made the implementation of ITU-R BS.1387 feasible in a way that it produced results very similar to the reference ones.

FFT perceptual model processes the inputs in the frames of $N_F/2 = 1024$ samples. Filter bank model produces, for each 192 samples at the input, one value for *MOVs* which use that model, so the length of a frame of the filter bank model is 192. The most convenient way to process inputs is in frames of $3 \cdot 1024 = 16 \cdot 192 = 3072$ samples.

During the experiments of the subjective evaluation of audio quality, it was concluded that distortions and noises at the beginning and the end of the tested audio material do not have significant impact; therefore the criteria for exclusion of the momentary MOV values at the starting and ending frames are introduced in ITU-R BS.1387.

MOV	Applied criteria
$RmsModDiff_A$	Delayed averaging
$RmsNoiseLoudAsym_A$	Delayed averaging, Loudness threshold
$AvgLinDist_A$	Delayed averaging, Loudness threshold
$SNMR_B$	Data boundary
EHS_B	Data boundary, Energy threshold

Table 3.2 – Applied criteria for the selection of frames

Delayed averaging is used on all *MOVs* from the filter bank model ($RmsModDiff_A$, $RmsNoiseLoudAsym_A$ and $AvgLinDist_A$). The first 0.5 seconds, which is 125 blocks (i.e. $125 \cdot 192 = 24000$ samples) for $F_S = 48$ kHz, are ignored for the averaging in time.

Loudness threshold is used on the *MOVs* which evaluate loudness distortion ($RmsNoiseLoudAsym_A$, $AvgLinDist_A$). All frames, before total loudness (3.65) in the left or the right channel of both signals (reference and test) reached 0.1, and 50 ms after reaching this loudness, are ignored for the averaging in time. 50 ms encompasses 13 frames.

For the variables from the filter bank model none of the criteria is applied for the selection of the final frame, so the final frame is the one that corresponds with the last frame of the input signals.

Data boundary is used on *MOVs* from the FFT model ($SNMR_B$ and EHS_B). The beginning of data is the first frame in which five consecutive samples exist, whose sum is bigger than 200 (for 16 bits range) and in the same way the end of data is the last frame, in which there are five consecutive samples, whose sum is bigger than 200. The frames before the beginning and after the end of the data, that is, before and after the frames which fulfill the criterion, are ignored during the averaging in time.

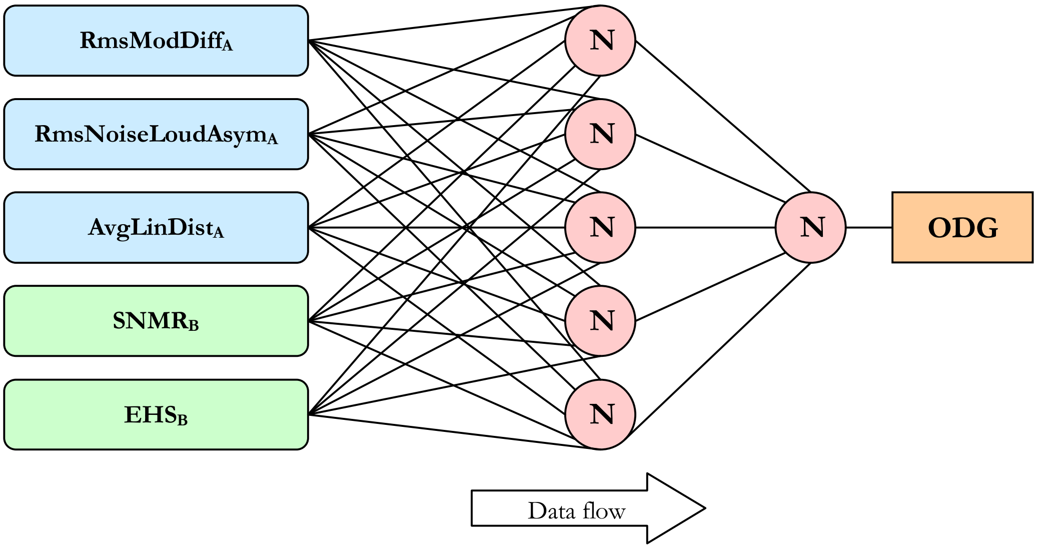
Energy threshold is a criterion, which when used avoids calculation of EHS_B in the frames with a very low energy. All the frames that fulfill the criterion are ignored:

$$\sum_{k=N_F/2}^{N_F-1} t_n[k] < 8000 \tag{3.91}$$

($t_n[k]$ are defined in (3.1)). It is required that criterion is fulfilled for both the reference and the tested signals in all channels (left and right or mono), and if at least one channel in one of the signals does not fulfill the criterion, EHS_B is calculated in that frame.

3.6. Neural network

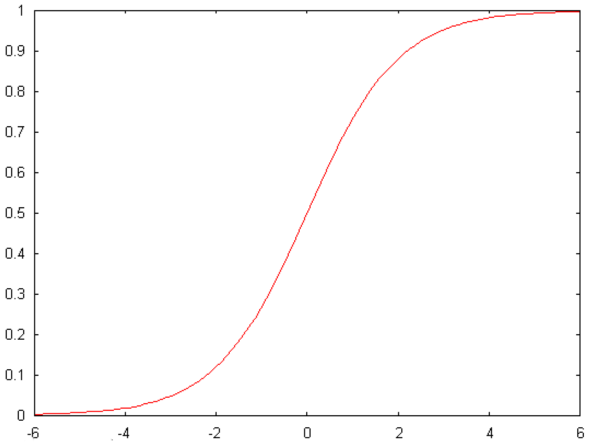
The values of the five *MOVs* are mapped by artificial neural network to *ODG - Objective Difference Grade* which represents the evaluation of the basic audio quality.



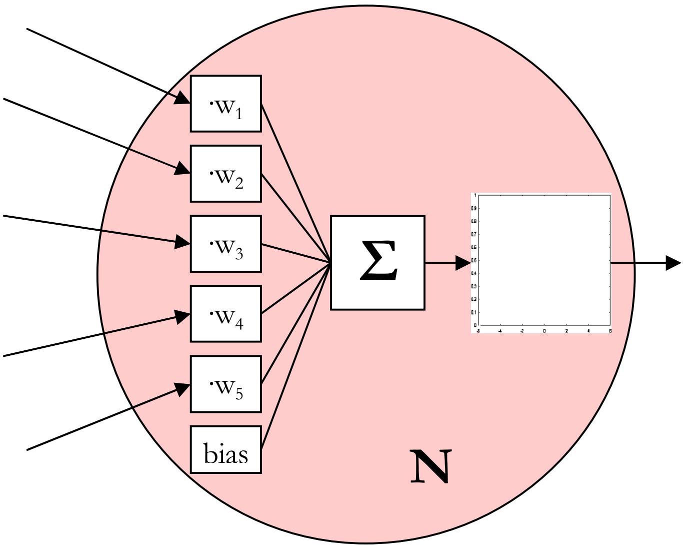
3.12 – Structure of the applied neural network

Neural network has five inputs and one hidden layer with five nodes in it. Activation function is a standard logistic function, which is an asymmetrical sigmoid:

$$sig(x) = \frac{1}{1+e^{-x}} \tag{3.92}$$



3.13 – Graph of the activating function (3.92)



3.14 – Structure of the nodes in a neural network

Non-linear activation function in the nodes of the hidden and output layers is required so that a neural network is capable of modeling a non-linear function.

First, the *MOVs* are scaled so that their values are within the interval [0,1]:

$$MOV'[i] = \frac{MOV[i] - a_{\min}[i]}{a_{\max}[i] - a_{\min}[i]} \qquad 0 \le i \le 4 \qquad (3.93)$$

The values $a_{\min}[i]$ and $a_{\max}[i]$, as well as the numbers of input nodes, which correspond to each of the *MOVs,* are given in the table 18 in [ITU01].

Scaled *MOVs* are then mapped by the neural network to ***Distortion Index - DI***:

$$DI = w_y[5] + \sum_{j=0}^{4}\left(w_y[j] \cdot sig\left(w_x[5,j] + \sum_{i=0}^{4} w_x[i,j] \cdot MOV'[i] \right)\right) \qquad (3.94)$$

based on which *objective difference grade* - *ODG* is defined:

$$ODG = b_{\min} + (b_{\max} - b_{\min}) \cdot sig(DI) \qquad (3.95)$$

Strictly speaking *sig*(*DI*) is the output which is produced by the neural network, being then mapped to ODG by the linear transformation. *Distortion index DI* was introduced as a measure independent of the five-grade impairment scale defined in [ITU97]. If the scale changes, it is enough to change monotonic mapping from *DI* to *ODG,* which is defined with (3.95).

Weight coefficients ($w_x[i,j]$, $w_y[j]$) and scaling coefficients of *ODG* (b_{min} and b_{max}) are shown in the tables 19-21 in [ITU01]. $w_x[5,j]$ and $w_y[5]$ are presented as *bias* in the nodes of the neural network.

3.7. Verification of the implementation

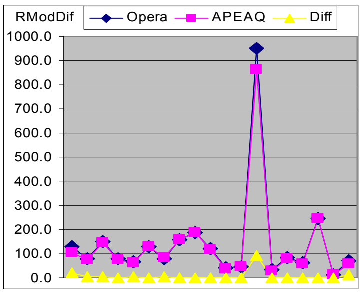
In order to consider the implementation as successful, ITU-R BS.1387 specifies 16 samples for testing and allowed error for *ODG* within +/-0.02. *ODG* and *DI* values are given in the table 23 in [ITU01]. It is expected that the relevant implementation *Opticom Opera* [OPT06] produces identical *ODG* values, but it is not the case here. However, those values are within the specified range:

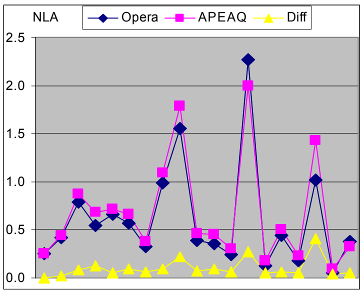
Name of the test		ODG BS.1387-1	ODG Opera	Abs. Difference	APEAQ Unmodif.	Abs. Difference	Test APEAQ	Abs. Difference
1	acodsna	-0.467	-0.464	0.003	-0.465	0.002	-0.459	0.008
2	bcodtri	-0.281	-0.283	0.002	-0.278	0.003	-0.278	0.003
3	ccodsax	-1.3	-1.298	0.002	-1.360	0.060	-1.325	0.025
4	dcodryc	NA	-0.415	NA	-0.479	0.064	-0.442	0.027
5	ecodsmg	-0.489	-0.490	0.001	-0.500	0.011	-0.475	0.014
6	fcodsb1	-0.877	-0.877	0.000	-1.047	0.170	-1.023	0.146
7	fcodtr1	-0.512	-0.516	0.004	-0.569	0.057	-0.538	0.026
8	fcodtr2	-1.711	-1.717	0.006	-1.801	0.090	-1.717	0.006
9	fcodtr3	-2.662	-2.662	0.000	-2.181	0.481	-2.618	0.044
10	gcodcla	-0.573	-0.575	0.002	-0.620	0.047	-0.616	0.043
11	hcodryc	NA	-0.126	NA	-0.144	0.018	-0.121	0.005
12	hcodstr	NA	-0.187	NA	-0.224	0.037	-0.207	0.020
13	icodsna	-3.664	-3.668	0.004	-3.631	0.033	-3.658	0.006
14	kcodsme	-0.029	-0.029	0.000	-0.026	0.003	-0.021	0.008
15	lcodhrp	-0.523	-0.524	0.001	-0.644	0.121	-0.585	0.062
16	lcodpip	-0.219	-0.220	0.001	-0.239	0.020	-0.238	0.019
17	mcodcla	-1.435	-1.438	0.003	-2.047	0.612	-1.424	0.011
18	ncodsfe	0.05	0.050	0.000	0.042	0.008	0.052	0.002
19	scodclv	-0.293	-0.297	0.004	-0.238	0.055	-0.221	0.072

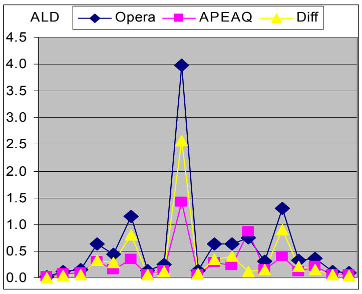
Table 3.3 – Deviation from the reference values

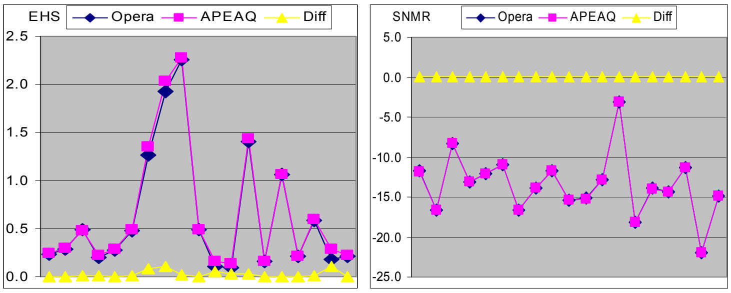
The initial implementation called ***APEAQ*** - *Advanced PEAQ*, implemented according to the recommendation *(APEAQ unmodif.* in the table 3.3) does not produce expected deviations. During the testing of different possibilities I managed to implement a test version, which produced fewer deviations. Criteria for the frame selection were changed, the spectrum without frequency weighting is input for EHS_B, root mean square in time is used for $AvgLinDist_A$, and linear average is used for $RmsMissingComponents_A$. The values for this version are in the column of the table 3.3 marked with *Test APEAQ*.

MOVs for the given examples are not shown in ITU-R BS.1387, but Opera values are available for comparison (19 examples, compared to 16 shown in the standard):

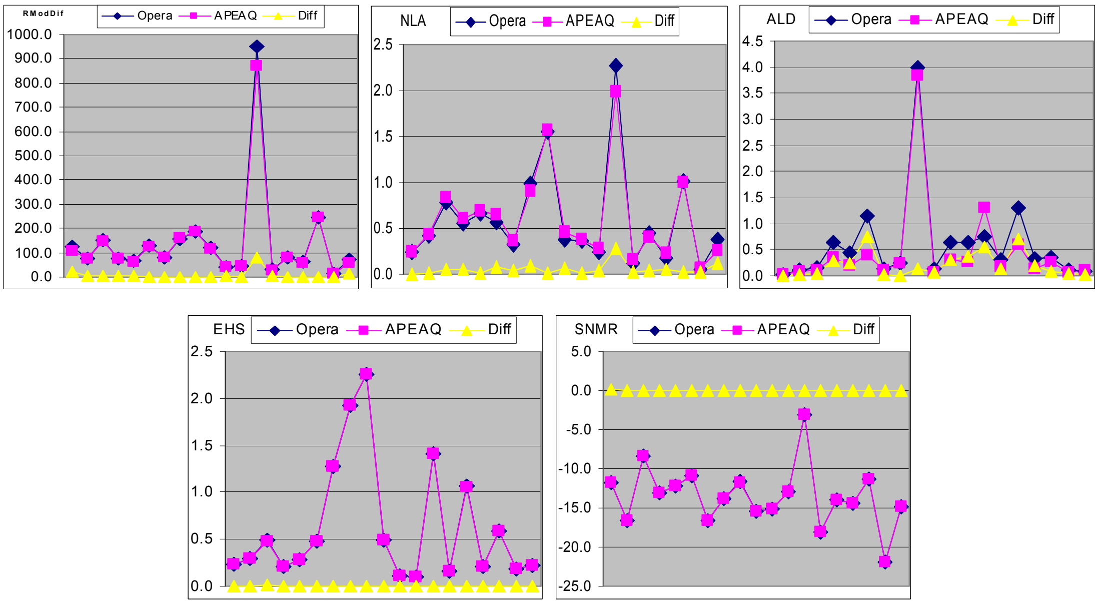








3.15 – Comparison of Opera MOV and unmodified APEAQ



3.16 – Comparison of Opera MOV and Test APEAQ

However, although changes produce the values which are more similar to the reference, there are no relevant reasons to use them. Having compared the results to three known implementations, whose outputs were provided by their authors (publicly accessible implementation does not exist), it was noticed that their *MOVs* and *ODG*s were practically the same as mine. Comparing of values for *excitation patterns* of 1 kHz sine signal also showed great similarities, where differences are only the result of variously implemented algorithms and the errors are within rounding accuracy.

As for further checking, two more *MOVs* were implemented from the PEAQ basic version. The results were the same, as far as the precision of algorithm and the errors within rounding accuracy are concerned, as in [LER02] and [KAB04]. The produced values were again different from the values produced by basic version of Opera.

Neural network can be tested if we use the *MOVs* produced by *Opera*. The testing showed that the neural network was properly implemented.

FFT model was obviously implemented according to the standard. Unfortunately, there are no available intersteps' values of the reference implementation, so it is impossible to determine which part of the filter bank model makes a difference. The filter for removing

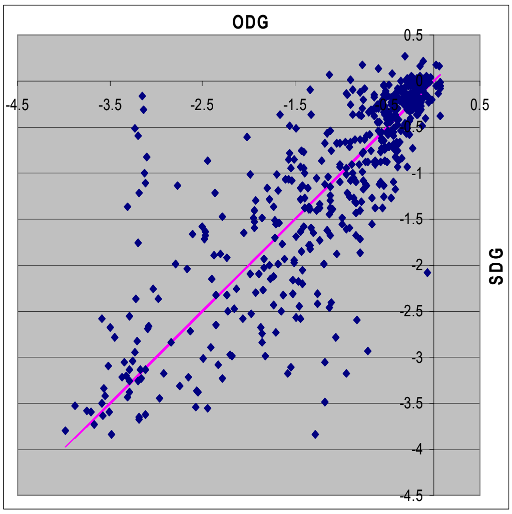
DC component was checked by comparing it with the outputs of available applications which can generate the same filter, and it turned out they were identical.

Based on the *MOV* values we can conclude that the difference occurs in one of the three parts: in the processing of excitation patterns, in the selection of frames for averaging in time and in the calculation of *RmsNoiseLoudAsym*$_A$ and *AvgLinDist*$_A$.

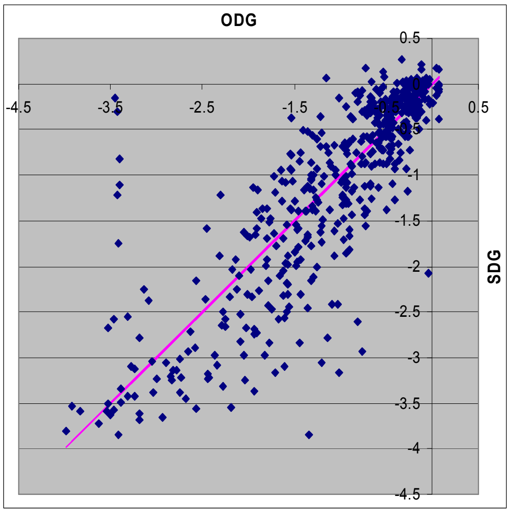
In [KAB06] it is stated that there are certain parts in the proposal which are not clear enough or those which contain some mistakes. None of the possible ways of implementation of the parts in question will produce significant differences in the final values of MOVs; neither will they produce the same results as those of the Opera.

After long and thorough attempts to implement the version which would produce results according to BS.1387, it raised some doubts whether the reference implementation is appropriate and to what extent it corresponds with the proposal ITU-R BS.1387.

APEAQ produces to some extent even better results than Opera. This is shown on scatter plots of *ODG* with *SDG*:



3.17 – The scatter plot of Opera



3.18 – The scatter plot of APEAQ (unmodif.)

as well as by observing the numerical values which define the quality of implementation:

	Opera	APEAQ
Correlation	0.813	0.835
% outside confidence interval	49.0	48.6
AES	2.501	2.277
Mean absolute error	0.417	0.390
Root mean square error	0.625	0.578

Table 3.4 – Comparison of the quality of Opera and unmodified APEAQ

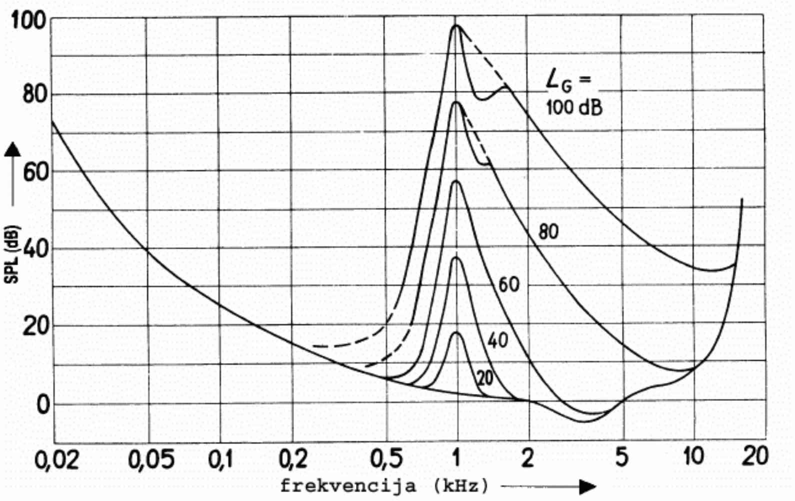
In this table are the results of the unmodified *APEAQ,* which was implemented according to ITU-R BS.1387. Results of the unmodified *APEAQ* will be presented in chapter 6.2, where the basis for the modified version is the version implemented according to ITU-R BS.1387, not the above mentioned test version.

4. Analysis of the implemented psychoacoustic models

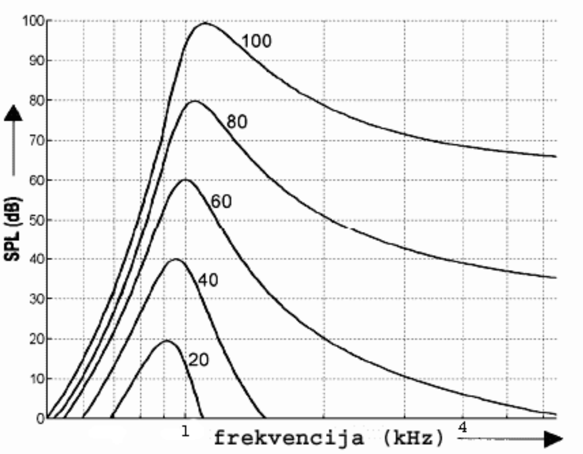
In this chapter the models and methods, which are a part of ITU-R BS.1387, will be analyzed, as well as the possibility for their improvement. The result of the analysis of the models and methods described in this chapter is ***modified APEAQ***.

4.1. Playback level

One of the distinct problems is the unknown playback level. Amount of masking in the frequency domain (3.20 and 3.38) depends on the masker's intensity and represents important element of a psychoacoustic model. This dependency also exists in the FFT model (graph 3.4) and in the filter bank model. Playback level is also important in determination of components that are above hearing threshold.



4.1 – Dependency of masking threshold on amplitude [THI99]



4.2 – Dependency of excitation on basal membrane on amplitude [ROB02]

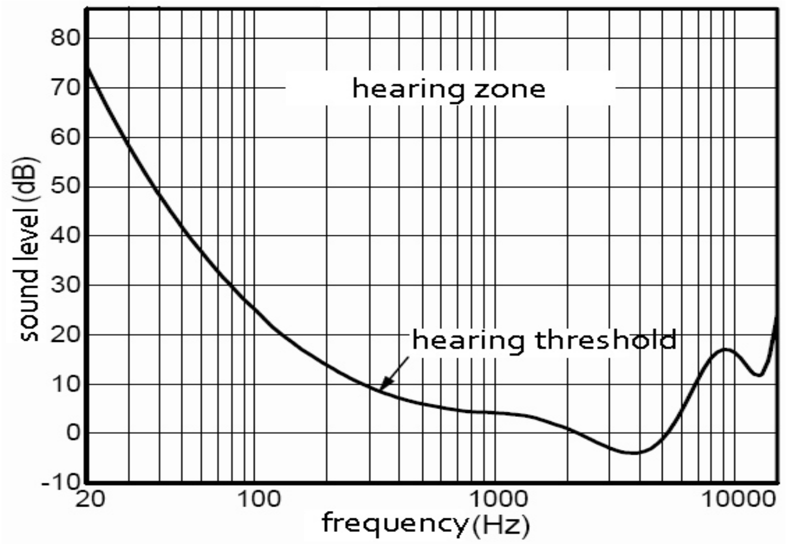
ITU-R BS.1387 does not consider the fact that input signals can have different levels, but states that they are all 0 dB FS (*full scale)* and of equal loudness. For determination of required amplification for the purpose of music reproduction on the same loudness level, the method described in [ROB01] is commonly accepted, which was developed by Robinson[10] in his Ph.D. Thesis [ROB02]. The implementation of this method, which is applicable on APEAQ, is called *WaveGain* and is available in [GAN05]. Modified APEAQ determines, using *WaveGain,* required amplification of reference and tested signals (separately). WaveGain gives this amplification with the scaling factor c_{gain} which is then multiplied with the input signal. Consequently, L_p which is used in (3.5) and (3.30) changes:

$$L_P = L_p + 10 \cdot \log_{10}\left(c_{gain}\right) \tag{4.1}$$

With this change, playback levels of the reference and the tested signal are leveled, so there is no need for them to be leveled again (3.53). Further spectrum adaptation (3.58) can hide distortions that occur in the tested signal, therefore for the calculation of $RmsNoiseLoud_A$ and $RmsMissingComponents_A$ excitation patterns are used without spectrum adaptation: $\tilde{E}_{S,ref}$ instead of $E_{P,ref}$ and $\tilde{E}_{S,test}$ instead of $E_{P,test}$. On $\tilde{E}_S$ is first applied interchannel masking, which will be described later in the text.

4.2. Absolute threshold of hearing

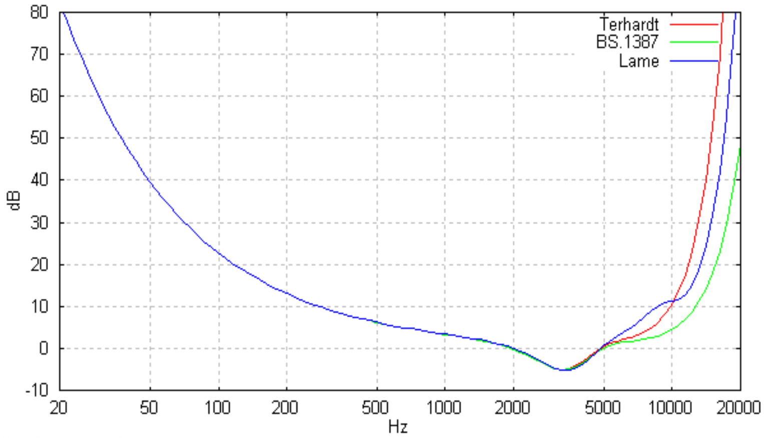
Threshold or limit of hearing is an indicator for energy which is required in order to hear a clear tone in an environment without noises. The threshold depends on a frequency, differs with each listener and declines on higher frequencies with age. Since only expert listeners, so-called 'golden ears', are selected as relevant in the objective evaluation of audio quality, it is the threshold of hearing of a young healthy listener, shown on the graph 4.3, that should be relevant for observation.



4.3 Absolute threshold of hearing [MIJ05]

[10] Ph.D. David John Michael Robinson, born in 1975., Rejnvort, Great Britain, http://www.david.robinson.org/.

The hearing threshold in BS.1387 is based on the approximation proposed in [TER79]. Only the degree which changes the upper frequency threshold of hearing has been changed (3.6).



4.4 – Comparing the approximations of the hearing threshold

In [LAM06] different approximation is used, which seems more true to actual values:

$$threshold / dB = 3.64(f / kHz)^{-0.8} - 6.8e^{-0.6(f/kHz-3.4)^2} + 6.0e^{-0.15(f/kHz-8.7)^2} + 0.0006(f / kHz)^4 \tag{4.2}$$

In adopting of this approximation we can see that the first term stays the same. Only frequency response is changed (3.7, 3.36), while internal noise stays the same (3.16, 3.45).

4.3. Approximation of the Bark scale

The transformation from time to frequency domain happens on the basal membrane of cochlea, in inner ear. Considering the anatomy of the basal membrane and its sensory receptors, frequencies are not uniformly distributed. If a frequency scale which corresponds to human hearing is used, and not the linear one, the effects, like simultaneous masking, can be easily explained and approximated.

4.3.1. Critical bands and the Bark scale

For creating a frequency scale of hearing several methods are used. ITU-R BS.1387 accepted the scale from [ZWI67] because it produced the best results. By using this scale, spectrum is divided in critical bands. It was derived from the experiments for determination of loudness. They showed that the loudness of narrowband noise with constant SPL had not changed, although frequency band of noise increased up to the width of a critical band. If band width of noise increases beyond critical band width then loudness increases, too. Consequently the experiments proved the dependency of critical band width on its central frequency. Further conclusion that was drawn was that a human ear is sensitive to about 24 critical bands. Idealized critical bands are presented in the table:

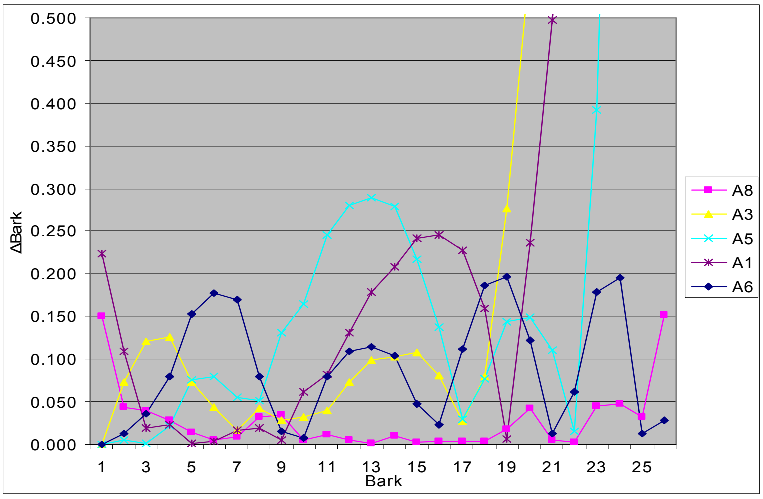
Bark	Cent.freq. [Hz]	Bound. Freq. [Hz]	My correction [Hz]
1	50	- 100	- 100
2	150	100 – 200	100 - 200
3	250	200 – 300	200 - 300
4	350	300 – 400	300 - 400
5	450	400 – 510	400 - 510
6	570	510 – 630	510 - 630
7	700	630 – 770	630 - 770
8	840	770 – 920	770 - 920
9	1000	920 – 1080	920 - 1090
10	1175	1080 – 1270	1090 - 1270
11	1370	1270 – 1480	1270 - 1500
12	1600	1480 – 1720	1500 - 1720
13	1850	1720 – 2000	1720 - 2030
14	2150	2000 – 2320	2030 - 2320
15	2500	2320 – 2700	2320 - 2740
16	2900	2700 – 3150	2740 - 3150
17	3400	3150 – 3700	3150 - 3750
18	4000	3700 – 4400	3750 - 4400
19	4800	4400 – 5300	4400 - 5360
20	5800	5300 – 6400	5360 - 6400
21	7000	6400 – 7700	6400 - 7770
22	8500	7700 – 9500	7770 - 9500
23	10,500	9500 – 12000	9500 - 12080
24	13,500	12000 – 15500	12080 - 15500
25	19,500 (17800)	15500 -	15500 - 20740

Table 4.1 – Critical bands [ZWI67]

4.3.2. Approximation of the Bark scale

Since audio coders and methods use more than 24 frequency subbands for objective evaluation, interpolation of values from the table 4.1 is required. For approximation of the Bark scale analytical functions are usually used. The existing approximations done by analytical functions are given in [CAR02]. Graphically compared errors for boundary frequencies are shown on the graph 4.5. The approximation which is used in BS.1387 is marked as A3 (in yellow) and it is clear that it is very bad compared to other ones. However, even the best one, by Traunmüller[11] (marked as A8), does not approximate the values from the table accurately enough. Also, another problem arose regarding this approximation, and it is concerned with the width of frequency subbands. If there are many of them, their width doesn't always decrease with the increase of the central frequency, which is a necessary request to be fulfilled as far as experimental results are concerned.

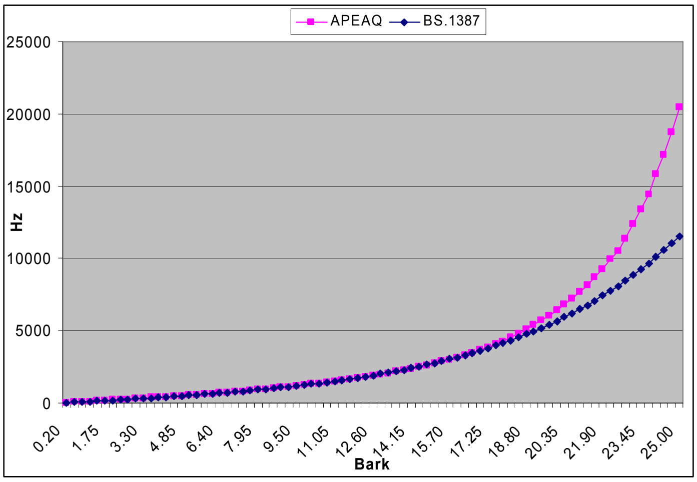
[11] Ph.D. professor Hartmut Traunmüller, born 1944. in Germany, Professor at University of Stockholm from1983.



4.5 – Comparison of errors in the Bark scale approximations

4.3.3. Spline approximation of the Bark scale

Spline interpolation is the only way for defining border and central frequencies, when the required number of subbands should be larger than 24. Cubic spline provides sufficient smoothness of the approximation [RAD91], which is an essential requirement. However, it was noticed that the width of subbands increases with the rise of a central frequency. In order to solve this problem and preserve required smoothness, slight changes were made in the table 4.1. Also, upper border for the 25th critical band was added, derived by extrapolation, and the central frequency of the 25th band was changed, which was anyway determined by extrapolation and at the same time too close to the highest audible frequency. By definition, the value of spline in each of the nodes is identical to the table values; therefore the error displayed on the graph 4.5 is 0, i.e. x axis. For the interpolation and the extrapolation, the package [ADV02] was used.



4.6 – Changes on the frequency scale

Graph 4.6 shows the changed frequency scale in relation to the original one from ITU-R BS. 1387-1. Central frequencies from the filter bank model are marked.

Due to the changes on the frequency scale, it is necessary to also change the length of filters $N[i]$ (3.34) and the distance between subbands (3.39) to $\Delta z = 0.62$ Barks. Filter length is defined by the formula (183) from [KAB06]:

$$N[i] = \frac{2 \cdot F_S}{B^{-1}\left(z_c[i] + \Delta z / 2\right) - B^{-1}\left(z_c[i] - \Delta z / 2\right)} \qquad 0 \le i < Z \qquad (4.3)$$

where $z_c = \mathrm{B}(f_c[i])$, and B is mapping from the linear frequency scale to the Bark scale. Central frequencies and lengths of the filters (i.e. length of their impulse responses) are given in the table:

#Filter	Cent.freq. f_c[i]	Filter length N[i]	#Filter	Cent. freq. f_c[i]	Filter length N[i]
0	51.0	1548	20	1968.6	509
1	113.0	1548	21	2159.2	492
2	175.0	1548	22	2372.5	429
3	237.0	1548	23	2603.8	387
4	299.0	1548	24	2864.5	371
5	361.0	1548	25	3155.6	298
6	423.0	1548	26	3472.5	279
7	488.3	1291	27	3858.6	252
8	566.1	1261	28	4267.7	190
9	641.5	1216	29	4792.2	182
10	728.9	1073	30	5410.7	146
11	818.8	1084	31	6057.2	132
12	911.2	969	32	6833.5	125
13	1011.5	939	33	7656.9	104
14	1113.6	856	34	8660.4	86
15	1233.7	807	35	9953.2	75
16	1356.6	753	36	11349.7	51
17	1491.3	683	37	13365.5	47
18	1640.8	626	38	15800.2	35
19	1796.5	591	39	18719.7	29

Table 4.2 - Central frequencies and filter lengths of modified APEAQ

In the FFT model of the PEAQ's basic version $Z = 109$ frequency subbands are used, and $\Delta z = 0.25$ Barks. In the advanced version $Z = 55$ is used due to performance issues. The number of subbands of the FFT model has influence only on $SNMR_B$ (3.81). The authors of ITU-R BS.1387 did not notice a big difference in *ODG* due to lower Z value, which they explained as the consequence of the FFT model being used in combination with the filter bank model. However, there is obviously greater correlation of $SNMR_B$ values with SDG, provided that more subbands are used. Therefore, in modified *APEAQ* it is set that $\Delta z = 0.25$ Barks, and subsequently $Z = 97$ because of the change of the scale. The loss in speed of the algorithm is minimal.

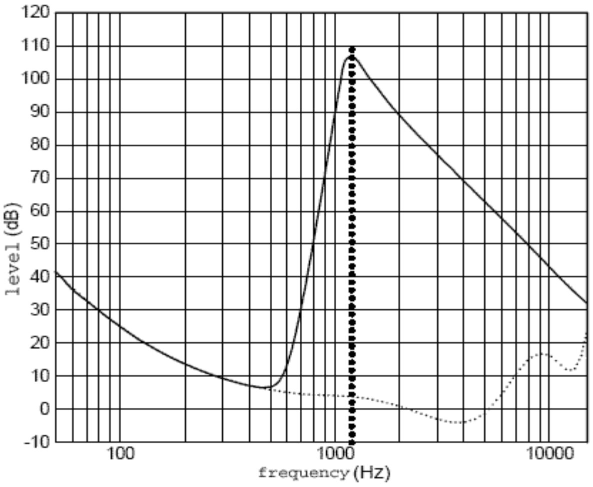
i	Boundary freq.	Cent. freq. f_c [i]	i	Boundary freq.	Cent. freq. f_c [i]
0	37.5 - 62.5	50.0	49	1884.9 - 1956.7	1920.5
1	62.5 - 87.5	75.0	50	1956.7 - 2037.5	2000.0
2	87.5 - 112.5	100.0	51	2037.5 - 2114.6	2075.7
3	112.5 - 137.5	125.0	52	2114.6 - 2190.4	2150.0
4	137.5 - 162.5	150.0	53	2190.4 - 2273.7	2231.7
5	162.5 - 187.5	175.0	54	2273.7 - 2363.8	2320.0
6	187.5 - 212.5	200.0	55	2363.8 - 2454.0	2408.4
7	212.5 - 237.5	225.0	56	2454.0 - 2547.5	2500.0
8	237.5 - 262.5	250.0	57	2547.5 - 2645.7	2596.1
9	262.5 - 287.5	275.0	58	2645.7 - 2751.9	2700.0

10	287.5 - 312.5	300.0	59	2751.9 - 2859.2	2804.9
11	312.5 - 337.6	325.0	60	2859.2 - 2956.8	2900.0
12	337.6 - 362.7	350.0	61	2956.8 - 3074.7	3015.1
13	362.7 - 388.4	375.5	62	3074.7 - 3212.5	3150.0
14	388.4 - 413.1	400.0	63	3212.5 - 3342.5	3276.6
15	413.1 - 439.7	426.4	64	3342.5 - 3469.5	3400.0
16	439.7 - 463.7	450.0	65	3469.5 - 3614.7	3541.1
17	463.7 - 491.5	477.5	66	3614.7 - 3778.1	3700.0
18	491.5 - 524.4	510.0	67	3778.1 - 3941.5	3858.6
19	524.4 - 553.7	538.9	68	3941.5 - 4088.0	4000.0
20	553.7 - 585.2	570.0	69	4088.0 - 4272.3	4178.8
21	585.2 - 616.3	600.6	70	4272.3 - 4500.0	4400.0
22	616.3 - 646.3	630.0	71	4500.0 - 4708.8	4602.9
23	646.3 - 679.5	662.7	72	4708.8 - 4911.9	4800.0
24	679.5 - 717.2	700.0	73	4911.9 - 5144.7	5026.8
25	717.2 - 752.3	734.6	74	5144.7 - 5425.0	5300.0
26	752.3 - 788.1	770.0	75	5425.0 - 5684.4	5553.1
27	788.1 - 825.1	806.5	76	5684.4 - 5937.9	5800.0
28	825.1 - 859.1	840.0	77	5937.9 - 6224.4	6079.4
29	859.1 - 897.9	878.4	78	6224.4 - 6550.0	6400.0
30	897.9 - 939.4	920.0	79	6550.0 - 6863.2	6704.3
31	939.4 - 978.8	959.0	80	6863.2 - 7169.1	7000.0
32	978.8 - 1020.4	1000.0	81	7169.1 - 7524.7	7343.8
33	1020.4 - 1062.0	1041.1	82	7524.7 - 7893.8	7700.0
34	1062.0 - 1101.9	1080.0	83	7893.8 - 8304.0	8094.9
35	1101.9 - 1146.7	1124.1	84	8304.0 - 8726.3	8500.0
36	1146.7 - 1198.3	1175.0	85	8726.3 - 9207.4	8961.9
37	1198.3 - 1246.1	1222.0	86	9207.4 - 9768.8	9500.0
38	1246.1 - 1295.0	1270.0	87	9768.8 - 10340.7	10048.8
39	1295.0 - 1346.2	1320.4	88	10340.7 - 10817.1	10500.0
40	1346.2 - 1396.6	1370.0	89	10817.1 - 11492.1	11147.6
41	1396.6 - 1451.1	1423.6	90	11492.1 - 12375.0	12000.0
42	1451.1 - 1508.1	1480.0	91	12375.0 - 13173.5	12765.9
43	1508.1 - 1565.8	1536.7	92	13173.5 - 13943.5	13500.0
44	1565.8 - 1630.1	1600.0	93	13943.5 - 14889.6	14406.3
45	1630.1 - 1691.9	1660.7	94	14889.6 - 16023.5	15500.0
46	1691.9 - 1752.5	1720.0	95	16023.5 - 17145.5	16571.4
47	1752.5 - 1819.3	1785.6	96	17145.5 - 18378.8	17747.3
48	1819.3 - 1884.9	1850.0			

Table 4.3 – Central and boundary frequencies in the FFT model of modified APEAQ

4.4. Frequency masking

Frequency masking, also known as simultaneous masking, happens when some existing sound of a certain frequency – masker, causes the change of hearing threshold for some other sounds, which are near in the spectrum. These sounds occur simultaneously. What happens is that the masker causes vibrations of the basal membrane not only within its own frequency, but also in the area surrounding it. In ITU-R BS.1387 the masking is modeled by the smearing of *excitation patterns* in the spectrum (3.19-3.25 and 3.37-3.42).



4.3 – Masking with an existing masker level of 110dB frequency 1200Hz [MIJ05]

There are four types of masking: *Noise-masking-tone NMT*, *Tone-masking-noise TMN*, *Noise-masking-noise NMN*, and *Tone-masking-tone TMT*. In each case noise in question is narrowband noise, whose spectrum is limited by the width of a critical band. *PEAQ* does not recognize explicitly different types of masking, but they are the result of the grouping into frequency subbands. TMT is the most complex one. The filter bank model successfully models TMT thanks to the filter's features and the fact that the smearing in spectrum takes place before any of the non-linear operations [THI99]. The grouping into frequency subbands also implicitly models the masking inside critical bands.

The masking between different critical bands is explicitly modeled. Masking models are different from one another depending on whether they take into consideration the amplitude of a masker and whether the transition between the slopes is smoothed. The soft transition between the upper and lower slopes is important if frequency resolution is high. Since frequency subbands are wide, even in the original version of the FFT model, it is not necessary that the soft transition is explicitly modeled. It is a result of the grouping process itself.

Masking dependency on the amplitude is a very important characteristic. Inside a frequency subband this dependency of masking on the masker's amplitude is again implicitly modeled. It is modeled among the subbands according to the formula (3.20) and (3.38) from [TER79]. In a model with such dependency it is important to know at which level a signal will be reproduced. Since we do not usually know the playback level, it is possible to use approximation of the worst case that can happen. On the other hand, if the playback level can be predicted, it is better to use the amplitude dependency masking. With introduction of *WaveGain* (chapter 4.1), the amplitude dependency becomes the best solution.

All in all, substantial improvement in the modeling of frequency masking does not seem to be possible.

4.5. Temporal resolution in the filter bank model

In order to speed up the execution of algorithms, temporal resolution is reduced at the output of the filter bank by selecting each I_S = 32th value (3.35). *Excitation patterns* are at the output of backward masking calculated at each I_{S2} = 6th input (3.44), so that sampling frequency of $\tilde{E}_S\,[i,n]$ equals $F_S/192$.

Since the widths of the last nine frequency subbands in PEAQ are bigger than $(F_S/32)/2 = 750$ Hz, *aliasing*, can occur. However, in [THI99] is stated that aliasing is not expected to happen on the signals for which PEAQ is used.

In the changed scale (table 4.2) in the last eight frequency subbands aliasing is likely to happen. By changing I_S to 16, only the last four subbands have got the width bigger than 1500 Hz, which corresponds with the half of the sampling frequency $F_S/16 = 3000$ Hz. This way the possibility of aliasing to occur is significantly reduced. In other words, the formula (4.3) defines the frequency range in which decay of a filter is less than -6 dB, while the full band is 24000 Hz. Aliasing is less likely to happen outside the band defined with (4.3).

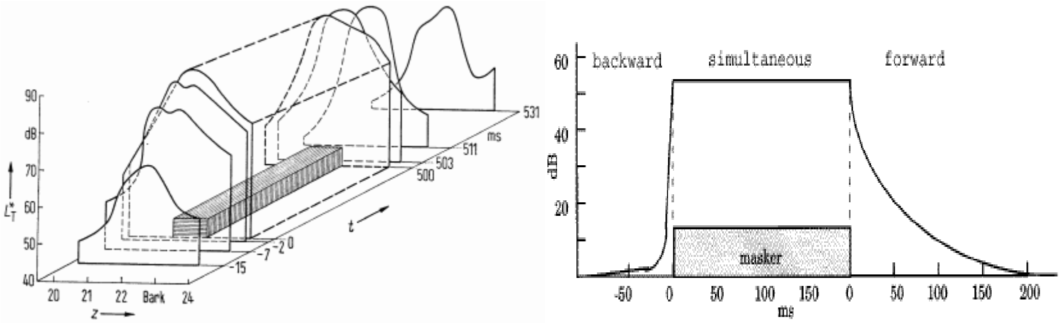
New value I_S requires a change of patterns' smearing in time, while the smearing in spectrum remains the same. The simplest adaptation, probably the best one too, is setting the sampling rate of the backward spreading outputs to 12 ($I_{S2} = 12$ u 3.44).

These changes reduce the speed of the algorithms. Nevertheless, even when implemented this way, the algorithm is much faster than the reference *Opera* (about 4 times faster).

4.6. Temporal masking

Temporal masking occurs when before and after a certain sound (masker) occurs, other sounds, which have similar frequency, are not heard. In PEAQ forward masking is modeled by first order low-pass IIR filter, which smears excitation patterns in time (3.28 and 3.48). Backward masking is modeled by low-pass FIR filter (3.44). Experiments showed that duration of forward masking is from 50 to 300 ms, and backward masking from 1 to 20 ms. Since the temporal resolution of FFT model is 21.3 ms, the modeling of backward masking is impossible in it.

In [THI99] is specified that changing parameters, which have influence on the duration of temporal masking, does not cause serious consequences regarding the result, and that was confirmed through the experiments with various values of the parameters (3.26, 3.44 and 3.46) in *APEAQ*. The time smearing effect on excitation patterns can be seen on the graphs (pictures 3.5 and 3.10), which correspond to the experimental results (picture 4.4).



4.4 – Masking curves [THI99, PAI00]

4.7. Binaural hearing

The modeling of binaural hearing in *PEAQ* is very simple – it is defined as the arithmetic mean of *MOV* for left and right channel.

$$MOV = \frac{1}{Number\_of\_Channels} \sum_{iChn=0}^{Number\_of\_Channels} MOV_{chn}[iChn] \tag{4.4}$$

MOVs are first averaged in time, so that they represent the evaluation of quality for each single channel on the whole tested signal.

Instead of the arithmetic mean, in modified APEAQ maximum is introduced:

$$MOV = \max_{iChn}\left(MOV_{chn}[iChn]\right) \tag{4.5}$$

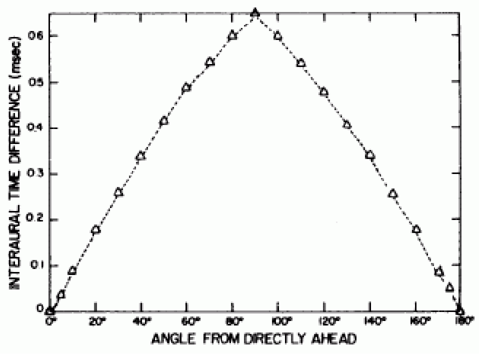
Interchannel masking is also implemented, according to [LAM06]. The chosen level of interchannel masking is -37 dB. Higher level is not used, because, during subjective evaluation, audio signals can be also listened to with headphones, where the interchannel masking is minimal. It is applied on the *smeared excitation patterns* of the bank filter model (3.48):

$$\begin{aligned} \widetilde{E}_{S,right}[i,n] &= \widetilde{E}_{S,right}[i,n] + 10^{-37/10} \cdot \widetilde{E}_{S,left}[i,n] \\ \widetilde{E}_{S,left}[i,n] &= \widetilde{E}_{S,left}[i,n] + 10^{-37/10} \cdot \widetilde{E}_{S,right}[i,n] \end{aligned} \tag{4.6}$$

These are minimal improvements of the binaural model. For a binaural model it is important to know the configuration of a sound source. Headphones and speakers produce very different stereo effects, and the arrangement of speakers is also important. From the results of the subjective tests, which are available, we do not know whether speakers or headphones were actually used for listening, so therefore it is impossible to implement a more accurate binaural model.

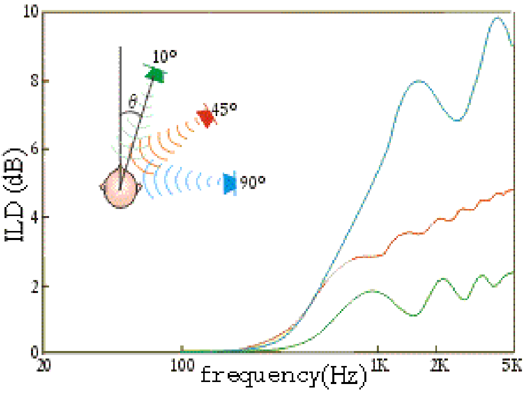
Sound localization has been considerably studied [GAR94, BIR05, HAR99, CIP05, MAI97]. Humans use a combination of clues in order to accurately determine the position of a sound source. In this thesis the most important ones will be described.

The sound that comes from a source which is not positioned right in front of or behind a listener does not reach both ears at the same time. This time difference is called *Interaural Time Difference - ITD*. The wave length of 1.5 kHz sine signal corresponds to the distance between the ears. Therefore, for clear tones with frequency over 750 Hz it is impossible to determine a unique *ITD*. On the other hand, for complex signals there is a way to determine *ITD,* even though all components are over 750 Hz. *ITD* is used for the determination of an angle in the horizontal plane.



4.5 – Dependency of ITD on the position of a sound source [ROB02]

On higher frequencies, human head makes a barrier for audio signals, so the signals that came from one side faded on the other. This phenomenon is quantified by *Interaural Level Difference – ILD*.



4.6 – Dependency of ILD on the frequency and the position of a sound source [HAR99]

The shape of a pinna also enables the localization. The pinna acts as an audio filter, and the frequency response of that filter depends on the direction from which the sound comes from (from both elevation and azimuth). This filtering effect is not heard as a change in audio spectrum, but is decoded by auditory system as spatial information, which enables us to discover the position of the source only with one ear. *HRTF - Head Related Transfer Function* which determines this filtering process can be found in [GAR94].

Movements of a head also help to determine the position of a source. Without movements, it would be difficult to tell whether a source was in front of us or behind us, or whether it was above or under the horizontal plane.

ITD, *ILD* and *HRTF* can be used for extracting of spatial information from a signal, but for that it is required to know in advance the configuration on which the sound would be reproduced. The lack of this required knowledge makes it impossible to integrate these methods in APEAQ.

4.8. Change in MOVs for distortion loudness

MOV AvgLinDist$_A$ has little correlation with *SDG*. Since its usage is not supported by the psychoacoustics, it is not used in the modified APEAQ. Once it is dismissed, there is no need for spectrum adaptation, and consequently the steps (3.49-3.58) are not carried out.

So-called spectrum holes can greatly affect the evaluation of audio quality [HYD06a]. They are frequency components of a signal, which exist in the reference signal, but not in the tested one. Components which are missing in the tested signal and additional distortions make a complex link, which determines the final measure of quality.

In *PEAQ* the added distortions are defined by *RmsNoiseLoud$_A$*, and the missing components are defined by *RmsMissingComponents$_A$*. Their simple combination (3.75) can be considered to be artificially imposed relationship of weight coefficients in 1:2 ratio for independent inputs into the neural network. The imposed ratio makes the approximation of a complex link which exists among these MOVs impossible. Their separation allows the training of a neural network to find independent coefficients which will best approximate the joined influence of these *MOVs* on *ODG*.

Dismissing *AvgLinDist$_A$* and separating *RmsNoiseLoud$_A$* and *RmsMissingComponents$_A$*, the number of inputs of the neural network remains the same, as well as its complexity.

The defining of loudness of distortions is the most important part of PEAQ. Among all the MOVs, *RmsNoiseLoudAsym$_A$* has the greatest correlation with SDG. Barbedo and Lopes [BAR05] proposed in their PEAQ modification the change of parameters for the calculating of *RmsNoiseLoud$_A$* (*AvgLinDist$_A$* and *RmsMissingComponents$_A$* are not used). Degree γ were changed from 0.23 to 0.08, and β was fixed to 1 (3.73). With these changes made, the new formula is:

$$N_L[i,n]=\left(\frac{E_{IN}[i]}{s_{test}[i,n]}\right)^{0.08}\left[\left(1-\frac{\max\left(s_{test}[i,n]\cdot E_{test}[i,n]-s_{ref}[i,n]\cdot E_{ref}[i,n],0\right)}{E_{IN}[i]+s_{ref}[i,n]\cdot E_{ref}[i,n]}\right)^{0.08}-1\right] \quad (4.7)$$

Thiede[12] [THI99] developed the formula for loudness of distortions from Zwicker's formula for loudness:

$$N=k\cdot\left(\frac{E_{Thres}[i]}{s_{thres}[i]\cdot E_0}\right)^{\gamma}\left[\left(1-\frac{s_{thres}[i]\cdot\left(E[i,n]-E_{Thres}[i]\right)}{E_{Thres}[i]}\right)^{\gamma}-1\right] \quad (4.8)$$

where k and E_0 are constants and thus do not affect the final result, because they are compensated by the change of weight factors of a neural network. E_{Thres} represents absolute threshold of hearing (ATH). It is not clear why Thiede changed absolute threshold of hearing with the internal noise. Since absolute hearing threshold is already included in the calculation of excitation patterns (and internal noise too), by changing E_{IN} with E_{Thres}, the formula (4.7) is now:

$$N_L[i,n]=\left(\frac{E_{Thres}[i]}{s_{test}[i,n]}\right)^{0.08}\left[\left(1-\frac{\max\left(s_{test}[i,n]\cdot E_{test}[i,n]-s_{ref}[i,n]\cdot E_{ref}[i,n],0\right)}{s_{ref}[i,n]\cdot E_{ref}[i,n]}\right)^{0.08}-1\right] \quad (4.9)$$

[12] PhD Thilo Thiede, born in 1967., Berlin, Germany.

Maximum is used in order to avoid negative values. If $s_{test}E_{test} < s_{ref}E_{ref}$ then $N_L[i,n] = 0$. For $s_{test}E_{test} < s_{ref}E_{ref}$ the following formula can be used:

$$N_L[i,n] = \left(\frac{E_{Thres}[i]}{s_{test}[i,n]}\right)^{0.08}\left[\left(1 - \frac{s_{test}[i,n]\cdot E_{test}[i,n] - s_{ref}[i,n]\cdot E_{ref}[i,n]}{s_{ref}[i,n]\cdot E_{ref}[i,n]}\right)^{0.08} - 1\right] \tag{4.10}$$

Second fraction can be simplified by division with $s_{ref}E_{ref}$. The second part of the formula is then determined by the excitation patterns' ratio of the tested and the reference signals. In the first fraction s_{test} is derived from s_{Thres} which compensated s_{Thres} in the second fraction of the formula (4.8). Since changes result in the compensation of s_{test} with s_{ref}, s_{test} is not required in the first fraction. The changed formula looks like this:

$$N_L[i,n] = \left(E_{Thres}[i]\right)^{0.08}\left[\left(\frac{s_{test}[i,n]\cdot E_{test}[i,n]}{s_{ref}[i,n]\cdot E_{ref}[i,n]}\right)^{0.08} - 1\right] \tag{4.11}$$

The thresholds of hearing are canceled out in the fraction, so E_{Thres} is a necessary factor for frequency weighting. Naturally, the improved threshold of hearing is used (4.2).

The formula (4.11) is used instead of (3.71) for the calculating of $RmsNoiseLoud_A$ and $RmsMissingComponents_A$. The excitation patterns from (4.6) are used in the calculation, E_{ref} is set to $\tilde{E}_{S,ref}$ and E_{test} is set to $\tilde{E}_{S,test}$ for $RmsNoiseLoud_A$, and E_{ref} is set to $\tilde{E}_{S,test}$ and E_{test} is set to $\tilde{E}_{S,ref}$ for $RmsMissingComponents_A$.

Thus defined *MOVs* have greater correlation with the subjective values, compared to the initial ones, but there are still the examples in which these *MOVs* evaluate bad quality, and the subjective evaluation of quality is very high. These examples have few frames on which *MOVs* are drastically above average ones. These great departures, of a very short intensity can cause big final values of *MOV* due to the root mean square averaging (3.76). The logical explanation would probably be that a listener can't hear these short distortions clearly. However, in [THI99] it is said that the root mean square averaging produces better results than the linear one.

The solution was found in the windowed averaging which is used for some variables in the basic version of PEAQ, with the selected window length of 20 ms, unlike the one of 100 ms in the basic version. The formula given in [ITU01] is:

$$N_{LRMS} = \sqrt{\frac{1}{N-L+1}\sum_{n=L-1}^{N-1}\left(\frac{1}{L}\sum_{m=0}^{L-1}\sqrt{N_{LM}[n-m]}\right)^4} \tag{4.12}$$

For the window of 20 ms, L is 5. N_{LM} is defined with (3.74).

5. Artificial neural network

In this chapter the applied algorithm for the training of the neural network and problems which arise during the process will be described.

5.1. Training of a neural network

The structure of the neural network described in chapter 3.6, remains unchanged in the modified APEAQ. The changes described in chapter 4, require determination of new weight coefficients, i.e. a training of the neural network. Since it is known from the subjective tests, which results should be produced, supervised training is applied.

The inputs of the neural network are values of five *MOVs* and the output is *ODG*[13]. The neural network models cognitive processes, that listeners use in order to translate distortions they hear in the tested signal into quality grade - *SDG*. *ODG* should be equal to *SDG* from the subjective tests, and possible deviations depend on the weight coefficients.

5.1.1. Incremental backpropagation

The training is done by an algorithm of incremental backpropagation [SMI99][MIL05]. *sig*(*DI*) is calculated for one of test signals and its deviation from the expected value:

$$r(DI) = sig(DI) - \frac{SDG - b_{\min}}{b_{\max} - b_{\min}} \tag{5.1}$$

where b_{min} and b_{max} are used for the linear transformation of *sig*(*DI*) into *ODG* (3.95). *h* represents input for activation functions of the nodes in the hidden layer:

$$h_j = w_x[5,j] + \sum_{i=0}^{4} w_x[i,j] \cdot MOV'[i], \qquad 0 \le j < 5 \tag{5.2}$$

The coefficients in the hidden layer are corrected according to gradient method:

$$\begin{aligned} \Delta w_x[i,j] &= MOV'[i] \cdot sig'(h_j) \cdot w_y[j] \cdot sig'(DI), \\ \Delta w_x[5,j] &= sig'(h_j) \cdot w_y[j] \cdot sig'(DI), \end{aligned} \qquad 0 \le j < 5, 0 \le i < 5 \tag{5.3}$$

$$w_x[i,j] = w_x[i,j] + c \cdot r(DI) \cdot \Delta w_x[i,j], \qquad 0 \le j < 5, 0 \le i \le 5 \tag{5.4}$$

The correction of the coefficients in the output layer is defined with the following formulas:

$$\begin{aligned} \Delta w_y[j] &= sig(h_j) \cdot sig'(DI), \\ \Delta w_y[5] &= sig'(DI), \end{aligned} \qquad 0 \le j < 5 \tag{5.5}$$

$$w_y[j] = w_y[j] + c \cdot r(DI) \cdot \Delta w_y[j], \qquad 0 \le j \le 5 \tag{5.6}$$

[13] Strictly speaking, *sig*(*DI*) is defined, from which *ODG* is determined by linear transformation.

Constant *c* defines the speed of convergence and is called the learning constant. Convergence speed is also determined by *r*(*DI*), which is proportional to deviation from the correct value.

For this training process, activation function needs to be differentiable. Standard logistic function (3.92) has its advantages, because there is a shortcut for calculating its first derivative, provided that the value of the function is known in the point x:

$$sig'(x) = \frac{e^{-x}}{\left(1+e^{-x}\right)^2} = sig(x) \cdot [1 - sig(x)] \tag{5.7}$$

After the correction of coefficients based on one test signal, the same procedure is applied with the next one. This procedure is continued until the last *Nth* signal. The correction based on all *N* signals makes one pass.

Before we start with the incremental backpropagation, randomly selected values for the initialization of weight coefficients are used. With these weight coefficients a neural network produces results which are equivalent to sheer guessing. A neural network improves its precision by the corrections based on one test at a time.

5.1.2. Standard measure of efficiency of a neural network

As a standard measure for neural network's efficiency root mean square error is used:

$$errorsum = \sqrt{\frac{1}{N}\sum_{n=0}^{N} r^2\left(DI_n\right)} \tag{5.8}$$

Root mean square error is, for the sake of efficiency, determined by accumulating during one pass, with the initialization at 0 before a pass. Thus defined *errorsum* is not accurate, but is good enough for representation of the convergence speed.

Beside a training set, ***validation set*** is used, too. Validation set is not used during the changing of weight coefficients. *errorsum* is also calculated for it and consequently it determines generalization ability of a neural network.

5.2. Problems in the training of a neural network

Here are some distinctive problems which are mutually related:

- Existence of local minima
- Determination of the learning constant
- Determination of initial value range
- Reliability of a training set
- Generalization ability of a neural network

The training process can be described as seeking for the minimum on a thirty-six-dimensional hypersurface (i.e. as many weight coefficients exist). It is unknown what this hypersurface looks like, so the only safe way to determine global minimum is the checking of each point on that hypersurface. However, even if the weight coefficients could be rationally quantized, the checking process of all the possible values would take too long, like for example with uniform quantization to only 10 different values there would be 10^{36} possibilities to be checked, which would take about 10^{13} years. The removal of equivalent permutations does not significantly reduce the number of possibilities. During the training process it was noticed that the uniform quantization would not approximate hypersurface well, not even close. Namely, in many areas of it, little changes of the coefficients can cause big difference in the neural network efficiency, while in the other areas there are big plateaus with small slopes.

The hypersurface of coefficients also includes a number of local minimums, making it impossible to know whether a local minimum is at the same time global, too. Incremental backpropagation can only determine a local minimum within the area of initial, randomly selected values. The definition of that area is determined by the degree of changes on the weight coefficients. The only way to determine global minimum is by the determination of many local ones and by comparing them as well.

During the training process it could be noticed that the learning constant between 0.1 and 0.0001 provided satisfactory convergence. High constant value speeds up the convergence, but it results in oscillations around a local minimum or causes overlooking of it, while small value slows the convergence down.

In the algorithm for the training process, possibility of the learning constant adaptation was also implemented. Its adaptation is based on the root mean square error (5.8). The calculation of the root mean square error considerably slows down the algorithm, because in this case the accuracy in its calculation is required, which is possible to achieve provided that while calculating the *errorsum* none of the coefficients is changed. *Errorsum* is calculated before and after one training pass, which makes the process at least three times slower. The following adaptations are implemented:

- If the error increases, the learning constant decreases.
- If the error considerably increases, the weight coefficients are set to initial values they had before the training pass
- If the error decreases, but too slowly, the learning constant increases.

The range for weight coefficient in ITU-R BS.1387 is wide: between -40 and +20. If such wide range is selected for the initial random values, local minimums that provide poor results are likely to occur often, while the small range often produces good local minimums, but leaves out a number of potentially better values for the coefficients. The range of final coefficients is usually wider than the initial one. The range of the initial values between -10 and +10 produces satisfactory results, but still with a large number of plateaus with small slopes and big minimums. In order to avoid these plateaus, the suspension of the propagation algorithm and the reestablishing with new random values were used. Criteria for this suspension were very big *errorsum* values during the propagation process, which were defined during observations of the training. Additional criterion is the evaluation of the root mean square error in a local minimum based on the extrapolation.

5.2.1. Generalization ability of a neural network

The main problem in the training process is the selection of test signals for the training input. Neural network achieves proper generalization if the difference between *ODG* and *SDG,* for the signals which were not available during the training, is small. For good generalization it is necessary that the training set is a representative sample of all possible inputs. Since it is impossible to predict all possible inputs, there should be as many test signals as possible with all kinds of distortions, which have been detected so far during the subjective tests.

For the improvement of generalization ability, the root mean square error on a validation set is also observed. The root mean square error on a training set decreases with the number of passes, while on the validation set it decreases up to a certain number of passes, after which it gradually rises again. From the rising point of *errorsum* on the validation set, the specialization for the training set begins and the generalization ability decreases, so it is necessary to stop the training process. This principle is called ***early stopping***.

The smaller the number of neurons, the greater the generalization ability is. Great number of neurons leads to the specialization of a network for the training set. The number of neurons in the hidden layer was taken from ITU-R BS.1387 and it was not changed, so that, above everything else, the changes of the psychoacoustic model could be fairly compared with PEAQ.

Moreover, there is the issue of *SDG* representing the average evaluation derived from the results acquired with the help of only a selected number of representative listeners. The deviations from the average grade among the listeners were substantially big. Naturally, some other group of expert listeners would provide different results. That is why SDG should always be presented with what is called the ***confidence interval*** - the interval in which there will be, with 95% expectancy rate, the expected SDG of the whole population.

Another major issue is badly-defined subjective tests which are available, because of the possible errors that may occur during the statistical processing. For instance, some examples notably illustrate how SDG was +5 with one listener, although the average value among all other listeners was -3, and even without these errors in processing, the reliability was still in question. The fact that this problem with subjective tests can’t be avoided is best illustrated in the fact that no listener provides the same evaluations during different listening sessions.

5.3. Selection of tests for the training process

Here are the results from the subjective tests which were available:

- MPEG90 [ISO90]. 50 test signals. 5 to 70 participants. Average confidence interval was 0.61, the biggest was 2.78. If we exclude the tests with 5 listeners, then average confidence interval was 0.48, and the biggest was 0.85.
- MPEG91 [ISO91]. 105 test signals. 40 to 93 participants. Average confidence interval was 0.41, the biggest was 0.80.
- MPEG95 [MEA95]. 132 test signals. 63 participants. Average confidence interval was 0.50, the biggest was 0.78.
- ITU92DI [ITU92]. 60 test signals. 23 participants. Average confidence interval 0.91, the biggest was 1.58.
- ITU92CO [ITU92]. 60 test signals. 19 participants. Average confidence interval was 0.85, the biggest was 1.33.
- ITU93 [ITU93]. 42 test signals. 33 participants. Average confidence interval 0.71, the biggest was 1.04.
- DB3 [ITU01]. 82 test signals. 27 to 171 participants. Average confidence interval was 0.70, the biggest was 1,27.
- Extension [AMO03]. 72 test signals. 14 to 29 participants. Average confidence interval was 0.79, the biggest was 1.39.

- Mp3_128 [AMO04a]. 72 test signals. 11 to 22 participants. Average confidence interval was 1.22, the biggest was 2.12.
- Multiformat [AMO04b]. 108 test signals. 11 to 27 participants. Average confidence interval was 0.56, the biggest was 1.13.
- Mares [MAR05]. 108 test signals. 18 to 30 participants. Average confidence interval was 0.29, the biggest was 0.71.

The number of the participants shows how many of them did the evaluation of certain test signals. In some tests certain signals were not evaluated by all the participants.

Tests MPEG90, MPEG91, MPEG95, ITU92DI, ITU92CO, ITU93 and DB3 were used during the development of PEAQ. Tests Extension, Mp3_128, Multiformat and Mares came out from users wish to objectively evaluate available audio codecs and were carried out independently from commercial influence, via Internet.

5.3.1. Listeners' expertise test

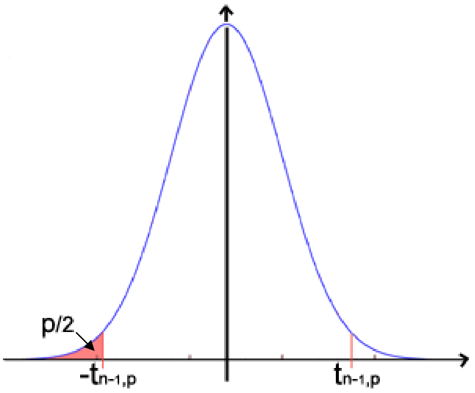
Some tests provided all the evaluations from all the participants, and not only the average grades. On these tests (DB3, Extension, Mp3_128, Multiformat and Mares) a statistical analysis was carried out. First of all it was tested whether a listener could be considered as an expert one. In Extension, Mp3_128, Multiformat and Mares tests two complementary methods for testing the expertise of a listener were used. The first method involves hidden test signals of substantially low quality, which an expert listener would definitely evaluate with equally bad grade. The second one requires from the listeners during ABX testing to determine which of the signals is the test signal and which is the reference one. In ABX testing listeners are given three signals, from which they know that one signal is the original one, and as for the other two they do not know which of them is the degraded (compressed) signal and which is the original one. After listening repeatedly they should be able to tell the difference between original and degraded signals with great certainty, otherwise their evaluation for a given signal would not be accepted as relevant [AMO05].

For DB3, statistical hypothesis test is used which states that the average SDG derived from this test is bigger or equal to 0. To check this hypothesis Student's t-distribution is used:

$$\frac{\overline{X}}{\overline{S}} \cdot \sqrt{N-1} < -t_{(N-1),0.10} \tag{5.9}$$

$$\overline{S}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{X})^2 \qquad\qquad \overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where N is a certain number of tests which listeners evaluated with SDG, and x_i represents individual grades. The selected critical region was 5%, and the number of non-expert listeners does not increase even with the critical region of 2.5%. The critical region of 5% was taken from [GRU92]. The table values for $t_{n,p}$ can be found in [MLA95] or [WIK06], where $t_{n,p}$ have different denotations, here is used the denotation from [MLA95]. If x_i for a listener are not good enough (5.9), the listener is considered not expert enough and none of their grades is used for defining an average SDG.



5.1 – Student’s t-distribution

The critical region size (marked in red p/2 on the graph 5.1) determines the probability when a non-expert listener can be characterized as an expert one. For $t_{(N\text{-}1),0.10}$, which is used in (5.9), is $p/2$ = 5%. The probability for an expert listener to be treated as a non-expert one is:

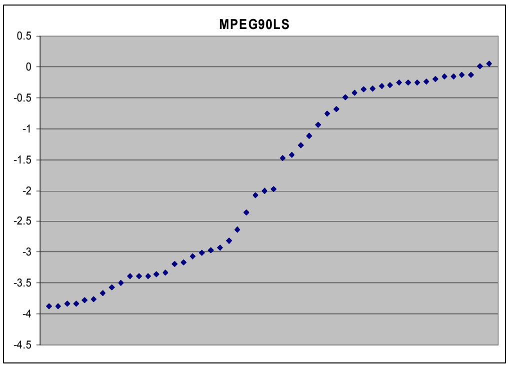
$$P\left(\frac{\overline{X}}{\overline{S}}\cdot\sqrt{N-1}\geq -t_{(N-1),0.10}\right)=P\left(\frac{\overline{X}-m}{\overline{S}}\cdot\sqrt{N-1}\geq -t_{(N-1),0.10}-\frac{m}{\overline{S}}\cdot\sqrt{N-1}\right)=P\left(t_{(N-1)}\geq -t_{(N-1),0.10}-\frac{m}{\overline{S}}\cdot\sqrt{N-1}\right)=\beta \tag{5.10}$$

Where m is the expected average grade of that listener on all possible relevant test signals, i.e. a measure of his/her expertise (the lower the m the bigger the expertise). $t_{N\text{-}1}$ has Student’s t-distribution, and β is proportional to m. If we take as an example that N = 41 and m = -1.0, we get that $\beta \approx 3\%$.

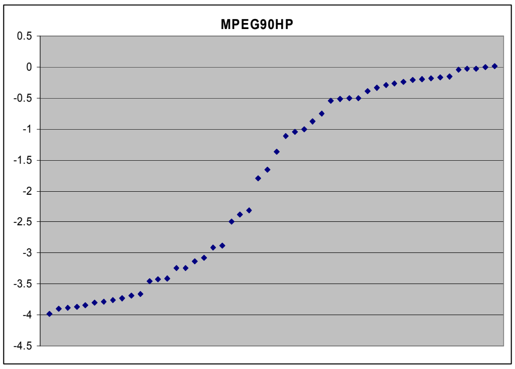
In a set of tests with seemingly transparent audio material, SDG is expectedly high. In those tests one should carefully decide which listeners are non-expert ones, e.g. by using smaller critical region. On the other hand, if the audio material covers the full scale range for SDG from -4 to 0, and if it is uniformly distributed, then a larger critical region can be used.

5.3.2. Selection of training, validation and testing sets

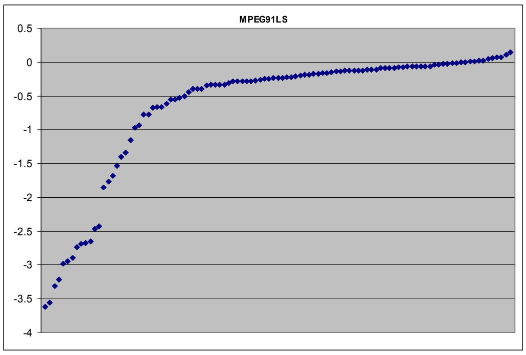
SDG range from the tests covers the whole interval from -4 to 0 (with some grades even bigger than 0):



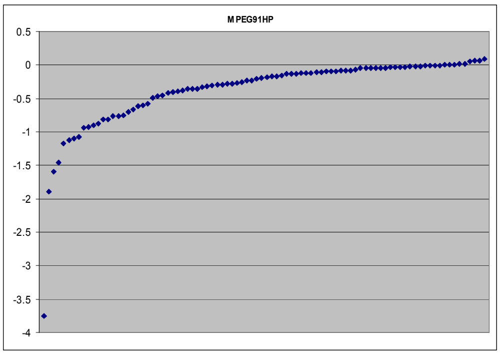
MPEG90 speakers



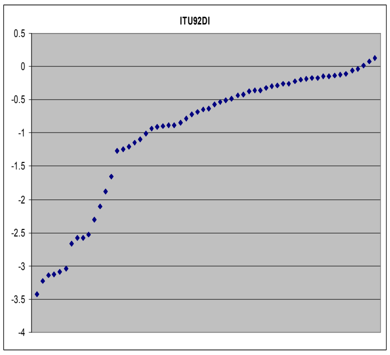
MPEG90 headphones



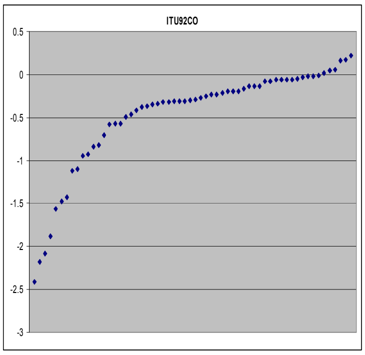
MPEG91 speakers



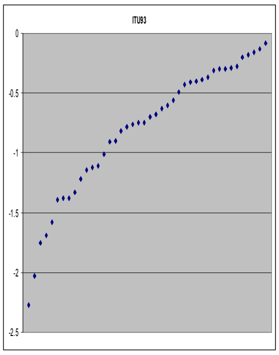
MPEG91 headphones



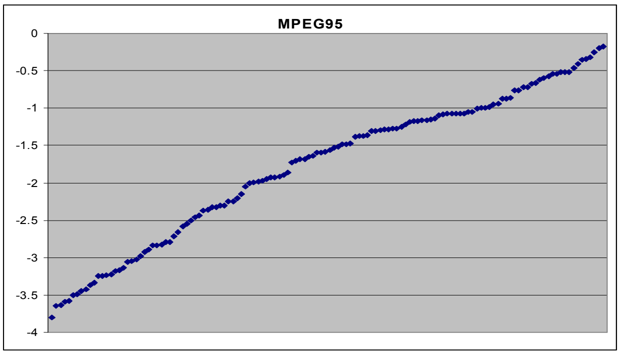
ITU92DI



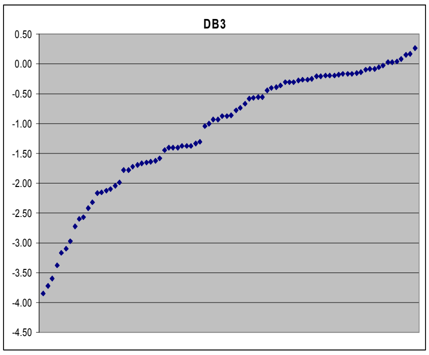
ITU92CO



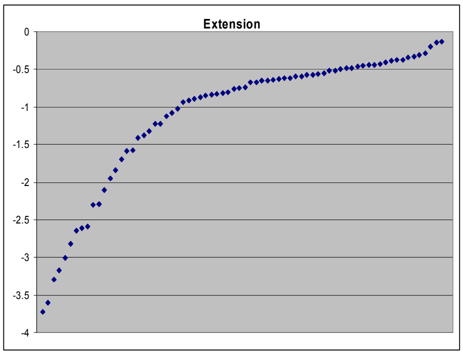
ITU93



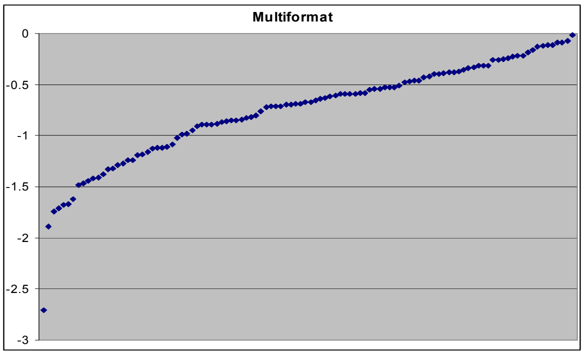
MPEG95



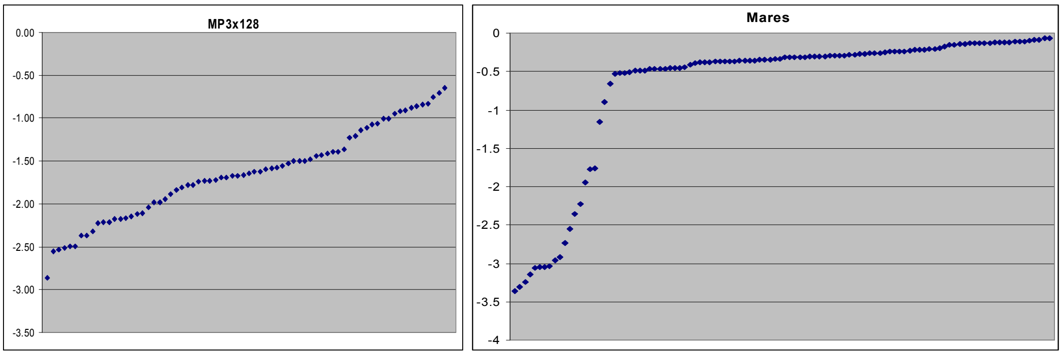
DB3



Extension



Multiformat



MP3_128 Mares

5.2 – SDG on available tests

For the training of a neural network in APEAQ the accessible tests from [ITU01] are used: MPEG90, MPEG91, MPEG95, ITU92DI, ITU92CO and ITU93. DB3 test is used as a validation set. Other tests, Extension, Mp3_128, Multiformat and Mares, are used for comparison of the generalization ability to *Opera.* Since these tests are completely independent from the tests in [ITU01] and they were not used during the training of the neural network for APEAQ, they represent the relevant set, based on which reliability of *Opera* and APEAQ can be compared.

5.3.3. Boundary values for MOV and ODG

New boundary MOVs were defined based on the MOVs values on the training set. These values are used for scaling (3.93) instead of those from ITU-R BS.1387. They were determined in a way that the highest correlation of each MOV with SDG was achieved.

i	MOV[i]	$a_{\min}[i]$	$a_{\max}[i]$	Coeff[*i*]
0	$RmsModDiff_A$	40.0	160.0	0.715
1	*RmsNoiseLoudA*	0.0	0.35	0.707
2	*RmsMissingComponentsA*	0.0	0.28	0.401
3	$SNMR_B$	-19.0	-2.5	0.696
4	EHS_B	0.15	0.77	0.596

Table 5.1 – Boundary values of MOVs

Also, new boundary values for ODG were defined:

$$b_{\min} = -3.96 \qquad b_{\max} = 0.03 \tag{5.12}$$

5.3.4. Cubic polynomial approximation of SDG

Then cubic polynomial was determined which approximates SDG well:

$$ODG_{MOV}[i] = 0.267 - 4.108 \cdot MOV'[i]$$

$$ODG_{lin} = \sum_{i=0}^{4} ODG_{MOV}[i] \cdot Coeff[i] \qquad 0 \le i \le 4 \tag{5.11}$$

$$ODG_{alt} = c_3 \cdot ODG_{lin}^3 + c_2 \cdot ODG_{lin}^2 + c_1 \cdot ODG_{lin} + c_0$$

MOV'[*i*] were scaled with the formula (3.93) based on constants from the table 5.1. *Coeff* represent correlation of MOV with SDG and are given in the table 5.1. The coefficients of the cubic polynomial were defined so that the correlation with SDG is maximized and are given in the following table:

C_3	C_2	C_1	C_0
0.01286	-0.14762	0.38298	-0.03

Table 5.2 – The coefficients of the cubic polynomial (5.11)

5.3.5. Introduction of fictive tests

In order to increase generalization ability, two fictive tests are introduced:

- Identity, for which all MOV values are equal to a_{min}, and SDG = b_{max}
- Worst, for which all MOV values are equal to a_{max}, and SDG = b_{min}

These fictive tests, aside from improving the generalization, also uniformly distribute errors along the full scale from b_{min} to b_{max}.

5.3.6. Criteria for excluding certain signals

Since it was noted that SDGs for certain signals probably do not reflect their real quality, and that MOV and the neural network cannot predict all possible values, criteria for rejecting certain signals were introduced:

- If $ODG_{alt} - SDG > 1.8$, listeners hear some distortions which APEAQ cannot predict
- If $SDG - ODG_{alt} > 1.0$, listeners did not hear some distortions
- If $SDG > SDG_{\text{Identitet}}$, listeners cannot hear distortions and evaluate reference signal incorrectly
- If the MOV values of the two tests are similar, and SDGs are very different. The function, which is represented by a neural network, cannot map two almost identical vectors into two different values of ODG
- Tests from MPEG90 with five listeners were not taken into consideration

These criteria improve generalization ability of a neural network and speed up its training. Based on these criteria, 38 from 418 tests were excluded (another 31 signals were not available), therefore 380 signals were used for the training of the neural network, 84 for the validation during the training process and 360 for the comparison with *Opera*.

5.4. Applied training process

Having considered all of the above, the following training process is applied:

- Random values between -10 and +10 are generated for each of the weight coefficients
- The coefficients are improved by applying incremental backpropagation
- If the root mean square error does not decrease fast enough, the training is aborted and new initial values are generated

- After 15000 iterations, the generated coefficients and values described in chapter 6.1. are recorded. The process is restarted by generating new initial values. Each of the recorded iterations represents a local minimum
- After a large number of local minimums were found (over 30000), the most suitable among them are selected based on the criteria described in chapter 6.1
- On the selected values of weight coefficients backpropagation is restarted, but this time with the adaptation of the learning constant and early stopping.
- Among all additional trainings it is those weight coefficients that are selected which will produce the best results for the validation set

The weight coefficients selected by this procedure are given in the following tables:

i	MOV[*i*]	$w_x[i,0]$	$w_x[i,1]$	$w_x[i,2]$	$w_x[i,3]$	$w_x[i,4]$
0	$RmsModDiff_A$	11.7308	-2.20653	-2.01372	-10.6767	-1.97304
1	*RmsNoiseLoudA*	-7.92159	2.13716	7.40097	2.94282	5.92348
2	*RmsMissingComponentsA*	5.49862	0.257564	0.497462	-4.8148	-0.20836
3	$SNMR_B$	2.42094	8.35004	-9.39781	-2.24411	-1.52524
4	EHS_B	-4.37367	2.81898	-12.985	6.41475	1.27332
5	*Bias*	-0.327587	-8.90636	4.61655	7.47879	-0.353271

Table 5.3 – The coefficients in the hidden layer

j	$w_y[0]$
0	-1.65981
1	-4.91063
2	1.43519
3	1.78166
4	-3.94657
5	3.56036

Table 5.4 – The coefficients in the output layer

These coefficients together with $a_{\max}$ (table 5.1), $b_{\min}$ and $b_{\max}$ (5.12) determine the new mapping from MOV into ODG (3.92-3.95).

6. Results

6.1. Criteria for the accuracy of objective grades

In order to determine the accuracy of ODG a few criteria are used. In ITU-R BS.1387, correlation, AES, tolerance scheme and number of tests for which |*ODG-SDG*| is high are used. I will introduce a few other very common criteria: root mean square error, mean absolute error and maximum error as well as a percentage of tests with ODG outside of the confidence interval - ***outliers.*** None of the criteria taken separately is reliable enough regarding the accuracy of the grades.

Root mean square error is defined with:

$$SKG = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(ODG_i - SDG_i\right)^2} \tag{6.1}$$

Mean absolute error is defined with:

$$SG = \frac{1}{N}\sum_{i=1}^{N}\left|ODG_i - SDG_i\right| \tag{6.2}$$

Maximum error is defined as the maximum value of $|ODG_i\text{-}SDG_i|$ among all tests. Knowing the value of maximum error makes the interpretation of a scatter plot easier, but it does not produce relevant information about the quality of ODG, because it is almost always determined with one distinctive *outlier.*

Correlation coefficient is defined with:

$$Correl = \frac{\sum_{i=1}^{N}(ODG_i - ODG_{avg})(SDG_i - SDG_{avg})}{\sqrt{\sum_{i=1}^{N}(ODG_i - ODG_{avg})^2\sum_{i=1}^{N}(SDG_i - SDG_{avg})^2}} \tag{6.3}$$

$$ODG_{avg} = \frac{1}{N}\sum_{i=1}^{N}ODG_i \qquad SDG_{avg} = \frac{1}{N}\sum_{i=1}^{N}ODG_i$$

Correlation coefficient represents a measure of a linear relationship of two random variables, in this case ODG and SDG. |*Correl*| value is between 0 and 1, where |*Correl*| = 1 if and only if there is linear dependency (if r=-1, that would mean that the correlation between 4-ODG and SDG equals 1). The closer |*Correl*| is to 0, the correlation of SDG and SDG is smaller.

It is not advisable to draw conclusions about ODG accuracy based only on the correlation coefficient. Each ODG which differs from SDG can have a big impact on the correlation, even though on most of the tests ODG is very similar to SDG [STA06]. It is recommendable, along with the correlation coefficient, to examine the *Scatter plot,* too.

AES - *Average error score* was introduced in ITU-R BS.1387 to implement different requirements for ODG accuracy dependent on the accuracy of SDG. Accuracy of SDG is determined by confidence interval. ***95% confidence interval*** is defined with the following formula ($t_{N,p}$ and S are defined in chapter 5.3.1):

$$\left(SDG - \frac{\bar{S}}{\sqrt{N-1}} \cdot t_{(N-1),0.05}, SDG + \frac{\bar{S}}{\sqrt{N-1}} \cdot t_{(N-1),0.05} \right) \tag{6.4}$$

$$IP = 2 \cdot \frac{\bar{S}}{\sqrt{N-1}} \cdot t_{(N-1),0.05}$$

The basis is the formula for the root mean square error, in which confidence interval IP_i is added:

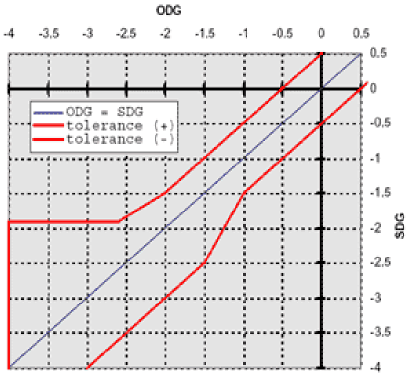
$$AES = 2 \cdot \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\frac{ODG_i - SDG_i}{\max(IP_i, 0.25)} \right)^2} \tag{6.5}$$

AES value range depends on the set of tests on which it is determined. Usually AES values range between 1.5 and 3.0 The smaller the AES, the more accurate ODGs are. AES must not be compared among different sets of tests.

Confidence interval is usually limited so that greater influence of tests, where the interval is of very low value, is avoided. The issue at hand is whether the minimum value of 0.25 is a good limitation value, because according to [THI99] it would be reasonable that the minimal interval corresponds with the precision which is achieved for ODG. The precision that is achieved is +/-0.5, which is 4 times greater than the one which corresponds with the 0.25 interval.

Tolerance scheme defines the allowed error for ODG depending on SDG and the confidence interval values [THI99]:

- If $SDG > -1.5$, tolerance is determined by the confidence interval
- If $SDG < -2.5$, allowed error is two times bigger in value than confidence interval
- For -2.5 < SDG < -1.5, tolerance is the interpolation between two values defined above
- For SDG < -1.9, it is allowed that ODG takes any value smaller than 1.9



6.1 – An example for the tolerance scheme for the interval 0.5 [THI99]

If all the grades shown on the scatter plot are inside a tolerance scheme, the model is considered good enough to be proposed as a standard [THI99].

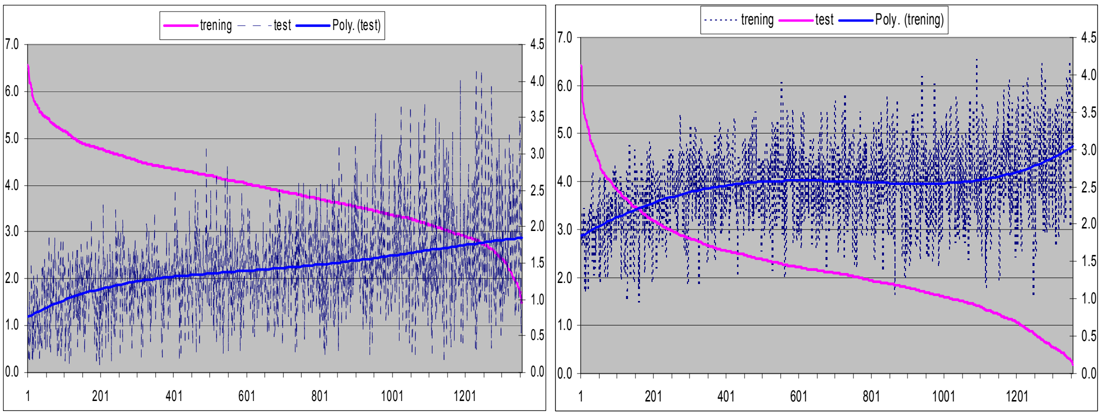
However, none of the analysed models, considered during the development of ITU-R BS.1387, could meet the required criterion. Since this criterion was not treated in ITU-R BS.1387, but was only proposed, it will not be used in the comparison of APEAQ with *Opera*. The scatter plots will be presented without a tolerance scheme, and the reader is left to personally evaluate the accuracy of ODG based on a scatter plot itself.

The number of tests, for which |*ODG-SDG*| is big, is another criterion which is used in ITU-R BS.1387. However, its weekness lies in the determination of how big an error should be to be considered as big enough. In ITU-R BS.1387 two values were selected, 1.0 and 1.5. Instead of this criterion the biggest error will be used. Also, the percentage of *outliers*, and scatter plots will be presented.

Percentage of outliers represents the percentage of tests, for which ODG is outside the confidence interval. Minimum interval is limited to 0.25, the same limit as the one used for AES.

6.2. Results

During the training of the neural network it was noticed that good weight coefficients produce bad results on the validation set and vice versa. Firstly, the training of the neural network for the unmodified APEAQ, implemented according to ITU-R BS.1387, was started. This was necessary for checking whether better weight coefficients can be generated than those given in the standard, because MOVs do not correspond to *Opera* values. However, it was impossible to get better results than those which were produced by the coefficients from ITU-R BS.1387.



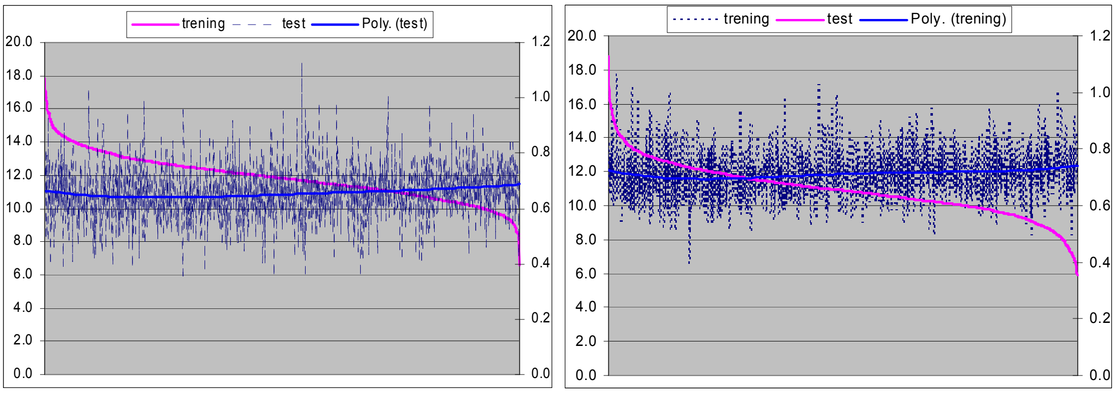
6.2 – The neural network's quality on the training and validation sets for APEAQ implemented acording to ITU-R BS.1387

The graphs show evaluated weight coefficients, which correspond with the local minimums, generated during the training process. The quality is evaluated based on the combination of criteria described in the previous chapter and represents only an approximate evaluation. In the left graph, values were sorted out according to the training set (red line without oscillations), while the values on the validation set have greater oscillations (marked in blue). The trend on the validation set was approximated by the cubic polynomial and is marked by blue line. In the right graph the same values are shown, but red color represents the validation set (on which they were sorted), and blue represents the training set.

It appears that good local minimums for the training set are bad for the validation set and vice versa. This indicates weak generalization ability of the neural network. The reason for this is probably either the fact that the training set is not representative enough or bad input data i.e. MOVs. This indication is different from the specialization on the training set in a local minimum, already described earlier in the text. It has greater significance, because it shows that each improvement on the training set produces bigger impairment on the validation set.

When for the input of a neural network the MOVs in the modified APEAQ are used, the generalization ability substantially increases. Great independency of quality on the validation set from the quality on the training set is achieved, which is shown on the graph 6.3.

Among the weight coefficients, which correspond with the local minimums, the ones that give best results on both validation and training sets were selected. Then the networks with those coefficients were furtherly trained and coefficients that produced best results on the validation set were chosen.



6.3 – The quality of the neural network on the training and validation sets for modified APEAQ

In the following tables there are the results produced by the chosen coefficients, with the comparison to *Opera* and the initial implementation of APEAQ without modification.

PEAQ	SKG	SG	Max.	AES	Correlation	% Outliers
Opera	0.625	0.417	**2.990**	2.501	0.813	49.0
APEAQ	**0.578**	**0.390**	3.288	2.277	**0.835**	48.6
APEAQ modif.	0.598	0.406	3.143	**2.217**	0.834	**47.8**

Table 6.1 – The comparison of APEAQ and Opera on the set of tests from ITU-R BS.1387

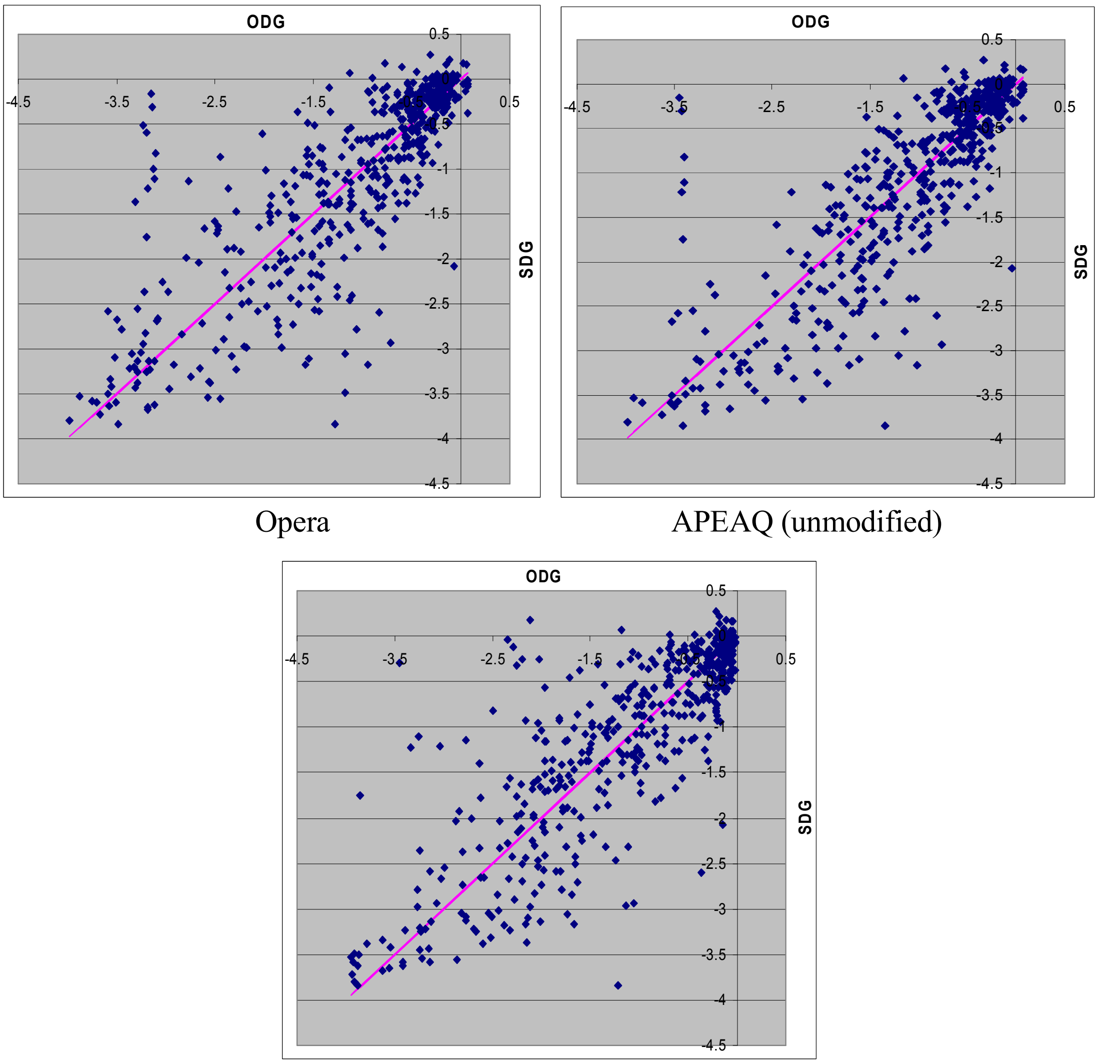
TEST	SKG	SG	Max.	AES	Correlation	% Outliers
Opera	0.533	0.396	2.305	1.462	0.774	37.1
APEAQ	0.568	0.435	**2.282**	1.584	0.752	43.9
APEAQ modif.	**0.531**	**0.384**	2.321	**1.187**	**0.789**	**31.1**

Table 6.2 – The comparison of APEAQ and Opera on the unknown set of tests

Tests MPEG90, MPEG91, MPEG95, ITU92DI, ITU92CO, ITU93 and DB3 were used in the training of the neural network. The comparison of APEAQ with *Opera* on this set is shown in the table 6.1. Tests Extension, Mp3_128, Multiformat and Mares were used, as unknown signals during the training process, for the comparison of the quality of *Opera* and APEAQ, i.e. for the comparison of their generalization ability. The results on the signals from these tests are shown in the table 6.2.

Notably good results of the modified version on the training set were produced as well as excellent results on the unknown set. The advantage of APEAQ, with the implemented improvements in the psychoacoustic model, was also confirmed on the scatter plots 6.5.

Better results on the training set were not achieved probably because of the small range that was chosen for initial random values of the coefficients. However, the training frequently produced bad local minimums with the wider value range. It remains unclear why APEAQ, which was implemented based on ITU-R BS.1387, produces better results than *Opera* on the training set.



Opera

APEAQ (unmodified)

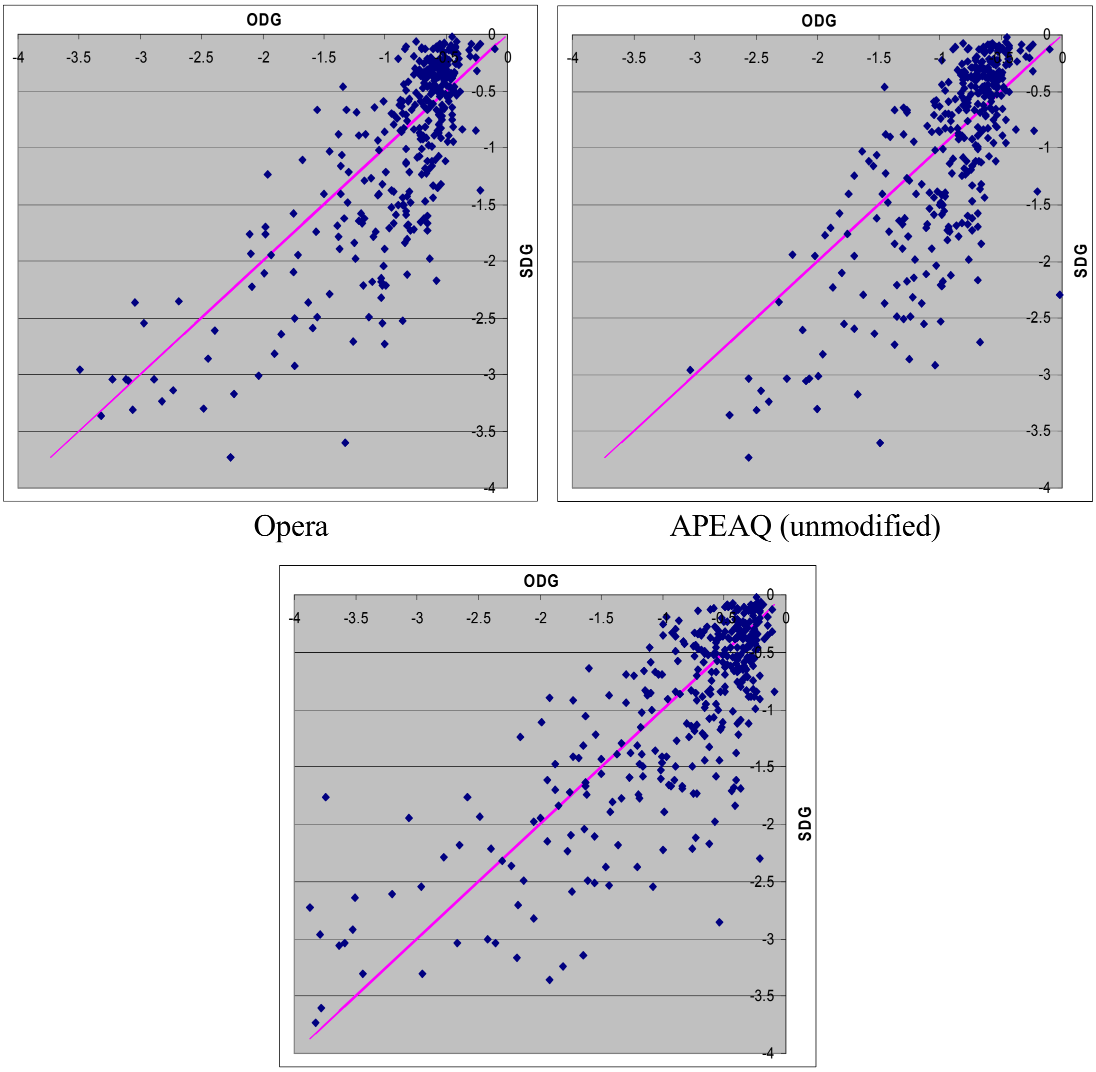
APEAQ (modified)

6.4 – The scatter plot of the training set

On the graphs there is an apparent similarity of *Opera* and unmodified *APEAQ*, which additionally confirms that ITU-R BS.1387 was well-implemented. On the other hand, the values of the most important MOV *RmsNoiseLoudAsym*$_A$ are considerably different (graph 3.15). This may be an indication for the overspecialization of the neural network from ITU-R BS.1387 for the training set.

Modifications clearly lead to completely different outputs. Whether this difference represents an improvement or not, can be checked on the set that was not used for the training process. The training set itself can create a false impression, especially if taking the fact that the neural network from ITU-R BS.1387 is potentially overspecialized for that set.

Only a part of DB3, as it was done in BS.1387, cannot in any case be used as a set for the checking of generalization ability. In the training process if a part of a set is used, then it can be expected that equally good results will be achieved on the rest of the set. If the checking is done on a set from a completely different subjective test, this deficiency does not exist.



Opera

APEAQ (unmodified)

APEAQ (modified)

6.5 – The scatter plot of the unused set in the training process

The similarity between the output of *Opera* and unmodified APEAQ is also obvious on the set of signals which were not used in ITU-R BS.1387. What is also obvious is that there is not a big differentiation between the values of ODG – a lot of signals of similar, yet different SDG values, have almost identical ODG values. Since a large number of signals were graded with over -1, that makes it a lot easier for the grouped values not to have a profound impact on the decrease of numerical quality evaluations (correlation, AES, etc.).

Although linear transformation can greatly improve the quality of unmodified APEAQ, we should bear in mind that ODG, applied in objective evaluation, cannot be transformed based on the knowledge of SDG.

Modified APEAQ again shows considerably different values than the models implemented according to ITU-R BS.1387 standard. ODG is uniformly distributed along the full scale, and errors do not show tendency to depend on the value of SDG.

Especially great improvement of accuracy was achieved in the value range from 0 to 0.5, where the precision of the modified APEAQ is about 0.5 better in value than the initial implementations. This range is especially important for checking the implementations of codecs on digital signal processors, because then very small impairments are expected in the values that need to be defined accurately.

6.3. An example of improvement for testing an encoder implementation

The application of Opera in the implementation of encoders, which was described in chapter 2.2.3, revealed some strange results that stand out in one of the examples. During compression of 48 kHz stereo signal at 32 kbps, extremely poor quality was expected as well as the lowest possible ODG. However, Opera gave for the signal *item4* ODG with -0.965. The remaining 38 signals, all 48 kHz stereo compressed at 32 kbps, had ODG -3.565 and -3.651. While listening to all the signals, it could be heard clearly how much the quality was degraded that exceeded the standards from ITU-R BS.1116 – degradation is obvious and even without listening of the original signals it can be graded with -4.

Apparently, *item4* has the highest values for $RmsModDiff_A$, $SNMR_B$ and EHS_B on the 39 observed signals, and the values $RmsModDiff_A$, $RmsNoiseLoudAsym_A$ and $AvgLinDist_A$ are very close or above the values for $a_{\max}$, that are given in the table 18 in [ITU01]. These values should produce very small DI and ODG, even smaller than on the other signals.

For further checking, the same signals were observed, compressed with the same encoder using a lower degree of compression. For instance, Opera produces ODG between -0.482 and -3.736 at 80 kbps, and *item4* has ODG -3.43, even though it is of better quality than at 32 kbps.

Item	RMD	NLA	ALD	SNMR	EHS	DI	ODG	ODG (Modif. Apeaq)
1	302.416	11.3922	23.6525	-2.52503	2.5429	-2.26975	-3.587	-3.949
2	845.682	15.4044	67.592	-0.66857	1.23492	-2.34464	-3.613	-3.949
3	364.418	18.1471	62.3069	-0.78897	1.97456	-2.29771	-3.596	-3.949
4	**2328.16**	**13.0793**	**42.7305**	**0.232303**	**4.02977**	**0.934389**	**-0.965**	-3.949
5	1037.47	15.0083	65.3284	-0.12783	1.59044	-2.34634	-3.613	-3.949
6	318.921	6.15702	16.4993	-2.59889	2.96678	-2.21164	-3.565	-3.949
7	350.767	12.937	64.753	-1.34553	1.7797	-2.29601	-3.596	-3.949
8	469.688	16.4164	95.3997	-1.74861	1.46638	-2.32562	-3.606	-3.949
9	407.764	6.94366	25.966	-0.60741	1.7712	-2.26821	-3.586	-3.949
10	355.434	13.0338	51.3101	-2.3449	0.921258	-2.3071	-3.600	-3.949
11	501.012	15.857	61.7712	-0.55699	1.06449	-2.32782	-3.607	-3.949
12	631.07	6.46758	6.83463	-8.62481	3.97515	-2.46445	-3.651	-3.769
13	297.76	8.81405	23.179	-3.18433	2.48949	-2.24814	-3.579	-3.949
14	504.871	13.0521	39.9461	-1.32555	2.72586	-2.32646	-3.606	-3.949
15	474.352	17.5839	81.8462	-1.14025	2.28442	-2.32309	-3.605	-3.949
16	314.15	18.926	19.5704	-3.71686	2.91566	-2.29577	-3.596	-3.949
17	426.054	10.2699	38.2428	-1.31148	2.20388	-2.30245	-3.598	-3.949
18	394.896	11.6986	68.9869	-0.743	1.36535	-2.30134	-3.598	-3.949
19	510.88	15.9354	84.8533	-1.09528	1.11731	-2.32983	-3.608	-3.949
20	442.658	23.2784	94.6844	-0.56298	1.53058	-2.31783	-3.603	-3.949
21	502.017	25.4275	118.571	-0.49166	1.01634	-2.32789	-3.607	-3.949
22	643.461	10.7136	57.129	-2.01889	3.48057	-2.33602	-3.610	-3.949
23	362.387	13.313	102.324	-1.90762	1.64632	-2.30494	-3.599	-3.949
24	470.855	10.4755	44.3287	-1.3368	1.943	-2.31552	-3.603	-3.949
25	596.536	10.9273	54.2096	-1.16559	2.16756	-2.33325	-3.609	-3.949
26	360.491	9.43045	32.2872	-1.40086	1.93329	-2.27403	-3.588	-3.949
27	395.931	7.91228	27.9388	-2.7387	2.66702	-2.28429	-3.592	-3.949
28	452.633	20.0675	73.1612	-0.69807	3.08351	-2.31576	-3.603	-3.949
29	382.219	18.1002	91.715	-0.76527	1.42282	-2.30524	-3.599	-3.949
30	469.792	9.88161	34.3289	-0.95775	1.98367	-2.31132	-3.601	-3.949
31	400.218	9.47353	39.6389	-0.44814	1.01696	-2.28949	-3.594	-3.949
32	320.238	5.52254	23.0483	-2.45172	2.22792	-2.20987	-3.565	-3.949
33	338.219	14.69	51.6628	-1.45698	2.12189	-2.29255	-3.595	-3.949
34	354.896	12.2765	32.5504	-1.04585	1.89952	-2.29069	-3.594	-3.949
35	348.909	12.5099	46.6074	-1.42553	2.63915	-2.28862	-3.593	-3.949
36	375.736	18.6252	102.093	-1.5424	1.33462	-2.30776	-3.600	-3.949
37	382.207	13.0357	91.7636	-0.96198	1.64371	-2.30442	-3.599	-3.949
38	296.185	12.1586	29.6931	-1.6335	2.25782	-2.26658	-3.586	-3.949
39	352.462	14.5863	131.1	-2.07754	1.35038	-2.30428	-3.599	-3.949

Table 6.3 – Checking results of the mp3 encoder

Unmodified APEAQ deals with the same problems as *Opera*, so these inconsistencies imply errors in the neural network given in ITU-R BS.1387. So far it has not been noticed that modified APEAQ manifests any of such problems – for the signals compressed by Fraunhofer's encoder at 32 kbps the smallest possible ODG was produced (except for *item12*, where ODG is small enough). Detailed check up on all the outputs of the encoder is not possible, because corresponding SDG values do not exist.

7. Conclusion

In this Master's Thesis methods of objective evaluation of audio quality and their use in an implementation of the encoder on a class of digital signal processors were analysed. ITU-R BS.1387 proposal was implemented with fair accuracy, worthy of the reference *Opticom Opera.*. Implementation called *APEAQ* is about six times faster than *Opera*, which considerably increases the possibility of its realisation, with execution in real time, on a *DSP* platform with limited resources. Unmodified *APEAQ* is even faster than the basic version of *PEAQ* in *Opera*, which was proposed by ITU for applications which require execution in real time. *APEAQ* is realisation of the advanced version, which is more accurate and complex than the basic version.

Existing models and methods, that are used for an evaluation of audio quality, were analysed. *APEAQ* was changed to increase accuracy, which is reflected in greater correlation of *MOVs* with *SDG*. These changes resulted in slower execution, so the modified version proved to be more suitable for use when accuracy is more important than the speed, while the unmodified version comes first if speed is a relevant factor.

It is possible for a certain combination of models and methods, where some *MOVs* do not achieve the highest degree of correlation with *SDG*, to produce highly accurate results. However, the testing of these combinations would require, within the given frame, a recurring training of the neural network with every change. Since such a process takes time, these combinations were not tested. It is also possible that a neural network would cover up flaws of such combinations.

The neural network was trained on a large number of signals, which helped improve its generalization ability. The testing of the neural network on a subset of tests used during the training process proved to be a bad evaluation method of its quality, although that subset was not used in the training process. It is particularly bad if the subset used for testing is a representative sample of a training set and vice versa. This validation method was often used in a development of objective evaluation methods, which consequently raised an issue of the accuracy of quality evaluation given in their descriptions. In this Master's thesis the neural network was tested on a large set of tests, independent from the training set, producing fairly accurate results.

Found weight coefficients of the neural network produced results which were better than those of the initial *APEAQ* and reference *Opera* implementations. On the other hand, these are probably not the best possible coefficients' values. Ultimately, it is impossible to find the best coefficients due to the nature of determination process itself.

Neural network models an average listener, although such a subject does not really exist. One of the possible ways of improving a neural network is the determination of a few coefficients, in other words one neural network per each listener, and then the arithmetic mean of derived *ODGs* would be used. This method would be suitable for real subjective tests, but I must point out the fact that while working on this Master's thesis, the results of subjective tests which would have secure this method's realization, were not available. A step further would have been the determination of the perceptual model for each listener, also requiring complete testing results, which are not publicly accessible.

Neural network can cover up distinctive flaws in other parts of an objective evaluation method, which again can manifest themselves in the testing of certain signals - *outliers*, so it would be better to use some other method for modelling of cognitive processes. This wasn’t made possible, because of the insufficient knowledge about the cognitive processes which take place in a subjective evaluation. Attempts to create objective methods without a neural network can be found in contemporary scientific works, but so far there have not been any results which would prove the advantage of a different way of modelling cognitive processes.

The final results, even though they are better than the initial ones, are not accurate enough, i.e. there is still a need for subjective tests in some cases. The reason for this may be the lack of the implemented models and methods, but it may also be the inconsistency of the results used in the subjective tests.

Nevertheless, as a result of the analysis elaborated in this thesis, the accuracy was improved as well as the increase of speed, in comparison to reference implementation, which subsequently provided better application in the implementation of an encoder on digital signal processors.

8. BIBLIOGRAPHY

[ADV02] Advanced Systems Design and Development : “XlXtrFun™ Extra Functions for Microsoft Excel”, http://www.ozgrid.com/Excel/excel-interpolate-cubic-curve-fit.htm, 2002

[AMO03] Amorim, R. : “128kbps extension Public Listening Test”, public listening test, http://www.rjamorim.com/test/128extension/results.html, 2003.

[AMO04a]Amorim, R. : “MP3 at 128kbps public Listening Test”, public listening test, http://www.rjamorim.com/test/mp3-128/results.html, 2004.

[AMO04b]Amorim, R. : “Multiformat at 128kbps public Listening Test”, public listening test, http://www.rjamorim.com/test/multiformat128/results.html, 2004.

[AMO05] Amorim, R. : “Listening Test – Conduction Handbook”, http://www.rarewares.org/rja/ListeningTest.pdf, 2005.

[BAR05] Barbedo, J.; Lopes, A. : “A New Cognitive Model for Objective Assessment of Audio Quality”, *J. Audio Eng. Soc.*, vol. 53, pp. 22-31, 2005.

[BAU01] Baumgarte, F.; Lerch, A. : “Implementation of Recommendation ITU-R BS.1387”, *Delayed Contribution, Document 6Q/18-E*, 2001.

[BIR05] Birchfield, S.; Gangishetty, R. : “Acoustic Localization By Interaural Level Difference”, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, 2005.

[BRA00] Brandenburg, K.; Popp, H. : “An Introduction to MPEG Layer-3”, *EBU Technical Review*, 2000.

[BRA87] Brandenburg, K. : “Evaluation of quality for audio encoding at low bit rates”, *Contribution to the 82nd AES Convention*, London, 1987.

[BRA92] Brandenburg, K.; Sporer, T. : "NMR and Masking Flag: Evaluation of quality using perceptual criteria", *Proceedings of the 11th international AES Conference on audio test and measurement*, pp. 169-179, Portland, 1992.

[BRA99] Brandenburg, K. : “MP3 and AAC explained”, *Proceedings of the 17th AES International Conference on High Quality Audio Coding*, Signa, Italy, 2003.

[CAR02] Carter, P. : “Structured Variation In British English Liquids: The Role Of Resonance. Appendix 1: The bark scale”, Ph.D. Thesis, *University of York*, York, 2002.

[CIP05] CIPIC Interface Laboratory : “Spatial Sound”, http://interface.cipic.ucdavis.edu/CIL_html/CIL_whatis.htm, 2005

[CON02] Conway, E.; Zhu, Y. : “Applying Objective Perceptual Quality Assessment Methods in Network Performance Modeling”, *11th International Conference on Computer Communications and Networks*, 2002.

[DEE04] Deepak, K. S. K. : “An improved perceptual quality metric for highly to moderately impaired audio”, Master’s Thesis, *New Mexico State University*, Las Cruces, 2004.

[ENE98] Enerstam J., Peman, J. : "Hardware Implementation of MPEG Audio Real-Rime Encoder", Master's Thesis, *Lulea University of Technology*, 1998.

[GAN05] Gan, M. : http://www.rarewares.org/files/others/wavegain-1.2.6.zip, 2005.

[GAR94] Gardner, B.; Martin, K. : "HRTF Measurements of a KEMAR Dummy-Head Microphone", *MIT Media Lab*, Cambridge, 1994.

[GAY03] Gayer, M.; Lohwasser, M.; Lutzky, M. : "Implementing MPEG Advanced Audio Coding and Layer-3 encoders on 32-bit and 16-bit fixed-point processors", *Proceedings of the 115th AES Convention*, New York, 2003.

[GLI82] Glišić, Z.; Peruničić, P. : "Zbirka rešenih zadatak iz verovatnoće i matematičke statistike", *Naučna knjiga*, Belgrade, 1982.

[GOR04] Marković, G.; Lukač, Ž.; Bandić, N.; Đekić M. : "Optimizacija algoritma kvantizacije pri implementaciji MP3 kodera na APX DSP procesoru", XLVIII Konferencija ETRAN, Čačak, 2004.

[GOT03] Gottardi, G. : "Interpretazione e implementazione in ANSI C della raccomandazione ITU-R BS.1387-1 : Perceptual Evaluation of Audio Quality", *Universita Politecnicà Della Marche*, 2003.

[GRU92] Grusec, T.; Thibault, L.; Beaton, R. : "Sensitive methodologies for the subjective evaluation of high quality audio coding systems", *Proceedings of the AES UK DSP conference*, pp. 62-76, London, 1992.

[GRU95] Grusec, T.; Thibault, L.; Soulodre, G. : "Subjective Evaluation of High-Quality Audio Coding Systems: Methods and Results in the Two-Channel Case", *Proceedings of the 99th AES Convention*, New York, 1995.

[HAR99] Hartmann, B. : "How We Localize Sound", *American Institute of Physics*, 1999.

[HYD04] HydrogenAudio forum : "Average Linear Distortion too high, OBJECTIVE aac tests", http://www.hydrogenaudio.org/forums/index.php?showtopic=13590, 2004.

[HYD06] HydrogenAudio forum : "Multiformat Listening Test @ 128 kbps", http://www.hydrogenaudio.org/forums/index.php?showtopic=40607&st=125, 2006

[HYD06a] HydrogenAudio forum : "Analysis by synthesis, split from joint stereo", http://www.hydrogenaudio.org/forums/index.php?showtopic=43039&hl=holes&st=25, 2006.

[HUB03] Huber, R. : "Objective assessment of audio quality using an auditory processing model", Ph.D. Thesis, *University of Oldenburg*, Oldenburg, 2003.

[ILI99] Ilić, V. : "Neuronske mreže", http://solair.eunet.yu/~ilicv/neuro.html, 1999.

[ISO90] "MPEG/Audio test report", Document MPEG90/N0030, *International Oganization for Standardisation*, Geneva, 1990.

[ISO91] "MPEG/Audio test report", Document MPEG91/N0010, *International Oganization for Standardisation*, Geneva, 1991.

[ISO92] "Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s, part 3: Audio", International Standard ISO/IEC 11172-3, *International Oganization for Standardisation*, Geneva, 1992.

[ISO98] "Information Technology – Generic Coding of Moving Pictures and Associated Audio Information, part 3: Audio", International Standard ISO/IEC 13818-3, *International Oganization for Standardisation*, Geneva, 1998.

[ITU01] Recommendation ITU-R BS.1387-1 : "Method For Objective Measurements Of Perceived Audio Quality", *International Telecommunications Union*, Geneva, 2001.

[ITU92] "CCIR Listening Tests – Basic Audio Quality of Distribution and Contribution Codecs, Sweden" , CCIR-Doc. 10-2/24, *International Telecommunications Union*, 1992.

[ITU93] "CCIR Listening Tests – Network Verification Tests without Commentary Codecs, Canada and Italy" , CCIR-Doc. 10-2/43, *International Telecommunications Union*, 1993.

[ITU97] Recommendation ITU-R BS.1116-1 : "Methods For The Subjective Assessment Of Small Impairments In Audio Systems Including Multichannel Sound Systems" , *International Telecommunications Union*, Geneva, 1997.

[JUS89] JUS N.N4.103 : "Elektroakustika : Uređaji audio-sistema, opšti termini, definicije i računske metode", *Savezni zavod za standardizaciju*, 1989.

[KAB04] Kabal, P. : "PQevalAudio", http://www-mmsp.ece.mcgill.ca/Documents/Software/Packages/AFsp/PQevalAudio.html

[KAB06] Kabal, P. : "An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality", *McGill University*, Montreal, 2006.

[KAR85] Karjaleinen, M. : "A new auditory model for the evaluation of sound quality of audio systems," *Proceedings of the ICASSP-85*, pp. 608–611, 1985.

[KEY99] Keyhl, M.; Schmidmer, C.; Watcher, H. : "A combined measurement tool for the objective, perceptual based evaluation of compressed speech and audio signals", *Proceedings of the 106th AES Convention*, Munich, 1999.

[LAI01] Lai, Hung-Chih : "Real-time Implementation of MPEG-1 Layer 3 Audio Decoder on a DSP Chip", Master thesis, *National Chiao-Tung University*, Taiwan, 2001.

[LAM06] LAME mp3 encoder : http://lame.sourceforge.net/

[LER02] Lerch, A. : "EAQUAL – Evaluation of Audio QUALity", http://wiki.hydrogenaudio.org/index.php?title=EAQUAL, 2002.

[MAI97] Maijala, P. : "Better Binaural Recordings Using the Real Human Head", *Helsinki University of Technology*, Helsinki, 1997.

[MAR05] Mares, S. : "Public, Multiformat Listening Test @ 128 kbps", public listening test, http://www.maresweb.de/listening-tests/mf-128-1/results.htm, 2005.

[MEA95] Measer, D.; Kim, S.-W., : “NBC time/frequency module subjective tests: overall results”, MPEG NO973 MPEG95/208, *International Oganization for Standardisation*, 1995.

[MIJ05] Mijić, M. : “Uvod u audiotehniku”, http://telekomunikacije.etf.bg.ac.yu/predmeti/te5aus/, 2005.

[MIL05] Milosavljević, M. : “Neuronske Mreže”, *Elektrotehnički fakultet Beogradskog Univerziteta*, Belgrade, 2005.

[MLA95] Mladenović, P. : “Verovatnoća i statistika”, *VESTA – Matematički fakultet,* Belgrade, 1995.

[NIE93] Nielsen, L. : “A Neural Network Model for Prediction of Sound Quality”, *Technical University of Denmark*, 1993.

[NIE94] Nielsen, L. : “Objective Scaling of Sound Quality for Normal-Hearing and Hearing-Impaired Listeners”, Ph.D. Thesis, *Technical University of Denmark*, 1994.

[OPT06] OPTICOM GmbH : http://www.opticom.de/products/opera.html

[PAI00] Painter, T.; Spanias, A. : “Perceptual coding of Digital Audio”, *Proceedings of the IEEE*, vol. 88, No. 4, pp. 451-513, 2000.

[PAI92] Paillard, B.; Mabilleau, P.; Morissette, S.; Soumagne, J. : “PERCEVAL: Perceptual Evaluation Of The Quality Of Audio Signals”, *J. Audio Eng. Soc.*, vol. 40, pp. 21-31, 1992.

[PAN95] Pan, D. : “A Tutorial on MPEG/Audio Compression”, *IEEE Multimedia Journal*, 1995.

[PRO01] Project P905: “AQUAVIT - Assessment of Quality for audio-visual signals over Internet and UMTS”, *British Telecommunications plc, Telecom Italia S.p.A., Deutsche Telekom AG*, 2001.

[RAD91] Radunović, D. : “Numeričke metode”, Građevinska knjiga, Belgrade, 1991.

[RAI02] Raissi, R. : “The Theory Behind MP3”, http://www.rassol.com/cv/mp3.pdf, 2002.

[REI04] Reiss, J.; Sandler, M. : “Audio Issues in MIR Evaluation”, *Proceedings of the 5th International ISMIR 2004 Conference*, Barcelona, 2004.

[ROB01] Robinson, D. : “Replay Gain - A Proposed Standard”, http://replaygain.hydrogenaudio.org/, 2001.

[ROB02] Robinson, D. : “Perceptual model for assessment of coded audio”, Ph.D. Thesis, *University of Essex*, Essex, 2002.

[SAL03] Salomonsen, K.; Sogaard, S.; Larsen, E. P. : “Design and Implementation of an MPEG/Audio Layer III Bitstream Processor”, Master’s Thesis, *Aalborg University*, 2003.

[SAR02] Sarle, W. : “comp.ai.neural-nets FAQ”, http://www.faqs.org/faqs/ai-faq/neural-nets/part1/preamble.html, 2002.

[SCHR79] Schroeder, M. R.; Atal, B. S.; Hall, J. L. : “Objective measure of certain speech signal degradations based on masking properties of human auditory perception”, *Frontiers of Speech Communication Research*, pp. 217-229, New York, 1979.

[SHLI96] Shlien, S.; Soulodre, G. : “Measuring the Characteristics of ‘Expert’ Listeners”, *Proceedings of the 101th AES Convention*, Los Angeles, 1996.

[SMI99] Smith, S. W. : “The Scientist and Engineer's Guide to Digital Signal Processing”, *California Technical Publishing*, San Diego, 1999.

[SOU98] Soulodre, G. A.; Grusec, T.; Lavoie, M.; Thibault, L. : “Subjective Evaluation of State-of-the-Art 2-Channel Audio Codecs”, *J. Audio Eng. Soc.*, vol. 46, pp. 164-176, 1998.

[SPO95] Sporer, T.; Gbur, G.; Herre, J.; Kapust, R. : "Evaluating a Measurement System", *J. Audio Eng. Soc.*, vol. 43, pp. 353 – 362, 1995.

[STA06] “Electronic Statistics Handbook”, StatSoft Inc., http://www.statsoft.com/textbook/stathome.html, 2006.

[STA99] “Glossary of Testing, Measurement, and Statistical Terms”, *Riverside Publishing*, Itasca, 1999.

[SAL04] Šalovarda, M.; Bolkovac, I.; Domitrović, H. : “Comparison of audio codecs using PEAQ algorithm”, *Faculty of Electrical Engineering and Computing, University of Zagreb*, Zagreb, 2004.

[TER79] Terhardt, E. : “Calculating Virtual Pitch”, *Hearing Research*, vol. 1, pp. 155-182, 1979.

[THI00] Thiede, Th.; Treurniet, W. C.; Bitto, R.; Schmidmer, C.; Sporer, T.; Beerends, J. G.; Colomes, C.; Keyhl, M.; Stoll, G.; Brandenburg, K.; Feiten, B. : “PEAQ – The ITU Standard for Objective Measurement of Perceived Audio Quality”, *J. Audio Eng. Soc.,* vol. 48, pp. 3–29, 2000.

[THI96] Thiede, Th.; Kabot, E. : “A New Perceptual Quality Measure For Bit Rate Reduced Audio”, *Proceedings of the 100th AES Convention*, Copenhagen, 1996.

[THI99] Thiede, Th. : “Perceptual Audio Quality Assessment Using a Non-Linear Filter Bank”, Ph.D. Thesis, *Technical University of Berlin*, Berlin, 1999.

[VUK97] Vuković, N. : “Verovatnoća i statistika”, http://statlab.fon.bg.ac.yu/srb1/knjige/knjiga.html, Belgrade, 1997.

[WIK06] Wikipedia: “Student’s t-distribution”, http://en.wikipedia.org/wiki/Student'27s_t-distribution, 2006.

[WIL00] Treurniet, W. C.; Soulodre, G. A. : “Evaluation of the ITU-R Objective Audio Quality Measurement Method”, *J. Audio Eng. Soc.,* vol. 48, pp. 164–173, 2000.

[ZWI67] Zwicker, E.; Feldtkeller, R. : “Das Ohr als Nachrichtenempfänger”, *Hirzel Verlag*, Stuttgart, 1967.

[ZWI91] Zwicker, E.; Zwicker, U.T. : “Audio Engineering and Psychoacoustics: Matching Signals to the Final Receiver, the Human Auditory System”, *J. Audio Eng. Soc.*, vol. 39, pp. 115-126, 1991.