Perceptual Audio Quality Assessment using a Non-Linear Filter Bank

(Gehörbezogene Qualitätsbewertung von Audiosignalen unter Verwendung einer nichtlinearen Filterbank)

vorgelegt von Dipl. Ing. Thilo Thiede

Vom Fachbereich Elektrotechnik der Technischen Universität
Berlin zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
- Dr. Ing. –

genehmigte Dissertation

Berlin 1999

D 83

Tag der Abgabe: 27.10.1998
Tag der wissenschaftlichen Aussprache: 19.04.1999

Promotionsausschuß:
Vorsitzender: Prof. Dr.-Ing. Heinrich Klar
1. Berichter: Prof. Dr.-Ing. Peter Noll
2. Berichter: Prof. Dr.-Ing. Manfred Krause

Abstract

Perceptual Audio Quality Assessment using a Non-Linear Filter Bank

This thesis describes a new method for the objective measurement of perceived audio quality. The method is based on a non-linear filter bank which provides a good approximation of auditory filter shapes and even models the level dependence of these filter characteristics. Unlike other measurement schemes, the quality estimation is not solely based on models for steady-state signals, but considers also the temporal structure of the envelopes of the auditory filter outputs. A further improvement compared to other measurement methods is a separation between linear and non-linear distortions. This takes into account the fact that imbalances in the frequency response of an audio device are less annoying than the same amount of non-linear distortions like, for example, quantisation noise. The computational complexity of the filter bank implemented in this method is lower than for most other filter banks applicable to perceptual measurement. The method has proven to be superior to most other measurement methods used in this field and a large part of it will be included in the ITU-recommendation „*method for objective measurements of perceived audio quality*“. Especially the part of this recommendation that addresses applications requiring maximum possible accuracy („*advanced version*“) is mainly based on this method.

Zusammenfassung

Gehörbezogene Qualitätsbewertung von Audiosignalen unter Verwendung einer nichtlinearen Filterbank

Die vorliegende Arbeit beschreibt ein neues Verfahren zur gehörbezogenen Qualitätsbewertung von Audiosignalen. Das Verfahren verwendet eine nichtlineare Filterbank, die eine gute Nachbildung der im Gehör vorliegenden Filtercharakteristiken liefert und dabei auch die Pegelabhängigkeit der Frequenzauflösung des Gehörs berücksichtigt. Im Gegensatz zu anderen Meßverfahren basiert das verwendete Gehörmodell nicht mehr ausschließlich auf Modellen für abschnittsweise stationäre Signale, sondern berücksichtigt auch die zeitliche Struktur der Hüllkurven der gefilterten Signale. Ein weiterer Vorteil gegenüber anderen Verfahren liegt in der vorgenommenen Trennung zwischen linearen und nichtlinearen Verzerrungen, durch die der Umstand berücksichtigt wird, daß Fehler, die durch einen ungleichmäßigen Frequenzgang entstehen, meist sehr viel weniger störend sind als nichtlineare Verzerrungen wie z. B. Quantisierungsrauschen. Die verwendete Filterbank erfordert im Vergleich zu anderen in gehörangepaßten Meßverfahren verwendbaren Filterbänken einen relativ geringen Rechenaufwand. Die neue Meßmethode ist den meisten anderen Meßverfahren für diesen Anwendungsbereich deutlich überlegen, und wird zu großen Teilen in die ITU-Empfehlung zur objektiven Messung der wahrgenommenen Tonqualität („*method for objective measurement of perceived audio quality*“) eingehen. Insbesondere der Teil der ITU-Empfehlung, der sich auf Anwendungen bezieht, für die eine maximale Genauigkeit erforderlich ist, besteht zum überwiegenden Teil aus dem hier vorgestellten Verfahren.

Contents

ABSTRACT **3**

ZUSAMMENFASSUNG **5**

CONTENTS **7**

1. INTRODUCTION **11**

2. FUNDAMENTAL PRINCIPLES OF PERCEPTUAL MEASUREMENT **13**

- 2.1 SOME CHARACTERISTICS OF THE HUMAN AUDITORY SYSTEM 13
  - 2.1.1 Masking 13
  - 2.1.2 Loudness Perception 18
  - 2.1.3 Auditory Frequency Scales 20
  - 2.1.4 Other Effects 21
- 2.2 PSYCHOACOUSTICAL MODELS 22
  - 2.2.1 Zwicker’s Model for the Calculation of Perceived Loudness 23
  - 2.2.2 Moore’s Model for the Calculation of Partial Loudness 24
  - 2.2.3 Analytical Expressions for Psychoacoustical Phenomena 26
  - 2.2.4 Summary and Conclusions 32
- 2.3 CONCEPTS OF PERCEPTUAL MODELS 34
  - 2.3.1 Masked Threshold Concept 35
  - 2.3.2 Comparison of Internal Representations 36
  - 2.3.3 Analysis of Linear Error Spectra 37
- 2.4 PERCEPTUAL MEASUREMENT METHODS 37
  - 2.4.1 NMR 39
  - 2.4.2 PAQM 39
  - 2.4.3 PERCEVAL 40
  - 2.4.4 POM 40
  - 2.4.5 OASE 40
  - 2.4.6 Comparison of Different Concepts 41
  - 2.4.7 Shortcomings of Existing Models 45

3. APPLICATION OF FILTER BANKS **46**

- 3.1 COMPARISON WITH FFT-BASED METHODS 46
  - 3.1.1 Temporal Resolution 47
  - 3.1.2 Spectral Resolution 47
  - 3.1.3 Summary 48
- 3.2 REQUIREMENTS ON THE FILTERS 48
- 3.3 FILTER BANKS USED IN PERCEPTUAL MODELS 49
  - 3.3.1 One-Third-Octave Filters 49
  - 3.3.2 FIR Filters 49
  - 3.3.3 BARK-Transform 50
  - 3.3.4 IIR Filters 51
  - 3.3.5 Warped Filters 52
  - 3.3.6 FTT (Fourier-Time-Transform) 55
  - 3.3.7 Gammatone Filter Banks 56
- 3.4 WAVELET-TRANSFORMS 57
- 3.5 DESIGN OF A NON-LINEAR FILTER BANK 58
  - 3.5.1 Modelling Level-Dependent Excitation Patterns 58
  - 3.5.2 FIR Filters using Recursive Algorithms 61
- 3.6 COMPUTATIONAL EFFICIENCY 68

3.7 FLEXIBILITY 70
3.8 CONCLUSIONS 70

4. DIX - A NEW PERCEPTUAL MEASUREMENT METHOD 71

4.1 OBJECTIVES AND REQUIREMENTS 72
4.2 OVERVIEW 73
4.3 PERIPHERAL EAR MODEL 74
  4.3.1 Structure of the Filter Bank 74
  4.3.2 Pre-Filtering and Scaling 76
  4.3.3 Sampling in the Time and Frequency Domain 77
  4.3.4 Rectification 79
  4.3.5 Time Domain Smearing 80
  4.3.6 Subsampling 80
  4.3.7 Threshold in Quiet 81
  4.3.8 Characteristics of the Filter Bank 82
4.4 ALIGNMENT AND ADAPTATION 88
  4.4.1 Introduction 88
  4.4.2 Time Alignment 89
  4.4.3 Dynamical Level and Pattern Adaptation 89
4.5 EVALUATION OF ENVELOPE MODULATIONS 93
4.6 MODEL OUTPUT VALUES 95
  4.6.1 Measures for Non-Linear Distortions 95
  4.6.2 Measures for Linear Distortions 98
  4.6.3 Measures for Changes in the Temporal Structure 99
  4.6.4 Temporal Averaging 101
  4.6.5 Selection of Valid Sequences of the Audio Signal 102
4.7 OPTIMISATION OF THE MODEL 103
  4.7.1 Auditory Frequency Scales 103
  4.7.2 Sampling in the Time and Frequency Domain 104
  4.7.3 Simultaneous Masking 104
  4.7.4 Temporal Masking 105
  4.7.5 Dynamical Level and Pattern Adaptation 105
4.8 SUMMARY 106

5. PEAQ - THE NEW STANDARD FOR OBJECTIVE MEASUREMENT OF PERCEIVED AUDIO QUALITY 107

5.1 INTRODUCTION: STANDARDISATION WITHIN THE ITU 107
5.2 COMBINING OUTPUT VALUES OF DIFFERENT MODELS 108
  5.2.1 Advanced Version 108
  5.2.2 Basic Version 108
5.3 DESCRIPTION OF THE COMBINED MODEL 109
  5.3.1 Outline 109
  5.3.2 FFT-Based Ear Model 111
  5.3.3 Filterbank-based Ear Model 113
  5.3.4 Model Output Variables 113
  5.3.5 Temporal and Spectral Averaging 118
5.4 OPTIMISATION OF THE MODEL 118
  5.4.1 Forward Masking 118
  5.4.2 Dynamic Level and Pattern Alignment 118

6. ESTIMATION OF PERCEIVED BASIC AUDIO QUALITY 119

6.1 GRADING SCALES FOR SUBJECTIVE AUDIO QUALITY ASSESSMENT 119
  6.1.1 The Five-Grade Impairment Scale 119
  6.1.2 SDG (“Subjective Difference Grade”) 120
  6.1.3 ODG (“Objective Difference Grade”) 120
6.2 ACCESSIBLE DATABASES 120
  6.2.1 Database 1 121
  6.2.2 Database 2 121

6.2.3 Database 3 121
6.2.4 Other Databases 121
6.3 MAPPING FROM MODEL OUTPUT VALUES TO OBJECTIVE DIFFERENCE GRADES 122
6.3.1 Polynomial Mapping Functions 122
6.3.2 One-Dimensional Mapping Functions 122
6.3.3 Multidimensional Mapping Functions 123
6.4 APPLICATION OF AN ARTIFICIAL NEURAL NETWORK 125
6.5 DEFINITION OF A DISTORTION INDEX 126

7. PERFORMANCE OF THE MEASUREMENT METHOD 127

7.1 APPLIED CRITERIA 127
7.1.1 Correlations between Model Predictions and Subjective Gradings 127
7.1.2 Average Prediction Error 128
7.1.3 Absolute Error Scores 128
7.1.4 Tolerance Scheme 128
7.2 PREDICTIONS OF COMPARATIVE TESTS AMONG AUDIO CODECS 129
7.3 ITU-R COMPARATIVE TEST 1996 131
7.3.1 Results of the First Phase 131
7.3.2 Second Phase: Fitting the Models to the Database 133
7.3.3 Results of the Third Phase 133
7.3.4 Conclusions 133
7.4 ITU-R VALIDATION TEST 1997 134
7.4.1 Optimisation and Preselection 134
7.4.2 Results of the First Phase 136
7.4.3 Second Phase: Fitting the Models to the Database 137
7.4.4 Results of the Third Phase 139
7.5 SUMMARY 146

8. SIMULATING PSYCHOACOUSTICAL EXPERIMENTS WITH PEAQ 147

8.1 PSYCHOACOUSTICAL MODELS AND PERCEPTUAL MEASUREMENT 147
8.2 SIMULTANEOUS MASKING 147
8.2.1 Masking Properties of Tones and Noises 147
8.2.2 Additivity of Masking 151
8.3 TEMPORAL MASKING 152
8.4 SUMMARY 153

9. OUTLOOK 154

10. SUMMARY 155

11. REFERENCES 156

12. ACKNOWLEDGEMENT 160

13. REMARKS 160

14. APPENDIX 161

14.1 INDEX 161
14.2 TABLE OF FIGURES 163
14.3 ABBREVIATIONS 166

1. Introduction

The quality of an audio signal which has been subjected to any kind of processing is determined by the subjective impression that a normal hearing person perceives when listening to the processed signal. A purely technical measure can give only a very rough estimate of the audio quality as long as it does not take into account characteristics of the human auditory system.

The most important phenomenon in human hearing, with respect to processing and measurement, is the occurrence of *masking*. When two signals are located sufficiently close to each other both in time and frequency, the weaker signal may become inaudible due to the presence of the stronger signal. The signal level up to which signal components are inaudible due to masking is called the *masking threshold*. Masking occurs between different components of the original signal as well as between the original signal and introduced distortions. In the first situation, the masked parts of the original signal need not to be processed. In the second situation, the masked distortions do not influence the perceived quality, and thus should not be taken into account by a measurement device.

Certain audio signal processing methods, so-called *perceptual coders*, make extensive use of masking phenomena and therefore must also consider masking in the measurement of the processed signals. Such perceptual coders perform a drastic irrelevance reduction to achieve a high coding gain. Signal components that are assumed to be imperceptible are not transmitted, and the coding noise is spectrally shaped according to the masking threshold of the audio signal. Traditional quality measures (e. g. signal to noise ratio or harmonic distortions), which cannot separate these inaudible artefacts from audible distortions, cannot be used to assess the performance of these coders. The only reliable method to assess the basic audio quality of perceptually coded signals have been subjective listening tests. Because listening tests are very time consuming and expensive, there has been a strong demand for new measurement methods which are capable of yielding a reliable estimate of the perceived audio quality.

Measurement methods designed for this purpose model parts of human auditory perception and are therefore called *perceptual measurement methods*. They detect and assess audible artefacts by comparing the output of the codec with the original signal. The development of perceptual measurement methods started almost at the same time as the development of perceptual codecs. The first perceptual measurement method was published in 1979 [SCH79], and has been used as a tool in the development of speech codecs. The principles of such methods had already been developed more than a decade earlier (even before the use of digital computers became practical) in the measurement of perceived loudness [FIS64]. In the field of audio coding, perceptual measurement methods have been introduced several years later with the development of the NMR (noise-to-mask ratio) measurement tool [BRA87].

At this time, the main focus was set on the application of the measurement results for the improvement of audio codecs. Consequently, the psychoacoustical models were restricted to concepts applicable to coding, and the absolute accuracy was less important than worst case considerations. Partly as a result of the standardisation of audio codecs (e. g MPEG 1 and 2, Layer I-III) in the 90s, the interest in perceptual measurement has increased, and several new measurement methods have been published ([BEE92], [PAI92], [COL93] etc.). The psychoacoustical models incorporated in these methods are more detailed than those used in coding. The application of the measurement methods is not restricted to codec development anymore, but also includes the estimation of the perceived overall quality of coded signals.

The most obvious shortcoming of existing perceptual measurement methods is the limited temporal resolution and, as a result, the neglect of temporal effects in the perception of distortions. Before the start of the current work (1994), there was a small number of proposals for perceptual measurement methods with enhanced temporal resolution, but these methods were still based on models for steady-state signals. Furthermore, because of their high computational complexity, the performance of these models was never verified against a sufficiently large set of real world data.

The new measurement method provides not only a perceptually adapted temporal and spectral resolution but also evaluates fluctuations in the temporal envelopes of each signal component. Therefore, it is probably the first measurement method for the assessment of audio codecs that is based on an auditory model that goes beyond the perception of steady-state signals. Unlike earlier models with enhanced temporal resolution, the computational complexity is low enough to allow for a validation of the method against a large database of test signals.

2. Fundamental Principles of Perceptual Measurement

The goal of perceptual measurement is to produce a technical measure for a perceptible quantity that corresponds to the intensity of the sensory perception evoked by this quantity. In the context of this work, perceptual measurement is always related to auditory perception. Nevertheless, this term may also be related to other kinds of perception like visual or tactual sensation. Perceptual measurement on audio signals requires a model of the human auditory system. Such a model can reflect the way sounds are processed by the human auditory system, but it is more important that it reflects the way sound events are perceived by a normal listener (the first approach implicitly includes the latter, but not vice versa).

This chapter will give a brief introduction into human auditory perception, followed by an overview on existing perceptual measurement methods.

2.1 Some Characteristics of the Human Auditory System

The typical application of a perceptual measurement method is to predict whether distortions introduced by a signal processing device (normally a perceptual codec) are audible, and, if so, how annoying these distortions are. Such predictions will only be possible if there is an unambiguous relationship between measurable characteristics of an audio signal and the way it is perceived by a normal hearing listener. This relationship should not depend on individual preferences of the listeners, but must be based on physiological properties of the human auditory system. Several investigations, for example the works of Zwicker [ZWI67], [ZWI90], have shown that many aspects of human auditory perception are in fact almost independent of individual preferences. The most important ones are the occurrence of masking, the perception of loudness, and the perception of pitch. These characteristics of the human auditory system can be modelled in order to get an objective estimate of perceived audio quality. Most of these characteristics can be approximated by analytical expressions which, for example, have been proposed in the works of Terhardt ([TER79], [ZWI80], [TER92]).

The listener-independent characteristics of auditory perception form a low-level model of the human auditory system. Besides the works of Zwicker, the works of Moore [MOO89] also include descriptions for most aspects of auditory perception, which form a slightly different model of the auditory system. Even though the results of the experiments carried out by Moore et al. are often considered to correspond better to the physiological structure of the auditory system, the model proposed by Zwicker has proven to work rather well when applied to perceptual coding and perceptual measurement.

2.1.1 Masking

The limited spectral and temporal resolution of the ear in combination with a limited dynamic range produces a phenomenon called *masking*. When two signals are sufficiently close to each other both in time and frequency, the weaker signal may become inaudible due to the presence of the stronger signal. The signal component that is masked is called *maskee* and the signal component that masks another one is called *masker*. Even though the correct way of looking at masking phenomena would be in the time-frequency plane, it is very usual to look at masking as two separate effects - one in the frequency domain and one in the time domain. If masking only depends on the location in the frequency domain, i. e. masker and maskee are presented at the same time, it is often called *simultaneous masking*. If masking depends mainly on the location in the time domain, i. e. masker and maskee have a similar spectral shape, it is called *temporal masking*. Temporal masking is again looked at as two separate effects: *forward masking* and *backward masking*. In the case of forward masking (also: *post-masking*), signal components are masked after termination of the masker, and in the case of backward masking (also: *pre-masking*), signal components are masked before the onset of the masker. The energy level below which a signal component is masked by other signal components is either called *masked threshold* (which implies that masking is looked at from the side of the maskee) or *masking threshold* (which implies that masking is looked at from the side of the masker). Both terms are equivalent. Apart from the location of masker and maskee in the time-frequency plane, the masking threshold in the case of forward masking also depends on masker duration.

Whereas simultaneous masking and forward masking are easy to understand, backward masking seems to be a more complicated phenomenon, because it implies, that a loud signal can mask another signal before the former one is actually present. It is usually explained by the assumption that loud signals are processed faster than weak signals and may therefore overtake the maskee during the processing of the signal, either on the auditory nerve or later on in the higher levels of the auditory system. However, this assumption is not necessary for the explanation of backward masking.

Masking can generally be explained by looking at the auditory system as a signal analyser with a (naturally) finite spectral and temporal resolution, determined by the filters used, and a finite accuracy, determined by the resolution in the presentation of the signal levels. A signal component can only be detected when the difference between the entire signal (masker plus maskee) and the signal without the

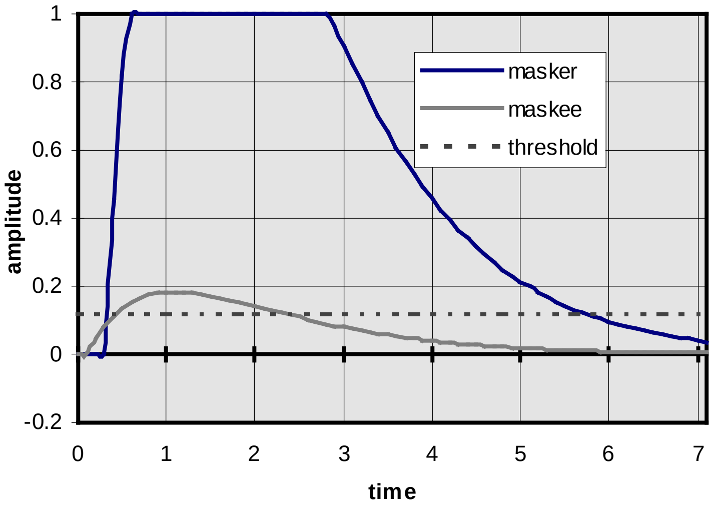


Fig. 2.1: Explaining backward masking by filter shapes and threshold. Even though the maskee starts before the masker, at the point where it exceeds absolute threshold the masker is already larger than the maskee.

maskee at any location in the time-frequency plane is larger than the amplitude resolution of the system. As the filter response to the maskee will need a certain time until it reaches the absolute threshold, masking can already occur when the onset of the masker starts within this time period (Figure 2.1). In general, the shape of the masking functions in such a model is determined by the shape of the spectral and temporal resolution function of the filters.

An addition to this model is the assumption of temporal (and spectral) integration. Even if the difference between two signals never exceeds the threshold given by the accuracy of the system at any isolated location, it can still be detected when averaging over a certain region of the time-frequency plane. This would reduce the resolution of the system but considerably lower the threshold of detectability, and it is very likely that the auditory system performs a similar processing. When incorporating temporal integration, the dependence of the masking functions on time-frequency resolution and integration time becomes more complex. Instead of the local values of masker and maskee, the areas below a certain time period of masker and maskee (the integration time) determine the masked threshold. Even though it is not as obvious as in the above model, this model also explains the occurrence of all kinds of spectral and temporal masking, and it even works without any assumptions about the absolute threshold and the ascent of the temporal filter shapes.

Among the three categories of masking, simultaneous masking has been examined most frequently and in more detail than temporal masking effects. The details of temporal masking seem also to be of minor importance for perceptual coding (even though transparent audio coding would be impossible without temporal masking). Measuring temporal masking is more difficult than measuring simultaneous masking because the two dimensions of the time-frequency plane are not really equivalent: whereas the frequency axis is limited to a well defined range, the time axis is virtually infinite. The absolute position on the time axis should thus not have any influence on frequency domain effects, whereas the absolute position on the frequency axis clearly influences temporal effects. Hence, simultaneous masking can be measured independently of time domain effects, whereas temporal masking cannot be separated from spectral effects. As a result, the measurement of temporal masking needs to be carried out for different centre frequencies, and the effort of such a listening experiment is rather high. Additionally, temporal masking cannot be measured without making assumptions about its origin and it will always be hard to decide whether detection is based on temporal or on spectral processing in the ear[1]. Furthermore, it requires test signals with a very well defined location in both time and frequency domain. Due to the Heisenberg relation this is only possible up to a limited extent. In the case of forward masking, this is not really a problem because the observed time constants are large enough (ca. 100 ms) to allow test signals with a sufficiently tight spectrum without introducing too much uncertainty in the temporal structure. In the case of backward masking, the observed time constants are so small (between one and 20 ms) that it cannot be measured reliably for narrow band signals.

---

[1] To be more precise: it actually can be measured, but it is not necessarily linked to the temporal resolution of the auditory system.

a) Masking Thresholds of Tones, Noises and Pulses

Masking thresholds are usually measured either as a function of time or of the centre frequency of the maskee, whereas level and centre frequency of the masker are held constant and are taken as parameters. Such *masking curves* are for example found in [FAS76] for temporal masking (Figure 2.2) and in [ZWI67] for simultaneous masking (Figure 2.3). These masking curves can be approximated by two-sided exponentials when represented as energies over a perceptual frequency scale (usually the *critical band scale*, which will be described in Section 2.1.3). The low frequency slope is very steep and depends only slightly on the masker level. The high frequency slope is much flatter and strongly depends on masker level. At low levels it is almost as steep as the low frequency slope, whereas it becomes almost flat at very high masker levels.

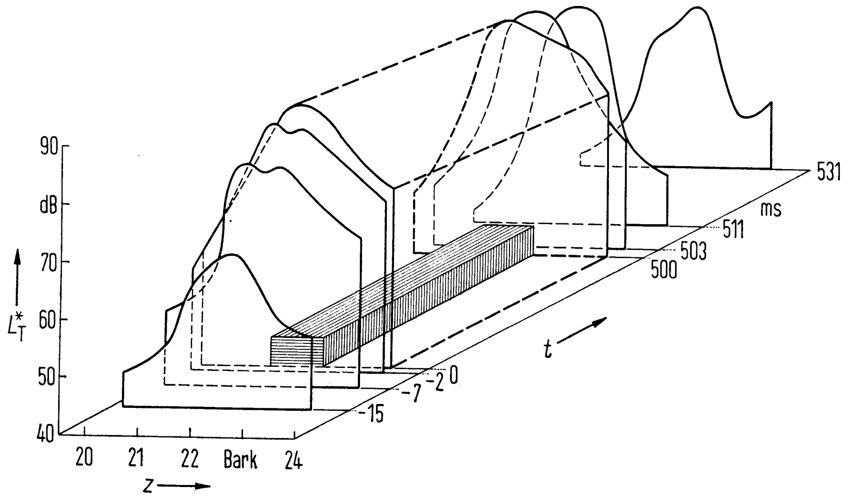


Fig. 2.2: Masking pattern for a narrow band noise pulse (from [FAS76]). The unit Bark corresponds to the critical band scale that will be described in Section 2.1.3.

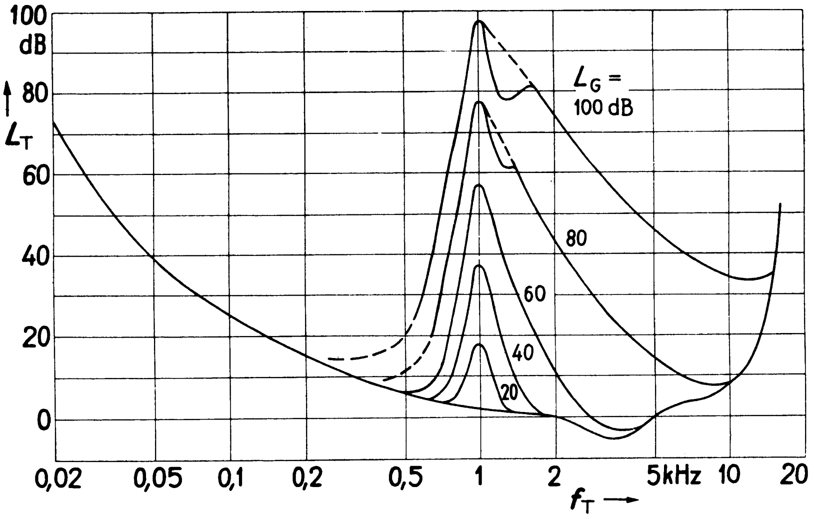


Fig. 2.3: Masking curves for simultaneous masking of tones by narrow band noise (from [ZWI67]).

The level difference between a signal and the maximum of the masking threshold it produces is called *masking index* [ZWI90]. It depends on the centre frequency of the masker but is assumed to be independent from the signal level. The linear representation of the masking index is called *threshold factor* [ZWI90].

When the roles of masker and maskee are interchanged in a masking experiment, i. e. when the maskee is held constant and the required masker level is measured for different frequencies of the masker, the resulting curves are called *psychophysical tuning curves*. They correspond to *neural tuning curves* which are derived from a similar experiment carried out by direct measurement of neural excitations instead of listening experiments [MOO89]. If masking is assumed to occur at a certain ratio between masker and maskee at one location on the basilar membrane, tuning curves can be interpreted as the level-inverted frequency response of virtual filters, the so-called *auditory filters*. Thus, masking curves, tuning curves and auditory filter shapes are different representations of the same effect. Without the level dependence of masking, auditory filter shapes would correspond to frequency-inverted masking curves (Figure 2.4).

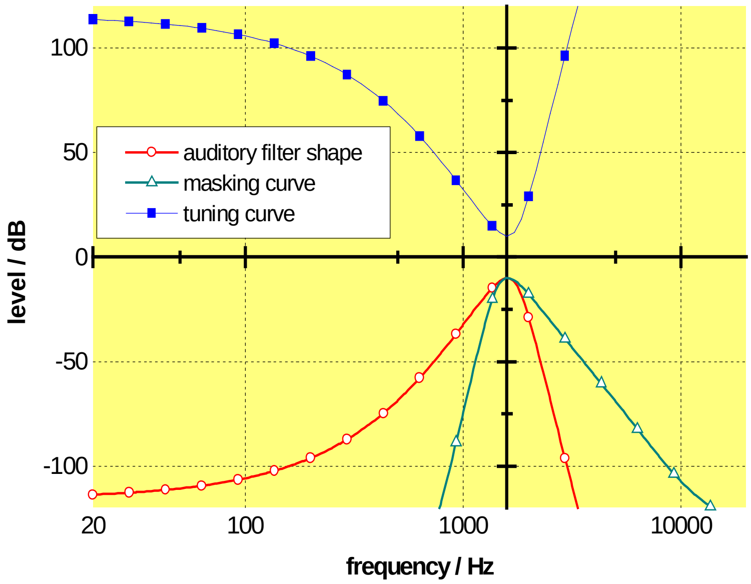


Fig. 2.4: Correspondence between masking curves, tuning curves and auditory filter shapes.

The amount of masking produced by a noise-like stimulus is typically much higher than the amount of masking produced by a tonal stimulus. This effect is known as the *asymmetry of masking* and has for example been described in [HEL72]. In the case when a pure tone is masked by a narrow band noise, the masking index is between -2 and -6 dB [ZWI67]. In the reversed constellation, the masking index can be lower than -20 dB and strongly depends on masker frequency. According to [SCH79] it can be approximated by the equation[2]

---

[2] This approximation was developed in the context of speech coding. Therefore, it should only be used for a frequency range as used in speech coding (i. e. below 3.4 kHz). At higher frequencies, this equation would result in a masking index of down to -40 dB, which appears to be unrealistic.

$$S/dB = -15.5 - z/Bark \quad \Big| \quad 0 \le z \le 25, \qquad (2.1)$$

where $z / Bark$ indicates the position on the *critical band scale* (see Section 2.1.3).

Masking between pure tones is difficult to measure because the detection threshold is mostly determined by beats and combination tones. When modelling masking between different signal types, usually only the type of the masker is considered but not the type of the maskee.

b) Additivity of Masking

If masking is produced by several signal components located at different positions in the time-frequency plane, the estimation of the resulting masked threshold is rather difficult. The most obvious assumption would be that the overall masking threshold is either given by the sum or by the maximum of the masking thresholds produced by each individual signal component. These assumptions have not been corroborated by listening experiments [HUM89]. In fact the masking threshold produced by a multi-component signal is much higher than the sum of the thresholds produced by the individual components. For certain signal constellations, the dependence between the thresholds produced by the individual signal components and the threshold produced by the complete signal can be approximated by a power law:

$$thres_{total} = \left(\sum thres_{individual}{}^{\alpha}\right)^{\frac{1}{\alpha}} \Bigg|_{\alpha \approx 0.2} \qquad [HUM89] \qquad (2.2)$$

Nevertheless, there is still no general model of this phenomenon[3].

2.1.2 Loudness Perception

a) Phon Scale and Equal Loudness Contours

The subjectively perceived loudness of an audio signal depends not solely on its sound pressure level but also on other signal characteristics like for example its frequency distribution. This effect is taken into account by replacing the decibel scale by the so-called *phon scale*. The loudness of a signal given in the unit *phon* corresponds to the sound pressure level in decibels of a pure tone at 1 kHz which produces the same perceived loudness as the measured signal.

When the sound pressure level of a pure tone that produces a certain loudness is plotted as a function of frequency, the resulting curves are called *equal loudness contours* (see Figure 2.5). The equal loudness contours for a sound pressure level of 0 dB and 60 dB at 1 kHz are of particular interest: the former corresponds to the *absolute threshold* of hearing (also: *threshold in quiet*) and the latter represents the frequency weighting at intermediate listening levels.

---

[3] In the power law proposed in [HUM89] the exponent had to be fitted to each individual experiment.

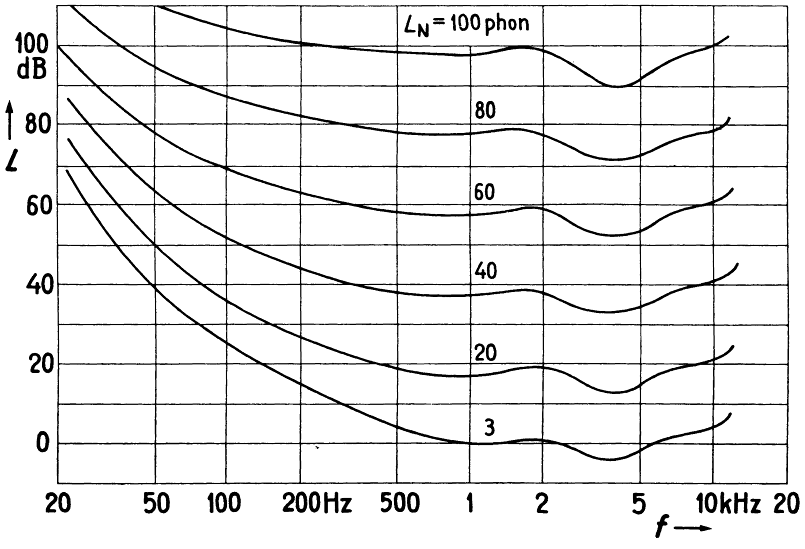


Fig. 2.5: Equal loudness contours (from [ZWI67]).

b) Critical Bands

Apart from its centre frequency, the bandwidth of a signal also has a significant influence on its perceived loudness. A wide band signal is usually perceived as being louder than a narrow band signal at the same sound pressure level and centre frequency. The influence of bandwidth on loudness perception has been investigated by listening tests where band pass noises or multi-tone signals were presented with different bandwidths while the total energy of the signal was kept constant. Such tests [ZWI67] have shown that, up to a certain bandwidth, perceived loudness is almost independent of bandwidth, whereas for larger bandwidths the perceived loudness increases with the bandwidth of the signal. Zwicker [ZWI67] interpreted this by the assumption that the auditory system groups signal components into a limited number of frequency bands, the so-called *critical bands*. The width of these bands depends on their position on the frequency scale. Below 500 Hz, the critical bandwidth is approximately 100 Hz. At higher frequencies, the width of a critical band is in the range of 20% of the centre frequency.

c) Sone Scale

Even though the perceived loudness of a sound naturally increases with increasing sound pressure level, the relation between sound pressure and loudness is not linear. Perceived loudness is usually measured in the unit *sone*, which has been introduced by Stevens [STE36]. The sone scale is derived from the observation that in order to double the perceived loudness of a sound, its sound pressure level has to be increased by approximately 10 $\mathrm{dB}_{\mathrm{SPL}}$, which corresponds to an increase of signal energy by a factor of ten. The relation between the loudness N and the sound pressure is thus given by a power law (Eq. 2.3). One sone is defined as the loudness corresponding to a pure tone at 1 kHz and a sound pressure level of 40 $\mathrm{dB}_{\mathrm{SPL}}$. Consequently, two sone correspond to 50 phon, three sone correspond to 60 phon and so on.

$$N/sone = 2^{\frac{1}{10}(L/phon-40)} = 2^{-4}\cdot 2^{\log_{10}\left(\frac{I_{1kHz}}{I_0}\right)} = \frac{1}{16}\cdot\left(\frac{I_{1kHz}}{I_0}\right)^{0.3} \quad \text{[ZWI67], (2.3)}$$

I_{1kHz} : intensity corresponding to an equally loud sine tone at 1 kHz
I_0 : intensity of a 1 kHz tone at the absolute threshold.

Zwicker [ZWI67] measured the same relation with narrow band noise instead of pure tones and found a slightly different exponent of 0.23. The difference was interpreted as a result of the level dependence of the auditory filters.

d) Difference Limen

Another aspect of loudness perception is the minimum perceptible level difference. It is either referred to as *difference limen* (DLI) or as *just noticeable difference* (JND). The JND slightly depends on level and frequency but is often approximated by assuming a constant JND of one decibel [CRE85]. Nevertheless, for some signal constellations the JND may be significantly lower.

e) Partial Loudness

As the most obvious property of a sound is its loudness, the perceived loudness of an audible distortion is one of the most important quality indicators for a processed audio signal. Therefore, the loudness perception for signal components which are close to the masking threshold produced by other signal components is of particular interest for perceptual models. If a signal is close to the masking threshold, its loudness does not change at once from its current value to zero. Only if the level of a signal component is much larger than the masking threshold it is perceived with the same loudness as when presented alone. When approaching the masked threshold, the loudness is gradually reduced until it becomes zero when reaching the threshold. In the range where the loudness of a signal is reduced by the masker but it is not yet completely masked the perceived loudness is called *partial loudness* and the effect that above the masked threshold the loudness of the maskee is reduced is called *partial masking*. There are only few models for the estimation of partial loudness. One has been introduced in the speech quality measure described in [SCH79] and another one has been proposed recently by Moore et al. [MOO97].

2.1.3 Auditory Frequency Scales

Many hearing phenomena are easier to explain in the frequency domain than in the time domain. This has also a physiological correspondence because the basilar membrane performs a kind of time-frequency decomposition. However, a linear frequency representation neither corresponds to the pitch perception of a human listener nor is it suitable to explain frequency domain effects of hearing, like, for example, simultaneous masking. A logarithmic frequency representation is slightly better fitted to human hearing, but is still not satisfactory. An auditory frequency scale can be derived in numerous ways:

- By measuring the location of maximum deflection of the basilar membrane for pure tones at different frequencies.

- By measuring the width of the critical bands observed in loudness perception (*critical band scale*, unit: *Bark*). A distance of one Bark on the critical band scale corresponds to the width of a critical band. The audible frequency range corresponds to approximately 24 critical bands [ZWI67].
- By measuring the ratio between the subjectively perceived pitch of pure tones. Actually, this would be the only pitch scale in the literal sense, even though also other auditory frequency scales are often referred to as pitch scales . Pitch is normally given in the unit *mel* [ZWI67]. The audible frequency range approximately corresponds to a range from zero to 2400 on the *mel scale*.
- From the just noticeable frequency difference. This results in the so-called *spectral increment scale* (unit: *SPINC*) as proposed by Terhardt [TER92]. The audible frequency range approximately corresponds to a range from zero to 2000 on the SPINC scale.
- From the area covered by the masking curve produced by a narrow band signal. This area divided by the maximum of the masking curve would give the width of an auditory filter in case it would have a rectangular shape. This value is thus called *equivalent rectangular bandwidth* and the corresponding frequency scale is called *ERB scale* (unit: *ERB*) [MOO83]. The audible frequency range approximately corresponds to a range from zero to 38 on the ERB scale.

In [ZWI67], Zwicker assumes that these scales are identical except for a normalisation factor. The relation between critical band scale and mel scale is simply given by the relation that one critical band equals 100 mel.

$$pitch / mel = 100 \cdot critical\,band\ rate / Bark\,. \qquad [ZWI67] \qquad (2.4)$$

Patterson and Moore [PAT86] postulate a similar relation between the ERB scale and the location of maximum deflection of the basilar membrane:

$$location\ on\ BM = 0.9 \frac{mm}{ERB} \cdot ERB - rate\,. \qquad [PAT86] \qquad (2.5)$$

The ERB scale used by Moore [MOO83] differs slightly from the critical band scale defined by Zwicker [ZWI67], especially in the lower frequency range.

The main advantage of using an auditory frequency scale instead of a simple linear or logarithmic frequency scale is that it eases the modelling of frequency domain effects. For example, simultaneous masking can be approximated by two sided exponentials when using the critical band scale and the shape of the masking curves remains almost unchanged for different centre frequencies of the masker.

2.1.4 Other Effects

There are numerous other effects that influence auditory perception but are usually not modelled in perceptual measurement methods.

a) Non-Linearities

Some phenomena of hearing lead to the conclusion that the human auditory system produces non-linear distortions. Examples for such phenomena are the occurrence of beats and combination tones and the perception of missing fundamentals. However, the latter example can also be explained by spectral or temporal pattern recognition.

b) Stapedius Reflex

At high sound pressures, an acoustic reflex occurs that activates two small muscles in the middle ear, the *stapedius* and the *tensor tympani*. The activation of these muscles affects the transmission of the sound pressure to the inner ear in order to protect the ear from damage. As this reflex reduces the energy reaching the inner ear by approximately 20 dB [GRE76], it can have a considerable influence on the perception of loud signals. On the other hand, it is a reflex that is probably aimed to protect the ear. Therefore, one can expect that it occurs only at sound pressure levels high enough to potentially cause a damage to the ear. Of course, in listening tests dealing with high quality audio material such levels should never be used. Modelling the stapedius reflex is thus not considered as particularly important for perceptual measurement. Nevertheless it has occasionally been modelled in perceptual measurement (for example in [KAP93]).

c) Pattern Recognition

Pattern recognition effects are especially occurring in the context of binaural hearing. The most widely known phenomenon based on pattern recognition is the ability of detecting a signal that actually is below the masking threshold if it comes from another direction than the masker. As in such situations the masking threshold can obviously be lowered by binaural analysis, this effect is called *binaural masking level difference* (BMLD).

There are also pattern recognition effects that do not require binaural hearing. Normally the masking threshold produced by a multi-component masker is clearly higher than the sum of the masking thresholds caused by the individual components of the masker. In the case of amplitude modulated maskers the opposite may occur. Adding a new component to the masker may lower the masked threshold produced by the total masker. This effect is called *commodulation masking release* (CMR). Most hypotheses for the origin of this effect can, in general, explain the occurrence of this phenomenon, but not entirely predict the quantity of the effect. The simplest explanation is the assumption that the auditory system is comparing the envelope modulations among different auditory filters. If the same modulation structure is found at different filters, the auditory system “knows” that an amplitude modulated signal is present and may try to detect other signal components by assigning a higher weight on the moments where the temporal minima of the masker occur (*dip listening*).

2.2 Psychoacoustical Models

Psychoacoustical models are simulating certain properties of human hearing. Since loudness is the most obvious property of a sound, models of perceived loudness have always been of particular interest in psychoacoustical research. Many perceptual measurement methods use ideas which originate from such models, especially from Zwickers model for the computation of perceived loudness [ZWI67]. Another model of perceived loudness has been introduced recently by Moore, Glasberg, and Baer [MOO97]. Both models will be described in the following.

2.2.1 Zwicker's Model for the Calculation of Perceived Loudness

Zwicker [ZWI67] describes a model for the computation of perceived loudness that already includes most of the processing steps that are used in perceptual measurement methods. A simplified version of this model which approximates auditory filtering by the use of one-third octave filters has been widely used for loudness estimation in the field of noise prevention and has become a part of an international standard [ISO75].

In the first step the input signal is decomposed into the frequency domain and grouped into critical bands. This can be expressed as a function of energy density on the critical band scale where *I* denotes the energy (intensity) of the signal, and *A(z)* denotes the energy in the critical band *z*.

$$A(z)=\int_{z-0.5Bark}^{z+0.5Bark}\frac{dI}{dz'}\cdot dz' \qquad \left|\begin{array}{ll} \mathrm{z}: & \text{critical band rate / Bark}\\ \dfrac{dI}{dz'}: & \text{critical band energy density}\end{array}\right. \tag{2.6}$$

or as a function of spectral energy density

$$A(z)=\int_{f(z-0.5Bark)}^{f(z+0.5Bark)}\frac{dI}{df}\cdot df \qquad \left|\begin{array}{ll} \mathrm{z}: & \text{critical band rate / Bark}\\ \dfrac{dI}{df}: & \text{spectral energy density}\end{array}\right. . \tag{2.7}$$

The transfer function between outer sound field and inner ear is taken into account by applying a frequency dependent weighting function. The effect of simultaneous masking is interpreted as a result of a spread of neural excitations from the basilar area that corresponds to the frequency range of the sound stimulus into areas that actually should not respond to the sound stimulus. This is modelled by assigning the energy within each critical band an *excitation* in the same critical band and also excitations in the adjacent critical bands. The excitations assigned to adjacent critical bands are determined by the shape of the masking curves. This results in numerous excitations for each critical band (one that originates from the energy in the same critical band and several others that originate from the energies in the adjacent bands). The way how these partial excitations add up is not defined in [ZWI67]. In [ISO75], the excitation is determined by the maximum of the partial excitations, even though this does not correspond to the additivity of masking (see Section 2.1.1 b)). The resulting *excitation patterns* are transformed to a density function *N'(z)* which is defined by the postulate that the area between this function and the frequency axis yields the loudness *N* of the sound stimulus

$$N=\int_{0\,\mathrm{Bark}}^{24\,\mathrm{Bark}} N'(z)\,dz\,. \tag{2.8}$$

The transformation from the excitation pattern *E(z)* to the *specific loudness pattern N'(z)* is given by the warping function

$$N'(z) = 0{,}068 \cdot \left(\frac{1}{s} \cdot \frac{E_{thres}(z)}{E_0} \right)^{0.23} \cdot \left[\left(1 - s + s \cdot \frac{E(z)}{E_{thres}(z)} \right)^{0.23} - 1 \right] \frac{sone}{Bark}, \tag{2.9}$$

where

E : excitation
E_{thres} : excitation corresponding to the threshold in quiet
E_0 : excitation corresponding to a sound pressure level of 40 dB (scale factor)
s : threshold factor.

Finally, the overall loudness of the signal is calculated from Eq. (2.8).

2.2.2 Moore's Model for the Calculation of Partial Loudness

Based on psychoacoustical measurements using the *notched-noise method* [PAT76], Moore and Glasberg [MOO83] have proposed models for auditory filter shapes and bandwidths that differ somewhat from the critical band based model described in the above section. Moore, Glasberg, and Baer [MOO97] have recently published a complete method for the prediction of thresholds, loudness, and partial loudness which is based on these models. It includes generally the same processing steps as the model by Zwicker, but uses different approaches within each of the steps. Since the partial loudness is of much more interest for perceptual quality measurement than the absolute loudness, only this part of the model is described here.

Like in Zwickers model, in the first step the input signals are transformed to excitation patterns on an auditory frequency scale. The auditory frequency scale is determined by the equivalent rectangular bandwidth (*ERB*). ERB and ERB-rate are approximated by the formulas

$$ERB / Hz = 24.7 \cdot (1 + 4.37 \cdot f / Hz) \tag{2.10}$$

and

$$ERB\text{-}rate / ERB = 21.4 \cdot \lg(4.37 \cdot f / kHz + 1). \tag{2.11}$$

Auditory filter shapes are modelled by the so-called *ROEX*-filter shapes ("*ROunded EXponentials*"). The shape of a ROEX-filter is defined by

$$W(g) = (1 + p \cdot g) \cdot e^{-p \cdot g}, \qquad \text{[MOO93]} \tag{2.12}$$

where

$$g = \frac{|f - f_{centre}|}{f_{centre}} \tag{2.13}$$

and

$$p = \frac{4 \cdot f_{centre}}{ERB\left(f_{centre}\right)} . \qquad (2.14)$$

From the excitation patterns, loudness and partial loudness are calculated using different formulas for four different cases. These different cases are given by distinguishing between total signal levels (masker plus maskee) below and above 100 dB SPL, and between signals below and above masked threshold.

The partial loudness for local excitation levels between the masking threshold and 100 dB SPL is calculated according to

$$\begin{aligned} N'_{sig} &= C \cdot \left\{ \left[\left(E_{sig} + E_{mask} \right) \cdot G + A \right]^{\alpha} - A^{\alpha} \right\} \\ &\quad - C \cdot \left\{ \left[\left(E_{mask} \cdot (1+K) + E_{thresQ} \right) \cdot G + A \right]^{\alpha} . \right. \\ &\quad \left. - \left(E_{thresQ} \cdot G + A \right)^{\alpha} \right\} \cdot \left(\frac{E_{thresN}}{E_{sig}} \right)^{0.3} \end{aligned} \qquad (2.15)$$

The partial loudness for local excitation levels below the masking threshold is calculated according to

$$\begin{aligned} N'_{sig} &= C \cdot \left(\frac{2 \cdot E_{sig}}{E_{sig} + E_{thresN}} \right)^{1.5} \\ &\quad \cdot \left\{ \frac{\left(E_{thresQ} \cdot G + A \right)^{\alpha} - A^{\alpha}}{\left[\left(E_{noise} \cdot (1+K) + E_{thresQ} \right) \cdot G + A \right]^{\alpha} - \left(E_{noise} \cdot G + A \right)^{\alpha}} \right\} . \\ &\quad \cdot \left\{ \left[\left(E_{sig} + E_{noise} \right) \cdot G + A \right]^{\alpha} - \left(E_{noise} \cdot G + A \right)^{\alpha} \right\} \end{aligned} \qquad (2.16)$$

where

E_{sig} : excitation produced by the signal
E_{mask} : excitation produced by the masker
E_{thresQ} : excitation corresponding to the threshold in quiet (internal noise)
E_{thresN} : excitation corresponding to the masking threshold
G : low level gain of the cochlear amplifier
A : frequency dependent constant that determines the level dependence of loudness compression
C : scaling constant, C = 0.047
K : threshold factor
α : compression exponent.

and

$$E_{thresN}(f) = K(f) \cdot E_{mask}(f) + E_{thresQ}(f), \qquad (2.17)$$

$$G(f) = \frac{E_{thresQ}(500\,Hz)}{E_{thresQ}(f)}. \tag{2.18}$$

In [MOO97], the frequency dependent constants are only given as diagrams. They can, however, roughly be approximated by the following equations:

$$E_{thresQ}(f) = 1.4 + 0.4 \cdot 10^{0.3 \cdot (f/kHz)^{-0.8}}, \tag{2.19}$$

$$K(f) = 0.35 + 0.14 \cdot 10^{0.2 \cdot (f/kHz)^{-0.8}}, \tag{2.20}$$

$$\alpha(f) = 0.171 + \frac{0.032085}{0.1 + G(f)^{0.25}}, \tag{2.21}$$

$$A(f) = 2.8 + \frac{2}{0.1 + G(f)^{0.25}}. \tag{2.22}$$

In contrast to Zwickers approach, the specific loudness and specific partial loudness are not zero at the threshold of audibility, but have a finite value. Moreover, the low frequency roll-off of the threshold in quiet is only partly assigned to internal noise and reduced detector sensitivity, and parts of it are assigned to the middle ear transfer function, whereas it was entirely assigned to internal noise in all previously existing models.

2.2.3 Analytical Expressions for Psychoacoustical Phenomena

This subsection gives an overview on analytical expressions that are frequently used when modelling parts of the auditory system. The use of analytical expressions introduces slight inaccuracies into the ear models because they are only approximations of a measured effect. On the other hand, the confidence intervals of psychoacoustical experiments are in most cases so large that the inaccuracies introduced by such an approximation can be neglected when compared to the inherent inaccuracies of the psychoacoustical experiment. Using analytical expressions eases the implementation of a model. Moreover, they normally provide a rather good interpolation and extrapolation of the values they describe, and they allow to keep a model flexible because they are not linked to a fixed number of tabulated values.

a) Auditory Frequency Scales

- **Critical Band Rate**

The critical band rate is expressed in the unit *Bark* and is usually referred to by the symbol z. Schroeder [SCHR79a] proposed a very simple approximation to calculate the frequency that corresponds to a given critical band rate:

$$f/kHz \approx \sinh\left(\frac{z/Bark}{7}\right). \tag{2.23}$$

This approximation was designed for a frequency range relevant for speech coding. It only corresponds to the critical band rate as measured by Zwicker [ZWI67] for frequencies below 5 kHz. The formula can be easily inverted in order to get the reverse transformation from frequency to critical band rate:

$$z / Bark \approx 7 \cdot \operatorname{arsinh}\left(f / kHz \right). \tag{2.24}$$

Terhard and Zwicker [ZWI80] proposed another expression that approximates the critical band rate within the entire audible frequency range

$$z / Bark = 13 \cdot \arctan\left(0.76 \cdot f / kHz\right) + 3.5 \arctan\left[\left(\frac{f / kHz}{7.5}\right)^2\right]. \tag{2.25}$$

- **Critical Bandwidth**

In [ZWI80], Terhard and Zwicker also proposed an approximation for the critical bandwidth:

$$critical\ bandwidth / Hz = 25 + 75 \cdot \left[1 + 1.4 \cdot \left(f_{centre} / kHz\right)^2\right]^{0.69}. \tag{2.26}$$

- **Equivalent Rectangular Bandwidth**

The most frequently used auditory frequency scale except for the critical band scale is based on the equivalent rectangular bandwidth. According to [MOO83] it can be approximated by a second order polynomial

$$ERB / Hz = 6.23 \cdot \left(f / kHz\right)^2 + 93.39 \cdot f / kHz + 28.52 \tag{2.27}$$

and the corresponding frequency scale is defined by

$$ERB\text{-}rate / ERB = 11.17 \cdot \ln\left(\frac{f / kHz + 0.312}{f / kHz + 14.675}\right) + 43.0. \tag{2.28}$$

Later on, Glasberg and Moore [GLA90], [MOO97] proposed a slightly different approximation:

$$ERB / Hz = 24.7 \cdot \left(1 + 4.37 \cdot f / kHz\right) \tag{2.29}$$

and

$$ERB\text{-}rate / ERB = 21.4 \cdot \lg(4.37 \cdot f / kHz + 1). \tag{2.30}$$

According to [STU92] the ERB-scale can also be approximated by

$$ERB / Hz = 19.5 + 0.117897 \cdot f / Hz \tag{2.31}$$

and

$$ERB\text{-}rate\,/\,ERB = 18.31\cdot \lg(0.006046\cdot f\,/\,Hz+1)\,. \qquad (2.32)$$

The three different approximations are depicted in Figure 2.6

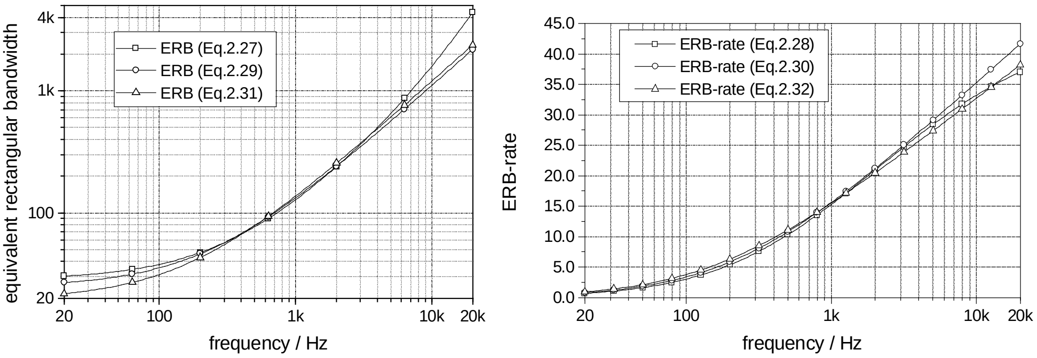


Fig. 2.6: Comparison of different approximations of the ERB-scale (left: bandwidth, right: ERB-rate).

- **Spectral Increment Function (SPINC)**

Terhardt [TER92] proposed another auditory frequency scale derived from the frequency discrimination threshold. However, this function is rarely used in perceptual measurement. It can be approximated by the equation

$$\phi(f)\,/\,spinc = 1414\cdot \arctan\left(\frac{f}{1414\,Hz}\right) \qquad [TER92]\,. \qquad (2.33)$$

b) Masking Curves and Excitation Patterns

In general, approximations for masking curves and excitation patterns are identical. The difference between both is only given by the way these functions are applied in the perceptual model.

- **Level Independent with a Smooth Peak**

Schroeder [SCHR79a] used an approximation which provides fixed upper and lower slopes with a slope rate of 25 dB/Bark at the lower slope and 10 dB/Bark at the upper slope (which corresponds to measured masking curves at intermediate sound pressure levels):

$$B(z\,/\,Bark) = 15.81 + 7.5\cdot(z+0.474) - 17.5\cdot\sqrt{1+\left(z+0.474\right)^2}\,dB\,. \qquad (2.34)$$

This formula models a soft transition between lower and upper slope.

- **Level-Dependent with a Sharp Peak**

Terhardt [TER79] proposed an approximation that takes the level dependence of simultaneous masking into account but does not provide a continuous transition between lower and upper slope:

$$B(z/Bark) = 10^{\frac{1}{10}\cdot S_1\cdot(z-z_{centre})} \quad | \quad z \le z_{\mathrm{centre}} \tag{2.35}$$

$$B(z/Bark) = 10^{-\frac{1}{10}\cdot S_2\cdot(z-z_{centre})} \quad | \quad z > z_{\mathrm{centre}} \tag{2.36}$$

where

$$S_1 = 27\frac{dB}{Bark}\ , \quad S_2 = \left(24 + \frac{230}{f_{centre}/Hz} - 0.2\cdot L/dB\right)\frac{dB}{Bark} \tag{2.37}$$

and

$z_{centre} = z(f_{centre})$: location of maximum excitation on the critical band scale.

- **Level-Dependent with a Smoothed Peak**

A generalisation of Schroeders formula for a masking curve with a soft transition between lower and upper slope is given by

$$B(z/Bark) = A_0 + \left(\frac{S_1 - S_2}{2}\right)\cdot(z+c_1) - \left(\frac{S_1 + S_2}{2}\right)\cdot\sqrt{c_2 + (z+c_1)^2}\ , \tag{2.38}$$

where

S_1 : lower slope rate
S_2 : upper slope rate
A_0 : peak excitation
c_1 : frequency shift
c_2 : width of the peak excitation.

This formula allows the combination of Schroeders modelling of a smooth transition between lower and upper slope with the level dependencies of simultaneous masking proposed by Terhardt [TER79]. It has been used in [DEU92] and in the perceptual measurement method described in [COL93].

- **Worst Case Approximation**

If it is not possible to model the level dependence of masking curves (e. g. when the playback level of the audio signal is not known in advance) the deviation between the correct (level-dependent) masking curve and the modelled (fixed) masking curve can be reduced by using a modified masking curve as proposed in [BRA89]. This masking curve is equal to or below the level-dependent masking curve in the entire range above the absolute threshold, whereas it is allowed to take any possible value in the range where the signal would be below the absolute threshold. Originally, this function is taken from a table and also takes the shape of the threshold in quiet into account. With the simplification that the absolute threshold is replaced by a constant value of zero decibels, an analytical expression for the upper slope of this *worst case* *masking curve* can be derived from the level dependence given by Terhardt [TER79] (Eq. 2.37):

$$B_{worst\ case}(\Delta z) = 10^{-\frac{S_2'}{10}\cdot\Delta z\cdot\left(\frac{1}{1+\frac{\Delta z}{5\,Bark}}\right)} . \tag{2.39}$$

where

$$S_2' = 24\frac{dB}{Bark} .$$

Since the lower slope is not level-dependent, the worst case masking curve is identical to the approximation given by Terhardt (see above).

c) Forward Masking

An approximation for the forward masking curves described in [ZWI67] was proposed by Kapust [KAP93].

$$D(t,T_m) = 1.0 - \frac{1}{1.35}\arctan\left[\frac{t/ms}{13.2\cdot(T_m/ms)^{0.25}}\right] , \tag{2.40}$$

where

T_m : masker duration,
t : time after the end of the masker .

d) Masking Index and Threshold Factor

All proposed approximations yield a masking index, S, in decibels. The threshold factor, s, is the linear representation of the masking index on an energy scale:

$$s = 10^{\frac{1}{10}\cdot S\ /\ dB} . \tag{2.41}$$

- **Masking Index for Pure Tones**

Schroeder [SCHR79a] approximated the dependence of the masking index on masker frequency by a linear relation:

$$S/dB = -(15.5 + z/Bark) . \tag{2.42}$$

- **Masking Index for Narrow Band Noise**

Kapust [KAP93] proposed an approximation to model the (slight) frequency dependence of the masking index for narrow band noise maskers:

$$S\,/\,dB = -2.0 - 2.05 \cdot \arctan\left(\frac{f}{4\,kHz}\right) - 0{,}75 \cdot \arctan\left(\frac{f^2}{2.56\,\mathrm{kHz}^2}\right). \qquad (2.43)$$

- **Masking Index as a Function of Tonality**

For maskers that are neither purely tonal nor clearly noise-like, Kapust [KAP93] proposed a weighted linear interpolation between the threshold indices calculated on the assumption of tonal and of noise-like signals. The weighting is determined by a tonality measure which is derived from the so-called measure of chaos, a measure of the distance between actual and linearly predicted phase and amplitude.

$$S\,/\,dB = \alpha \cdot S_{tone} + (1-\alpha) \cdot S_{noise} \quad , \qquad (2.44)$$

where

α : tonality measure
S_{noise} : masking index for noise-like maskers
S_{tone} : masking index for pure tone maskers .

e) Specific Loudness

The approximation most frequently used for calculating specific loudness is given in [ZWI67] :

$$N' = k \cdot \left(\frac{1}{s} \cdot \frac{E_{thres}}{E_0}\right)^{\gamma} \cdot \left[\left(1 - s + s \cdot \frac{E}{E_{thres}}\right)^{\gamma} - 1\right], \qquad (2.45)$$

where

k : scaling factor; in [ZWI67]: $k = 0.068$
γ : in [ZWI67]: $\gamma = 0.23$
s : threshold factor
E_{thres} : excitation corresponding to the absolute threshold
E_0 : normalisation factor.

f) Partial Masking

In [SCHR79a] Schroeder et al. proposed a very simple approximation for the effect of partial masking. In an informal listening test they matched the intensity of a partially masked signal to its intensity in the absence of a masker. The result could be approximated by

$$I'_{Noise} = \frac{I_{Noise}}{1 + \left(\frac{I_{Signal}}{I_{Noise}}\right)^2}. \qquad (2.46)$$

Moore et al. [MOO98] recently presented a much more complex model for the calculation of partial loudness. They used different formulas depending on whether the signal is above or below the threshold and depending on the absolute sound pressure level (see Section 2.2.2).

g) Threshold in Quiet

In [TER79] the absolute threshold of hearing is approximated by the formula

$$\begin{aligned} threshold/dB = {} & 3.64 \cdot \left(f/kHz\right)^{-0.8} - 6.5 \cdot e^{-0.6 \cdot \left(f/kHz - 3.3\right)^2} \\ & + 10^{-3} \cdot \left(f/kHz\right)^4 \end{aligned} . \tag{2.47}$$

This approximation is used in almost all perceptual measurement methods. It consists of three terms, one describing the lower frequency cut-off, one describing the increased sensitivity of the ear in the frequency range around 3 kHz and one describing the upper frequency cut-off. The first term (or at least part of it) is usually interpreted as a result of *internal noise* (caused by muscle activity, blood flow etc.), whereas the last two terms are interpreted as the transfer characteristic of middle and inner ear. Consequently, in perceptual models this equation is often divided into two parts: one called the *internal noise function* and one called the *middle ear transfer function*.

- **Middle Ear Transfer Function**

The middle ear transfer function is usually modelled by the equation

$$A/dB = -6.5 \cdot e^{-0.6 \cdot \left(f/kHz - 3.3\right)^2} + k \cdot 10^{-3} \cdot \left(f/kHz\right)^4 \quad , \tag{2.48}$$

where *k* determines the upper frequency cut-off (see above). According to the approximation of the absolute threshold given in Eq. (2.47), the value of k is one. As the upper frequency cut-off strongly depends on the age of the listener, *k* can assume different values when a different population of listeners is to be modelled. The upper frequency cut-off in the absolute threshold modelled by Eq. (2.47) starts rather early, i. e. the test listeners have obviously been rather old and can probably not be considered as "normal hearing listeners". For this reason, in [KAP93] *k* was set to a value of two in order to shift the upper frequency cut-off to higher frequencies.

- **Internal Noise**

The internal noise function is usually modelled by the left hand part of Eq. (2.47) (for example in [COL93]):

$$E_{internal\ noise}/dB = 3.65 \cdot \left(f/kHz\right)^{-0.8} . \tag{2.49}$$

2.2.4 Summary and Conclusions

Apart from the different formulas used for loudness calculation, the main difference between the psychoacoustical models proposed by Moore and Zwicker are the different auditory frequency scales and the different auditory filter shapes. Critical band scale and ERB-scale differ in both the distribution of the auditory bands and the width of the auditory bands. The ERB-scale used by Moore is derived from experiments that were especially designed in order to measure auditory filter bandwidths, whereas the critical band scale is derived from an effect which is influenced by auditory filter bandwidths, but may also depend on other stages of auditory processing. Thus, the ERB-scale should be a better representation of the frequency selectivity in the auditory system than the critical band scale. On the other hand, the experiments used to establish the ERB-scale were based on assumptions on the auditory filter shape which are not necessarily true. ERB-scale and ROEX-filters were derived under the assumption that the auditory filters are equally shaped and symmetrical when related to a logarithmic frequency scale. The filter shapes used by Zwicker and Terhardt were derived under the assumption that the auditory filters are equally shaped when related to an auditory frequency scale which is measured by experiments that do not require assumptions on the auditory filter shapes. The latter approach appears to be less influenced by experimental conditions than the first one.

Even if the ERB-scale is a better representation of the auditory filter bandwidths this does not necessarily mean that it is also a better starting point for a perceptual model. As long as the effects that are responsible for the deviations between auditory filter bandwidth and critical bandwidth are not incorporated in a perceptual model, they would be missing in a model based on the ERB-scale, whereas they would be taken into account when using the critical band scale (although they would be modelled in the wrong place).

With respect to the widths of the auditory filters, the critical bandwidth seems to be a less convenient criterion because it is probably based on spectral integration and the auditory filters may be clearly narrower than the critical bands. Here, the equivalent rectangular bandwidth might be a more appropriate criterion. The conclusion for this work is to use the critical band scale in order to determine the distribution of the auditory filter bands, but rather use a fraction of the critical bandwidth than the critical bandwidth itself to determine the width of the auditory filter bands.

2.3 Concepts of Perceptual Models

Human sound perception can roughly be described by a five stage approach (Figure 2.7). The outer sound field is transmitted to the inner ear (1) and decomposed into its spectral components (2). The sensitivity of the ear and its frequency selectivity are enhanced by active processes (3) which probably include a kind of feedback mechanism. The neural excitations at the inner ear are transmitted to the auditory centres of the brain through the auditory nerve and are translated into sensational quantities (4). The auditory centres also perform several kinds of pattern recognition mechanisms (5) which may again influence the formation of the sensational quantities.

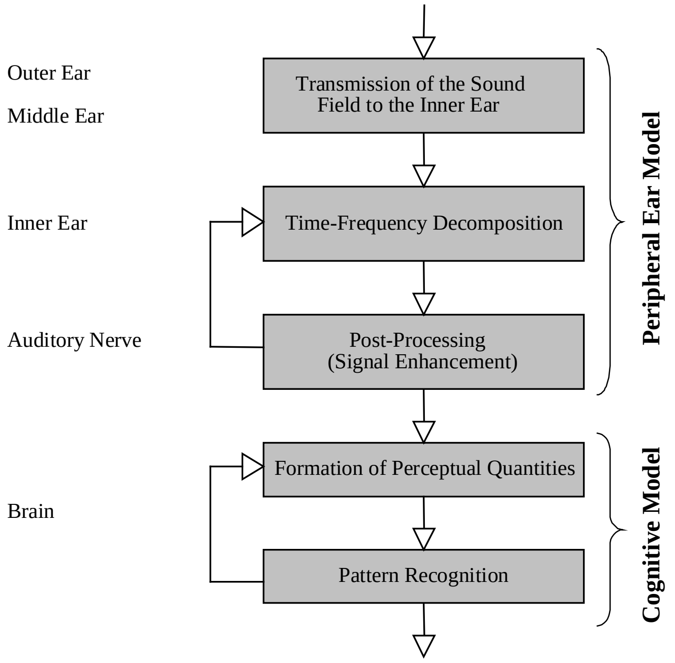

Outer Ear
Middle Ear
Transmission of the Sound Field to the Inner Ear
Inner Ear
Time-Frequency Decomposition
Auditory Nerve
Post-Processing (Signal Enhancement)
Peripheral Ear Model
Brain
Formation of Perceptual Quantities
Pattern Recognition
Cognitive Model

Fig. 2.7: Stages of Auditory Processing.

The first three stages in Figure 2.7 describe the translation from an outer sound field to neural excitations and the last two stages describe the further processing from these excitation patterns to sensations. The translation from outer sound field to neural excitations is almost independent of individual preferences and represents the part of sound perception that is mainly based on the physiological structure of the auditory system. In a perceptual model, these steps are called the *peripheral ear model*. In the later stages of auditory processing, individual preferences cannot be clearly separated from more common properties of auditory processing. These stages, which include pattern recognition processes and auditory streaming, are referred to as the *cognitive model*.

Auditory perception can be modelled by different approaches. These approaches are usually compromises between two extreme concepts:

- *Functional models* of the physiological processes that occur in the auditory system. Such a model would be entirely based on measurements of acoustical and mechanical properties of the ear as well as measurements of neural excitations on the auditory nerve and neural activity in regions of the brain that are involved in auditory processing.
- *Heuristic models* of observed properties of the auditory system. Such a model would be entirely based on listening test results. The structure of such a model does not necessarily correspond to the structure of auditory processing in the human ear.

In theory, the first approach could yield a perfect model for all possible auditory phenomena. On the other hand, such a model usually requires an extremely high computational effort. Furthermore, one cannot expect to know all underlying physiological processes incorporated in auditory perception in sufficient detail.

The latter approach is more practical and requires - as long as the number of auditory phenomena to be modelled is limited - a lower computational effort. Besides the lower computational effort, the main advantage of this approach is that it can have highest possible accuracy for particular aspects of hearing since it is directly derived from the phenomena it should reproduce. On the other hand, the more auditory phenomena are to be modelled with such an approach, the more complex it gets. Furthermore, different auditory phenomena might lead to contrary models. Therefore, most perceptual models are partly based on physiological processes and partly on phenomena that can be observed by listening tests. Among the two main concepts used in perceptual measurement methods (see Sections 2.3.1 and 2.3.2), the concept of *comparing internal representations* is closer to a functional model and the *masked threshold concept* is closer to a heuristic model.

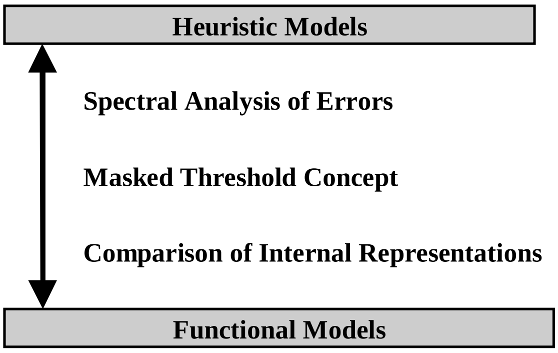

Heuristic Models
Spectral Analysis of Errors
Masked Threshold Concept
Comparison of Internal Representations
Functional Models

Fig. 2.8: Ordering of different concepts of perceptual models between heuristic and functional models.

2.3.1 Masked Threshold Concept

The *masked threshold concept* (also: "*noise signal evaluation*") has been used in the earlier perceptual measurement methods like the method described in [SCH79a] and the NMR [BRA87]. In such a model, the error signal, which is the difference between the original and the processed signal, is compared to the masking threshold produced by the original signal.

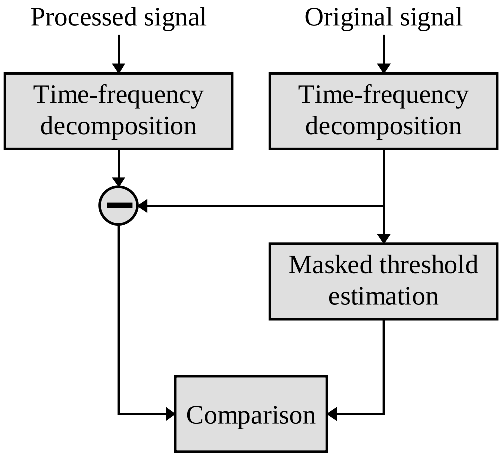

Processed signal
Original signal
Time-frequency decomposition
Time-frequency decomposition
Masked threshold estimation
Comparison

Fig. 2.9: Masked threshold concept.

Its main advantage is the possibility to derive the parameters of the model directly from masking experiments. Furthermore, such a model can be used in audio encoding without major changes. On the other hand, the possibility to model more complex auditory phenomena with this concept is rather limited.

2.3.2 Comparison of Internal Representations

The concept of *comparing internal representations* (also: "*comparison in the cochlear domain*") has been introduced in [KAR85] and is used in most of today's perceptual measurement methods (for example in [BEE92, PAI92, COL93]). It is based on the calculation of excitation patterns for both the original and the processed signal. The properties of the distortions (audibility, loudness, annoyance etc.) are estimated by comparing these excitation patterns. This concept is much closer to a functional model of the auditory system than the masked threshold concept and therefore is a better starting point for the modelling of more complex auditory phenomena.

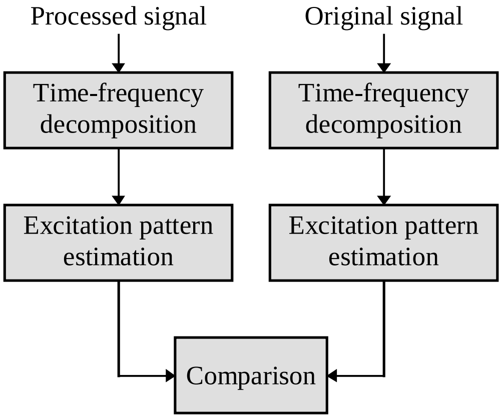

Processed signal
Original signal
Time-frequency decomposition
Time-frequency decomposition
Excitation pattern estimation
Excitation pattern estimation
Comparison

Fig. 2.10: Comparison of internal representations.

2.3.3 Analysis of Linear Error Spectra

Some effects, like, for example, the perception of fundamental frequency, are easier to model when using linear spectra instead of a basilar membrane model. This is considered as a third concept and referred to as *spectral analysis of errors*. Apparently, this concept cannot be used in filterbank-based models, but only in FFT-based models. As it is clearly far away from a functional model of the auditory system, a perceptual model cannot be solely based on this concept. Nevertheless, it can be a useful supplement to the above concepts because it yields some additional information about the character of the distortions which is difficult to obtain from other concepts.

2.4 Perceptual Measurement Methods

The common feature of all perceptual measurement methods is the modelling of masking effects. Simultaneous masking is always modelled by applying a spreading function that corresponds in shape to an average masking curve (Figure 2.12). Temporal masking effects are often not explicitly modelled, but emerge very roughly from the limited temporal resolution of the time-frequency decomposition. The degradations of the test signal are estimated by comparing the processed signal to the original signal, which serves as a reference. The test signals to be used are usually the same musical excerpts as used in the subjective evaluation of audio codecs. However, in principle, any kind of audio signals, including artificially created test signals, may be used.

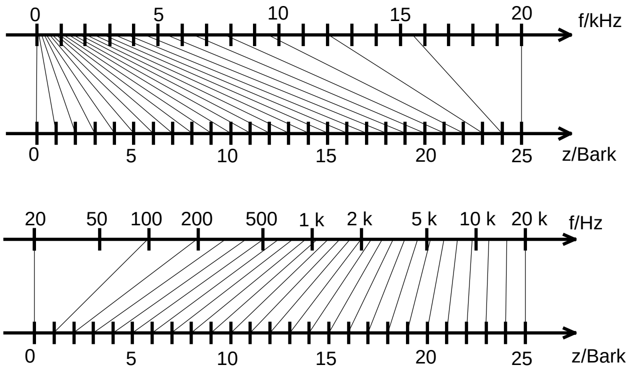


Fig. 2.11: Frequency warping when modelling an auditory frequency scale (upper: linear frequency scale, lower: logarithmic frequency scale)

The starting point of most perceptual measurement methods is a filter or weighting function to model the transfer function of outer and middle ear, followed by a short-term time-frequency decomposition. The frequency scale is warped in order to take the non-uniform spectral resolution of the auditory system into account (Figure 2.11). The main effect of this frequency warping, besides matching the auditory frequency resolution, is that the auditory filter shapes become almost uniform, and therefore can be modelled by a simple convolution with a spreading function (Figure 2.12). When the masked threshold concept is used, this convolution is applied to the original signal only. When the comparison of internal representations is used, this convolution is applied to both the original and the processed signal.

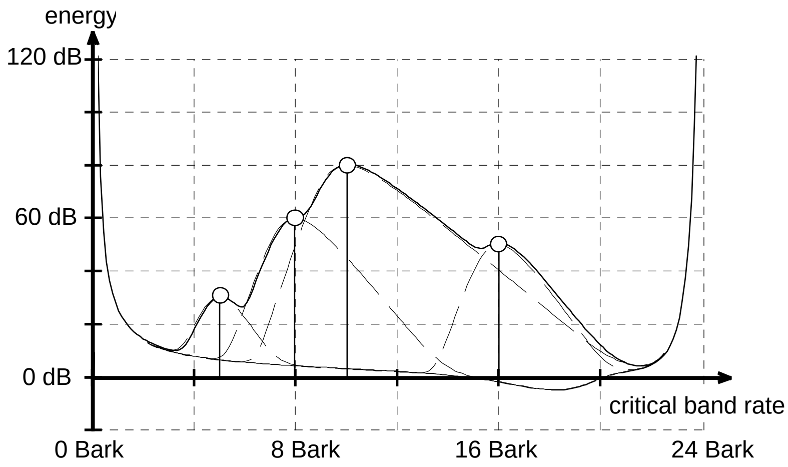


Fig. 2.12: Frequency domain smearing by convolving a linear spectrum with a spreading function (example: 4 pure tones + threshold in quiet, linear additivity).

If temporal masking is explicitly modelled, this will usually be done by a simple first order low pass filter, which performs an exponential temporal averaging. The perceptually adapted signal representations thus obtained are compared to each other in order to estimate the severeness of the degradations of the processed signal. The details of this comparison form the main differences among the existing perceptual measurement methods.

The predecessor of all perceptual measurement methods is Zwickers procedure of calculating perceived loudness (see Section 2.2.1). Hardware implementations of this algorithm were already made in the early 60s, like, for example, in [FIS64] where even the level dependence of auditory filter shapes had been modelled. The first applications of such algorithms for the estimation of signal degradations have been introduced by Schroeder et al. [SCH79] and by Karjalainen [KAR85]. These methods have been applied to speech codecs, but not to audio codecs.

- The measurement method *Speech Signal Degradation* [SCH79] calculates the loudness of the errors found in the processed signal reduced by the original signal that partly masks these errors. The time-frequency decomposition is performed via an FFT. Masking thresholds are approximated by a spreading function which is derived from a measured masking curve for a 1 kHz sine tone at a sound pressure level of 80 dB. It is the first known measurement method that uses the masked threshold concept.
- The measurement method *ASD* ("*Auditory Spectrum Distance*") [KAR85] calculates the difference loudness between the processed signal and the original signal. The time-frequency decomposition is performed by an FIR filter bank with a spectral resolution of half a critical bandwidth. Masking thresholds are modelled directly by the frequency response of the filters, which are designed to approximate the same masking curve as used in [SCH79]. Temporal masking is modelled by an asymmetric non-linear filter that corresponds better to temporal masking effects than the usual first order IIR-filters. ASD is the first known perceptual measurement method that uses the concept of comparing internal representations

Starting in 1987 with the introduction of the NMR measurement system [BRA87], numerous perceptual measurement methods for the quality evaluation of audio codecs have been developed. The most important methods are briefly described in the following subsections. A more detailed overview on perceptual measurement methods has been presented in [THI94a]-[THI94c].

2.4.1 NMR

The *NMR* ("*Noise-to-mask ratio*") measurement system [BRA87] explicitly computes an error signal, either as the difference between processed and original signal, or as the absolute difference between their amplitude spectra. Error signal and original signal are analysed within 27 non-overlapping frequency bands, using the masked threshold concept. The general structure of this method was based on a simplified version of the method described in [SCH79]. The masking curve has been modified in order to get as close as possible to the level-dependent masking curves known from psychoacoustical experiments, without explicitly modelling this level dependences ("*worst case masking curve*" see Figure 2.13 and Eq. 2.39). The low number of frequency bands together with a rather simple model of the auditory system eases a real-time implementation of this method, and it became the first commercially available stand-alone perceptual measurement system. The simplifications of the auditory model and particularly the low number of frequency bands limit the performance of this method when estimating perceived audio quality. However, the NMR system has successfully been used as a tool in codec development for several years.

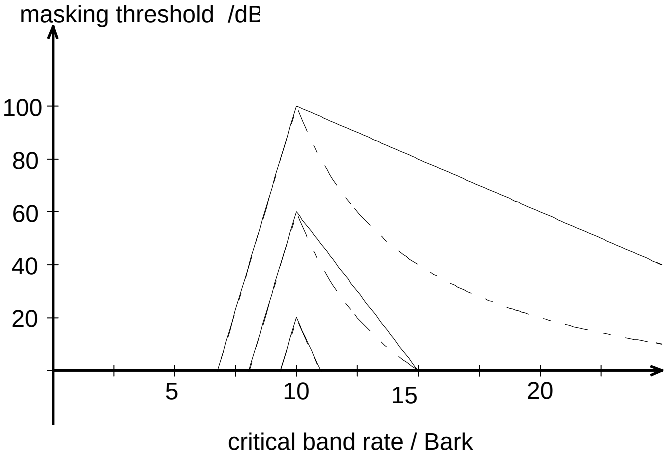


Fig. 2.13: Relation between the worst case masking curve(dashed lines) used in NMR and level-dependent masking curves (solid lines).

2.4.2 PAQM

The *PAQM* ("*Perceptual Audio Quality Measure*") [BEE92] incorporates a much more detailed auditory model than NMR. It models the level dependence of masking as well as the non-linear additivity among different masker components and the asymmetry between tonal and noise-like maskers. These effects are modelled by applying a power function (with an exponent smaller than one) to the local energy densities before spectral and temporal masking functions are applied (as proposed in [HUM89]). The main output of PAQM is the logarithm of the *noise disturbance*, which is the difference loudness between processed and original signal, based on the specific loudness calculation proposed by Zwicker [ZWI67].

The exponents used in the modelling of masking asymmetry and the exponent used in the calculation of specific loudness have been adjusted experimentally to achieve the highest possible correlation between model predictions and subjective gradings of coded audio signals. Consequently, PAQM showed higher correlations with subjective quality gradings than most other perceptual measurement methods, but the model is not entirely related to psychoacoustics anymore.

The performance of PAQM as a predictor of subjective quality gradings has been further enhanced by modifying the spectral and temporal averaging strategy. These additions, which are referred as cognitive correction [BEE94] or perceptual streaming [BEE96], are again mainly determined by experimental optimisation.

For the quality measurement of speech codecs, a modification of PAQM, the *PSQM* ("*Perceptual Speech Quality Measure*") [BEE93] has become an international standard [ITU97]. This modification deviates even more from classical psychoacoustics, as it does not even model simultaneous masking anymore.

2.4.3 PERCEVAL

PERCEVAL ("*PERCeptual EVALuation*") [PAI92] is mainly characterised by the use of an extremely high number of frequency bands and the computation of a detection probability as the main output parameter. Apart from the high number of more than 2000 filter bands, which is determined by the estimated number of hair-cells along the basilar membrane, the underlying ear model of PERCEVAL is rather simple. It models neither the level dependencies of auditory filter shapes, nor is temporal masking explicitly modelled. To overcome the restriction to distortions near threshold, which is implied by the approach of calculating a detection probability, later versions of PERCEVAL also used the average distance between the excitation patterns of processed and original signal as output parameter. When the model was tested within the ITU-R TG 10/4 competitive tests, several additional parameters have been included that were separately calculated for different frequency regions and mapped to a global quality measure using a neural network. However, this extended version has never been published except for those parts of the model that are incorporated in the ITU-R recommended measurement method PEAQ [THI98].

2.4.4 POM

The measurement method *POM* ("*Perceptual Objective Model*") [COL93] combines the statistical approach of calculating a detection probability with the detailed ear model used in PAQM, including non-linear additivity of masking and level-dependent spreading functions. It also models the more smoothed masking curves as used in [SCH79]. The number of filter bands is determined by the frequency discrimination threshold. The resulting number of bands is lower than in PERCEVAL, but still rather high. Like in PERCEVAL, the detection probability has later been complemented by adding an excitation difference measure, and like PERCEVAL the model has been extended by a large set of additional parameters which are mapped to a single quality measure by a neural network.

2.4.5 OASE

OASE (*Objective Audio Signal Evaluation*) is a filterbank-based perceptual measurement method that has been introduced in the end of 1996 [SPO96]. It is based on an FIR-filter bank with 241 frequency bands, which corresponds to a rather high frequency resolution of one-tenth critical bandwidth. The high computational effort of this filter bank is reduced by using a fast-forward-convolution algorithm and by pre-filtering the input signals using a tree-structured low pass cascade which allows to reduce the sampling rates in the lower filter bands. The filter shapes approximate the worst case masking curves used in NMR. Temporal masking was originally modelled by a non-linear low pass filter as proposed in [KAR85], but has later been replaced by a simpler IIR-filter. The main output of this model is the probability of detection of the excitation differences between processed and original signal.

2.4.6 Comparison of Different Concepts

a) Level Dependence of the Excitation Slopes

Auditory filter shapes (and hence also the masking curves) differ significantly among different masker levels, which can be taken into account either by applying a level-dependent spreading function or by using the worst case spreading function proposed in [BRA87]. In practice, the effect of level dependencies is reduced by the fact that the sound pressure levels used in listening tests are usually within a rather limited range. On the one hand side, the level is set as high as possible in order to avoid potentially audible distortions from being hidden due to the absolute threshold. On the other hand, very high sound pressure levels are uncomfortable for the listeners, and are therefore also to be avoided. However, especially in test items with a large dynamic range, the effect of level dependencies may still be rather large and should not be neglected in a perceptual model. As long as the playback level of the test items is known, the modelling via level-dependent spreading functions is preferable, because it is clearly much more accurate than other approaches. In case it is not possible or not desired to define the playback level of the test signal, the worst case masking curves from [BRA87] appear to be a reasonable alternative. The latter case may, for example, occur in codec evaluation where the overall performance of a codec for all possible playback levels may be more important than the performance for a pre-defined playback level.

b) Additivity of Masking

The way the non-linear additivity of masking is taken into account in existing perceptual models like PAQM or POM is not completely satisfactory. Even though the model from [HUM89], which is incorporated in both measurement methods, appears to be an improvement when compared to methods that neglect this effect, it is not fully consistent. First of all, such a model generally yields a much higher excitation for noise-like signals as compared to tonal signals, which also implies a remarkably increased loudness for such signals. This is clearly not realistic. Moreover, the exponent used in the power law given in [HUM89] for the additivity of masking depends strongly on the experimental context, which actually yields the conclusion that such a power law is not an appropriate model for this effect.

c) Number of Filter Bands

The optimum number of filter bands to be used is a parameter which is difficult to estimate on the basis of purely theoretical considerations. It can, however, be stated that extreme high filter numbers like, for example, used in PERCEVAL should not be required as most of the filters apparently do not carry any additional information. On the other extreme, filter numbers lower than the number of critical bands clearly will result in a loss of information. As the critical band concept in this respect relies on non-overlapping filters, which also means a loss of information, the minimum filter number should be somewhat higher than this. The auditory filter bandwidth may therefore be a more reasonable criterion for the minimum number of required filter bands.

For FFT-based models, a more practical criterion to determine the maximum number of filter bands emerges from the requirement that all frequency bands should contain relevant information. This is, the filter bands should at no place on the frequency scale be narrower than the frequency range covered by one line of the underlying time-frequency decomposition. For an FFT-length of 2048 tabs at a sampling rate of 48 kHz, which is the most typical setting for the existing perceptual models, this results in a number of slightly more than 100 filter bands. In case no rectangular bands, but more rounded shapes are to be modelled, this value may be doubled in order to allow for a 50 % overlap between adjacent bands.

Within this estimated range of 38 to 200 filter bands, the optimum number of filter bands also depends on the kind of output parameters to be calculated. If time-domain analysis, which is very likely performed in the auditory system, is neglected in the perceptual model, this shortcoming can be partly compensated by increasing the number of filter bands. Hence, even under the assumption that a number of 38 auditory filters would sufficiently represent the information accessible for the auditory system, a higher number of filter bands may be appropriate.

d) Comparison of Statistical and Deterministic Models

In theory, the statistically based approach of calculating a detection probability should yield better predictions than other approaches when the distortion is near threshold. When the distortion is clearly above threshold, the loudness of the distortion is a better quality indicator. Hence, it can be expected that a measure that combines both, detection probability for distortions close to threshold, and noise loudness for distortions above threshold, yields better predictions for the overall quality range than measures that use only one of these approaches. It can be shown that PAQM, as a result of the loudness compression, already comes close to such an approach. For small distortions, the noise disturbance can, theoretically, be monotonically mapped to a detection probability as used in [COL93]. Moreover, both noise loudness and detection probability result in a kind of noise-to-mask ratio for the case of small distortions. Therefore, it might be concluded that near masked threshold all known perceptual measurement methods essentially produce the same kind of output parameter. This also implies that the differences in the performance of the methods when applied on only slightly distorted signals are more likely to originate from the peripheral ear models and the (spectral and temporal) averaging strategies than from the kind of output parameters they produce.

- **Detection Probability for Small Distortions**

In POM [COL93], the probability for detecting the difference between two signals within one filter channel is calculated from their local energies, E_1 and E_2, according to

$$P(n) = 1 - 10^{-K \cdot 10^{\,0.2 \cdot \Delta SL(n)}} \qquad (2.50)$$

where $\Delta SL(n) = SL_2(n) - SL_1(n)$

$SL_{1,2}(n) = \log_{10}\left[E_{1,2}(n) + E_{thres}(n)\right]$ is the local energy level,

n denotes the filter channel,

and K determines the steepness of the threshold function.

The local probability of non-detection is

$$\overline{P(n)} = 10^{-K \cdot 10^{0.2 \cdot \Delta SL(n)}}, \tag{2.51}$$

which, after some manipulations, can be written as

$$\overline{P(n)} = 10^{-K \cdot \left[\frac{E_2(n)+E_{thres}(n)}{E_1(n)+E_{thres}(n)}\right]^{0.2}} \tag{2.52}$$

The total probability of non-detection is thus

$$\overline{P_{total}} = \prod_{\forall n} 10^{-K \cdot \left[\frac{E_2(n)+E_{thres}(n)}{E_1(n)+E_{thres}(n)}\right]^{0.2}} = 10^{-K \cdot \sum_{\forall n}\left[\frac{E_2(n)+E_{thres}(n)}{E_1(n)+E_{thres}(n)}\right]^{0.2}} \tag{2.53}$$

and the total probability of detection is given by

$$P_{total} = 1 - 10^{-K \cdot \sum_{\forall n}\left[\frac{E_2(n)+E_{thres}(n)}{E_1(n)+E_{thres}(n)}\right]^{0.2}}, \tag{2.54}$$

which can also be written as

$$P_{total} = 1 - 10^{-K \cdot \sum_{\forall n}\left[1+\frac{E_2(n)-E_1(n)}{E_1(n)+E_{thres}(n)}\right]^{0.2}}. \tag{2.55}$$

If the excitation difference is small, this can be approximated by

$$P_{total} = 1 - 10^{-K \cdot \sum_{\forall n}\left[1+0.2 \cdot \frac{E_2(n)-E_1(n)}{E_1(n)+E_{thres}(n)}\right]}, \tag{2.56}$$

and, if the number of filter bands is N,

$$P_{total} = 1 - 10^{-K \cdot N} \cdot 10^{-0.2 \cdot K \cdot \sum_{n=0}^{N-1}\left[\frac{E_2(n)-E_1(n)}{E_1(n)+E_{thres}(n)}\right]}. \tag{2.57}$$

This means that the calculated detection probability is a function of the average relative excitation difference.

- **Noise Disturbance (PAQM) for Small Distortions**

In [THI94a] it has been shown that for small distortions the compressed noise loudness used in PAQM [BEE92] corresponds to an average relative excitation difference as well. PAQM is based on a specific loudness difference using Zwickers loudness formula (Eq. 2.45) with an altered compression exponent γ:

$$\Delta N' = k \cdot \left(\frac{1}{s} \cdot \frac{E_{thres}}{E_0} \right)^{\gamma} \cdot \left[\left(1 - s + s \cdot \frac{E_{test}}{E_{thres}} \right)^{\gamma} - \left(1 - s + s \cdot \frac{E_{ref}}{E_{thres}} \right)^{\gamma} \right]. \tag{2.58}$$

Since γ is very small in PAQM, this can be approximated by

$$\Delta N' = k \cdot \left(\frac{1}{s} \cdot \frac{E_{thres}}{E_0} \right)^{\gamma} \cdot \gamma \cdot \left[\ln\left(1 - s + s \cdot \frac{E_{test}}{E_{thres}} \right) - \ln\left(1 - s + s \cdot \frac{E_{ref}}{E_{thres}} \right) \right], \tag{2.59}$$

which, after some manipulations, can be written as

$$\Delta N' = k \cdot \left(\frac{1}{s} \cdot \frac{E_{thres}}{E_0} \right)^{\gamma} \cdot \gamma \cdot \ln\left(1 + \frac{E_{test} - E_{ref}}{\frac{1-s}{s} \cdot E_{thres} + E_{ref}} \right). \tag{2.60}$$

For small distortions this can be approximated by

$$\Delta N' = k \cdot \left(\frac{1}{s} \cdot \frac{E_{thres}}{E_0} \right)^{\gamma} \cdot \gamma \cdot \frac{E_{test} - E_{ref}}{\frac{1-s}{s} \cdot E_{thres} + E_{ref}}. \tag{2.61}$$

The noise disturbance is given by

$$ND = \sum_{\forall n} \Delta N'(n) = k \cdot \gamma \cdot \sum_{\forall n} \left(\frac{1}{s} \cdot \frac{E_{thres}(n)}{E_0} \right)^{\gamma} \cdot \frac{E_{test}(n) - E_{ref}(n)}{\frac{1-s}{s} \cdot E_{thres}(n) + E_{ref}(n)}. \tag{2.62}$$

Except for a frequency dependent weighting

$$w(n) = k \cdot \gamma \cdot \left(\frac{1}{s} \cdot \frac{E_{thres}(n)}{E_0} \right)^{\gamma}, \tag{2.63}$$

and an attenuation of the absolute threshold

$$E_{thres}'(n) = \frac{1-s}{s} \cdot E_{thres}(n), \tag{2.64}$$

the noise disturbance for small degradations is thus a monotonic function of the relative excitation difference.

e) Loudness Compression

As derived above, one effect of the loudness compression used in PAQM is that it transforms the noise loudness into a kind of detection probability when the distortions are small. On the other hand, it does not only result in a loudness reduction of the distortions by the original signal in the range close to the masked threshold (which might model partial masking), but also for distortions far above masked threshold. The latter is clearly not desired, even though it might not be relevant for the quality estimation of high quality audio codecs, as such large distortions are probably beyond the range of the quality scale applied.

The loudness compression could be explained as an additional data compression that is performed when storing audio information in the short-time memory. This may be appropriate, because in a listening test for codec comparison according to the ITU-R recommendation BS 1116 [ITU97] the signals are not presented at the same time. Therefore, the processed signal has to be compared with the memorised original signal.

2.4.7 Shortcomings of Existing Models

Most existing models use an FFT for the time-frequency decomposition, and therefore require a scale transform from the linear frequency scale given by the FFT to the highly non-linear frequency scale used in auditory models. This results in a sub-optimal product of time and frequency resolution. Consequently, in one of both domains (normally in the time domain) a potentially insufficient resolution has to be accepted. As a result of the insufficient temporal resolution, most perceptual measurement methods are solely based on models for steady-state effects, and temporal effects, like modulations, are not incorporated in the models (however, this does not mean that such effects are completely ignored, as changes in the temporal structure are always linked to a corresponding change in the spectrum).

The missing evaluation of temporal effects makes it also necessary to model effects in the frequency domain which are best explained in the time domain, for example, the masking asymmetry between tonal and noise-like maskers and the non-linear additivity of masking.

None of the existing models provides a consistent transition between masking, partial masking and full perception of the distortions. Moreover, the differences between the perception of additive (non-linear) distortions and changes in the spectral envelope (linear distortions) are not fully reflected in any of the existing models. Even though this effect, which can be regarded as a kind of *auditory streaming*, is partly taken into account by the asymmetry introduced in late versions of PAQM [BEE94], there is not yet any satisfactory model of it.

3. Application of Filter Banks

Whereas filterbank-based algorithms have been frequently used in measurement methods applied to speech codecs (e. g. [KAR85], [HOL93], [HAN96]), almost all measurement methods applied to high quality audio codecs have used FFT based algorithms. One of the few exceptions was an algorithm proposed in 1989 by Kapust [KAP89] which was based on a combination of FIR filters and FFTs. Because of its high computational complexity, this algorithm was never verified against a sufficiently large set of “real-world data” (i. e. coded audio signals for which the perceived quality was known from subjective tests). It took until 1996 before two other filterbank-based measurement methods were published for which the computational complexity was low enough to allow for an extensive verification ([THI96], [SPO96]).

The main reason for which filter banks have not been applied in the evaluation of audio codecs for a long time is probably the comparably higher computational effort: as high quality audio signals are usually sampled at six times the rate of speech signals (48 kHz versus 8 kHz) and the frequency range to be evaluated is extended (24 critical bands instead of 15), the same algorithm would have approximately ten times the computational complexity when applied to high quality audio instead of speech.

3.1 Comparison with FFT-Based Methods

Filterbank-based ear models are more closely related to human sound processing than FFT-based ear models not only in the time domain but also in the spectral domain. Whereas the advantages in terms of temporal resolution are obvious and widely known, the advantages of filter banks in the processing of steady-state signals are often not recognised.

In FFT based ear models, the input signals are first decomposed into linear short-time spectra, which then are mapped to an auditory frequency scale. This so-called frequency-to-pitch mapping is carried out by summing up neighbouring spectral coefficients within equidistant segments of the auditory frequency scale. The result of this procedure is a staircase approximation of the frequency-to-pitch mapping function. In order to achieve equidistant segments, the width of the frequency bands, Δz must be much larger than the bandwidth corresponding to one spectral coefficient of the FFT, Δf. This requires a large number of spectral coefficients (i. e. a large frame length of the FFT) and a large width of the segments on the auditory frequency scale. Both

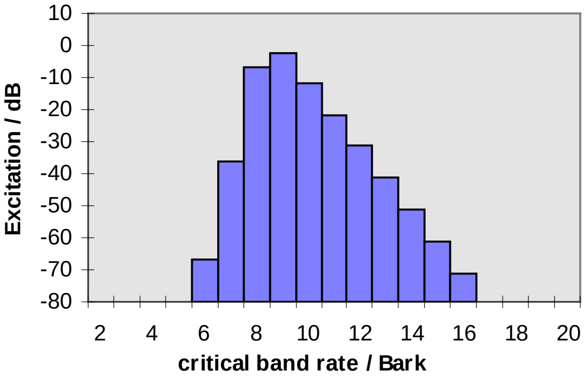


Fig. 3.1: Excitation pattern of a pure tone for an FFT based measurement algorithm that groups the spectral coefficients into rectangular bands.

together result in a small number of almost rectangular frequency bands in the signal representation (Figure 3.1). In the following considerations of spectral and temporal resolution, spectral resolution is defined as the equivalent rectangular bandwidth of a filter (or an FFT-line), which is given by the area below its amplitude response divided by its maximum value. Temporal resolution is defined accordingly by the equivalent rectangular bandwidth of the envelope of its impulse response.

3.1.1 Temporal Resolution

As an accurate frequency-to-pitch mapping requires a large number of spectral coefficients, the temporal resolution in an FFT-based ear model is limited by the corresponding frame length of the FFT. Typical values of the frame length are 1024 or 2048 samples (at a sampling rate of 48 kHz). This corresponds to a temporal resolution of 10 to 20 ms, which allows for a reasonably accurate model of forward masking but is not sufficient for the modelling of backward masking. Apart from its influence on the modelling of masking, the insufficient temporal resolution also destroys the temporal fine structure of the signal which accounts for roughness sensation and probably also contributes to other effects like the masking asymmetry between tonal and noise-like signals.

3.1.2 Spectral Resolution

In the frequency domain, the shortcomings of FFT based ear models are mainly caused by the staircase approximation of the frequency-to-pitch mapping. This can be illustrated by the following example: a single tone which is almost exactly on the border between two filter bands can either belong to the lower or the upper frequency band, depending on a very small frequency shift in the one or the other direction. It is thus treated as if its frequency would be either half a filter bandwidth higher or half a filter bandwidth lower than its actual frequency. The computed masking curves would always have the desired shape but can be shifted in frequency by almost one filter bandwidth in either direction.

This frequency shift can lead to severe deviations when estimating the amount of masking among narrow band signals. Especially at the steep lower slope of the masking pattern. Assume that the masker frequency is only slightly above the border between two filter bands and the frequency of the maskee just below the frequency border between two neighbouring filter bands. The distance between masker and maskee is then overestimated by one filter bandwidth. As a result, the amount of masking will be underestimated according to the steepness of the masking curve:

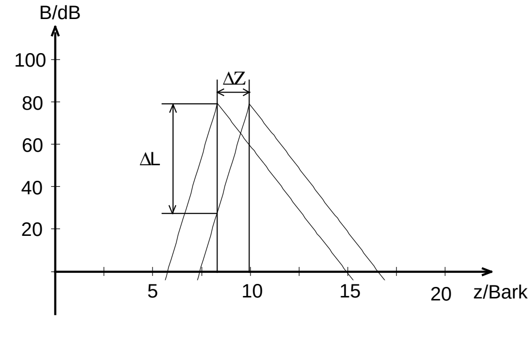


Fig. 3.2: Uncertainty in the estimation of the amount of masking between two narrow band signals.

$$\Delta L_{MAX} = S \cdot \Delta z \qquad \begin{array}{|ll} S: & \text{excitation slope (ca. } 27\,\frac{\text{dB}}{\text{Bark}}\text{)}. \\ \Delta z: & \text{filter bandwidth in Bark} \end{array} \tag{3.1}$$

In the opposite case, the amount of masking will be overestimated by approximately the same value (see also Figure 3.2).

In filterbank-based ear models the frequency-to-pitch mapping is implicitly included in the distribution of the filters. No grouping of frequency bands is necessary and thus the occurrence of rectangular filter bands is avoided. If the slopes of the masking curves are directly realised by the shape of the filters, the relation between masker and maskee is always correct, even when the filter bands are comparably wide. Therefore, especially when using a low number of filter bands, filterbank-based ear models are, in theory, clearly superior to FFT based ear models.

3.1.3 Summary

When using FFT-based ear models, there is always a trade-off between temporal resolution and accuracy in the spectral domain. Due to the non-linear frequency-to-pitch mapping, the product of temporal and spectral resolution is more than twenty times worse than theoretically possible at high centre frequencies (this value is determined by the ratio of the bandwidths between the auditory filters at very high and very low centre frequencies). For example, according to the Bark-scale given in [ZWI67] the critical bandwidth is 100 Hz at low centre frequencies and 2500 Hz at 10 kHz. If the transform length of an FFT is chosen in order to match this spectral resolution even at low frequencies, it will be 25 times higher than necessary at 10 kHz. Consequently, due to the inversely proportional relationship between spectral and temporal resolution, the temporal resolution will be 25 times worse than possible at this point. This can be looked upon as an unwanted data reduction. Because of this trade-off, the influence of temporal and spectral resolution on the performance of an ear model in real-world applications cannot be tested with an FFT-based model.

In a filterbank-based model, the product of temporal and spectral resolution can be as low as the minimum value given by the Heisenberg relation. Since this value is much smaller than the value one can expect to find in the human auditory system, it offers the chance of investigating the influence of temporal and spectral resolution without interactions. The outcome might be that the temporal resolution achieved with FFT-based models is sufficient, but it is also possible that an enhanced temporal resolution yields significant improvements. Within this work it will be shown that the latter is true.

3.2 Requirements on the Filters

In several aspects, the requirements on filter characteristics for measurement purposes clearly differ from the requirements on filter characteristics used for data reduction. Orthogonality is not required and not even desirable. Perfect reconstruction is not required, but the filtered signal representations should preserve all relevant information that may be contained in the original signal. The filter slopes do not need to be steeper than auditory filter shapes. On the other hand, a high stop-band attenuation is essential and cannot be replaced by aliasing cancellation properties (as often used in audio coding). In order to ensure that the model behaviour is independent of the position of the signal components relative to the filter bands, the filters should not be flat in the pass-band but continuously change between pass-band and stop-band.

These criteria also apply when FFT-based models are incorporated. For example, the required high stop-band attenuation implies that, unlike commonly used signal analysis techniques, a simple raised-cosine window is better than a Hamming-window because the latter one achieves a fast transition between pass-band and stop-band (which is not required) at the cost of stop-band attenuation (which is essential).

3.3 Filter Banks used in Perceptual Models

There are numerous filter bank techniques used in auditory models though most of them never have been applied in the particular field of perceptual quality measurement.

3.3.1 One-Third-Octave Filters

The filter bank schemes with the longest history not only in auditory modelling but in all fields of acoustics are one-third octave filter banks. They were originally realised as analogue filters and can nowadays be realised either as FIR or as IIR filters. In applications where temporal resolution is not important, they may also be realised using an FFT. The term one-third-octave filter does not relate to an exactly defined filter shape, but to a certain tolerance scheme into which the filter characteristics have to fit.

The filter bandwidths are proportional to the centre frequencies and thus correspond to an auditory frequency scale in the upper frequency range. By combining neighbouring filters at low centre frequencies, the Bark-scale can be approximated. This technique is used in the loudness measurement scheme by Zwicker which is standardised in ISO 532. However, one-third-octave filters yield only a rough approximation of auditory filter shapes, and their approximately rectangular filter shapes are not always desireable in auditory modelling (cf. Section 3.1.2).

3.3.2 FIR Filters

The main advantage of FIR filters is the fact that they can be easily designed to fit any required frequency response by simply mirroring the frequency response at the 0 Hz axis, sampling the frequency response and calculating its inverse discrete Fourier transform. The inverse DFT directly produces the coefficients of the corresponding linear phase FIR filter. The length of the DFT determines the accuracy with which the frequency response is approximated.

The linear phase allows to achieve equal phases for all filter bands by simply delaying the output signals by the difference between the delays of the current filter and the filter with the longest impulse response. This eases some parts of the post-processing, even though linear phase filters actually do not reflect the properties of auditory filters (which rather are assumed to have minimum phase).

The main problem with FIR filters is the comparably high computational effort. It can be reduced by certain computation techniques or by choosing particular filter structures.

a) FIR Filters and Fast Forward Convolution

One possibility of reducing the computational costs of FIR filters is the use of the so-called *fast forward convolution* (FFC). This technique is used for example in the measurement scheme described in [Spo96].

A fast forward convolution uses the correspondence of convolution and multiplication in the frequency and time domain respectively. Any filtering can be looked at as a convolution of the input signal with the impulse response of the filter. It can therefore also be performed by multiplying the Fourier transform of the signal with the frequency response of the filter and transforming the result back to the time domain by an inverse FT. Because of the periodicity of the DFT, when using a N-point DFT for the computation of a filter with N tabs, only one point of the result represents a useful output value of the filter. This would, of course, make no sense as a DFT is much more time consuming than the computation of one filtered output value. But when using a M-point DFT for a filter with N tabs and $M>>N$ and appending $M-N$ zeros to the filter's impulse response, $M-N+1$ useful output values can be calculated with one DFT. The optimum M is found by minimising the number of arithmetic operations required for the computation of one output value. This number is proportional to the ratio $[ld(M)\cdot M]/(M-N)$.

As the FFC is only efficient with a large FFT length, signal delays of more than one second can occur and a large amount of memory is required.

3.3.3 BARK-Transform

A measurement method based on the so-called BARK-transform was first published in 1989 by Kapust [KAP89] and described in detail in [KAP93]. It originally used multiple FFTs with different window functions to approximate the spectral and temporal resolution of the human auditory filters. The frequency-to-pitch mapping is therefore approximated by a staircase function. For a decomposition into 25 frequency bands it uses 11 different window functions. The frame length is 512 for all FFTs and the window functions are designed in order to approximate rectangular band pass characteristics. Some of the window functions are longer than the frame, which is possible because of the linearity of the FFT (*"polyphase FFT"* [CRO83]).

As only few of the output coefficients of the FFT spectra are used, the FFTs were finally replaced by complex FIR filters. These FIR filters are realised using a recursive algorithm [GOE68] that needs only half as much computations as compared with straightforward FIR implementations.

Nevertheless, the computational effort of this method is still extremely high. Even though the algorithm is twice as fast as a straightforward FIR filter implementation, it is much slower than a fast-forward-convolution. The staircase approximation of the frequency-to-pitch mapping and the use of rectangular frequency bands are additional disadvantages (cf. Section 3.1.2).

3.3.4 IIR Filters

Compared to DFT algorithms or FIR filters, the structure of an IIR filter corresponds better to the way natural systems work. Thus, one may expect that auditory filters can be more efficiently modelled by IIR filters than by FIR filters. Using the correspondence between IIR filters and analogue filters, it can be shown that such filters match the auditory frequency scale better than FIR filters. When expressed on a logarithmic frequency scale, the slopes of an analogue band pass of a given filter order are independent of the chosen centre frequencies. Analogue filters may therefore be considered as best fitting for frequency analysis on a logarithmic frequency scale. This also applies for digital IIR filters because any analogue filter of order N can be approximated by an IIR filter of the same order [TIE83] (though the approximation does not hold for frequencies close to half the sampling rate). As the auditory frequency scale is closer to a logarithmic scale than to a linear scale, it can be concluded that the use of IIR filters may be advantageous over FIR filters.

As the steep upper slope of auditory filters is assumed to be constant (when expressed in dB versus critical band rate) and the upper range of the critical band scale corresponds to a logarithmic frequency scale, filters with constant slopes are desireable. Typically, high pass and low pass filters optimised for a maximum flatness of the pass-band (Butterworth filters) show such a constant slope. The slope of such a band-pass filter, consisting of Butterworth high and low pass of order N, is 6·N dB per octave. As one octave corresponds to slightly less than four critical bands and the maximum slope of an auditory filter is in the range of 30 dB/Bark, a 40th order IIR filter provides sufficiently steep slopes. As can be seen in Figure 3.3, a 40th order band pass consisting of a low pass and a high pass with Butterworth characteristics yields a rather good approximation of an auditory filter at the upper slope. The lower slope in Figure 3.3 seems to be far too steep compared to the auditory filter, but at low signal levels the lower auditory filter slope might become almost as steep as the upper one. Hence, a filter with symmetric slopes is a good starting point for an auditory filter bank.

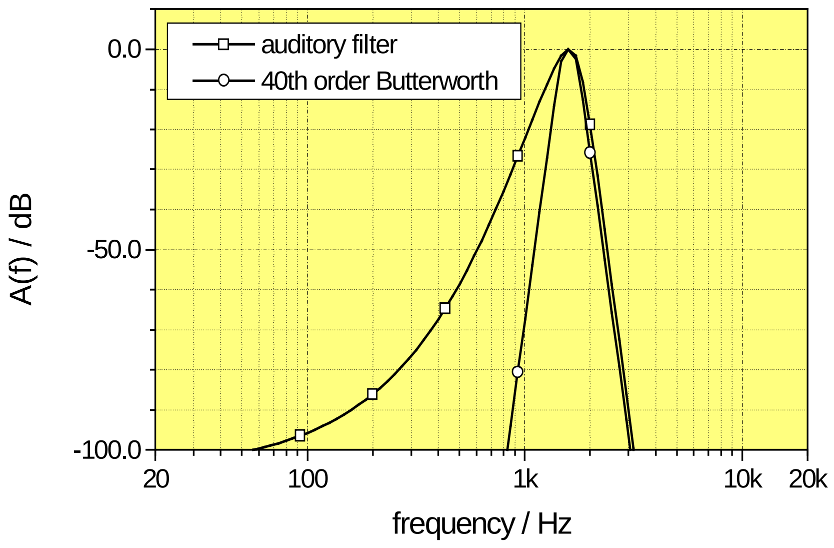


Fig. 3.3: Frequency response of an auditory filter at a centre frequency of 1600 Hz and a 40th order analogue filter consisting of a low pass and a high pass with Butterworth characteristic.

In the neighbourhood of poles and zeros the slopes of a filter can be much steeper than the above mentioned value of 6·N dB/octave (which is valid at a large distance from poles and zeros). Hence, in practice, remarkably lower filter orders are sufficient.

The main problem with IIR filters is that they are, in general, difficult to design. In case perfect (i. e. rectangular) band pass characteristics are desired, there are several approaches known from the design of analogue filters. The coefficients of the analogue filters can be transformed to coefficients for digital IIR filters by the *bilinear transform* [TIE83]. When a perfectly flat pass-band is desired (*Butterworth* approximation) and a constant slope rate is not considered as important, a 12th order band pass can already show sufficiently steep slopes within almost the full dynamic range of the ear (Figure 3.4). If some ripples in the pass-band can be tolerated (*Chebychev* approximation) even lower filter orders may be used.

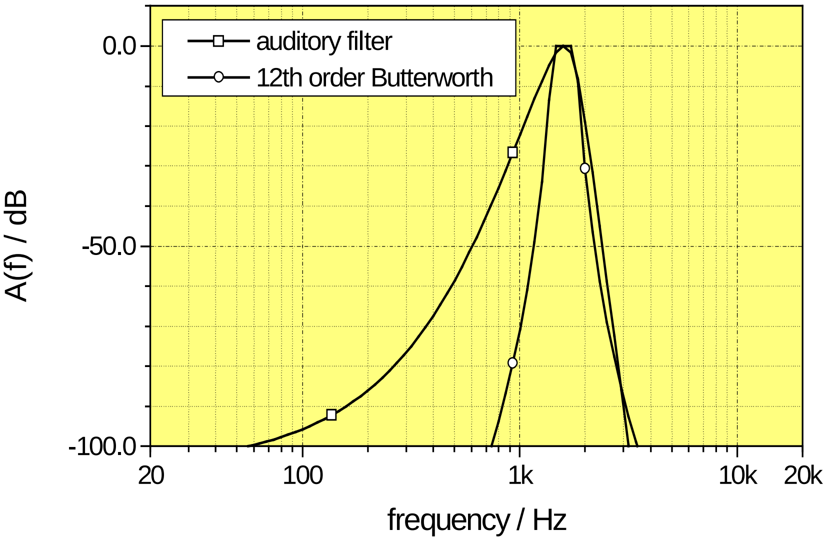


Fig. 3.4: Frequency response of an auditory filter at a centre frequency of 1600 Hz and a 12th order analogue band pass with Butterworth characteristic.

3.3.5 Warped Filters

The concept of signal processing on a warped frequency scale has frequently been used in speech [STR80] and audio coding [HÄR97]. In [LAI90] the term *FAMlet* (FAM: Frequency and Amplitude Modulated functions) was introduced for an orthogonal transform based on frequency warping. A brief review of warped filtering is given in [HÄR97].

When a signal passes through a delay line, the values at each element of the delay line correspond to the preceding values of the time signal. When the delay line is replaced by a chain of (identical) all pass filters, the momentary values at each element of the all pass chain correspond to the preceding values of a virtual, frequency warped time signal. The frequency warping is given by the phase response of the all pass. The result of a spectral analysis of these all pass outputs corresponds to a frequency analysis on a warped frequency scale.

This can easily be understood by comparing the delay element used in a linear frequency analysis with an all pass element used in the warped frequency analysis. The transfer function of the delay element is Z^{-1} or $e^{-j\omega T}$ (where T is the inverse of the sampling rate) The transfer function of an arbitrary all pass is $e^{j\varphi(\omega)}$. By comparing the arguments, it becomes obvious that an all pass can be considered as a delay element where either the time scale or the frequency scale is warped:

$$\varphi(\omega)=\omega'(\omega)T=\omega T'(\omega)\ , \tag{3.2}$$

$$T'(\omega)\ =\varphi(\omega)/\omega \tag{3.3}$$

and

$$\omega'(\omega)=\varphi(\omega)/T. \tag{3.4}$$

This means that for a frequency component $A(\omega_1,t)$ the delay of one all pass element is $\varphi(\omega_1)/\omega_1$ instead of T_{samp} (Eq. 3.3). The momentary values along the all pass chain correspond to a time signal with a frequency of $\omega_1'=\varphi(\omega_1)/T_{samp}$ (Eq. 3.4). Thus, when applying a Fourier transform to the momentary values along the all pass chain, the signal components at the frequencies ω_n appear in the spectral coefficients for ω_n'. Eq. (3.3) can be interpreted as a dispersion of the time signal. This means that different frequency components travel at a different speed. It is clear that frequency warping can only be used with FIR filters because IIR filters would require an infinite length of the all pass chain. However, the resulting filter characteristics are IIR filters because the impulse response of an all pass (with the exception of a simple delay) is infinite.

When frequency warping is used for coding purposes, orthogonality of the coefficients of the warped spectrum can be achieved by applying a pre-filter with a frequency response given by

$$W(\omega)=\sqrt{\frac{d}{df}\varphi(\omega)} \tag{3.5}$$

and resetting the all pass chain after each frame. For measurement purposes it is more convenient to omit the pre-filter and run the all pass chain continuously.

By designing an appropriate window function it is possible to define the spectral characteristics of the filter bands in the (warped) frequency domain. If the phase response of the all pass elements corresponds to the Bark-scale, a window function of the shape

$$Win(\omega)=\sin^2\left(\pi\cdot\frac{n}{N}\right)\cdot\frac{1}{1+\left(a\cdot\frac{n}{N}\right)^2} \tag{3.6}$$

will provide exponential slopes like those found in the excitation patterns of the auditory system. In this equation, a determines the slope rate and N is the window length in samples. In principle, it is even possible to realise asymmetric slopes by defining a complex window function

$$Win(\omega) = \sin^2\left(\pi \cdot \frac{n}{N}\right) \cdot \frac{1}{\left(1 + ja \cdot \frac{n}{N}\right) \cdot \left(1 - jb \cdot \frac{n}{N}\right)}, \qquad (3.7)$$

where a determines the lower slope rate and b determines the upper slope rate. In practice, this would raise problems at low frequencies because due to the flat upper slope of the excitation, the slope of the mirror frequency of a low frequency component will reach into the positive part of the spectrum.

Even though in theory almost any frequency warping function could be realised, the concept of frequency warping is only advantageous when the warping function can be realised with low computational effort, i. e. the order of the all pass elements should be low. Fortunately, already a first order digital all pass can yield a reasonably good approximation of the Bark-scale, when its characteristic frequency is chosen accordingly. The phase response of a first order all pass of the form

$$A(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \Bigg|_{z = e^{j \cdot 2\pi \frac{f}{f_{samp}}}} \qquad (3.8)$$

is

$$\varphi(f) = \arctan\left(\frac{\left(\lambda^2 - 1\right) \cdot \sin\left(2\pi \frac{f}{f_{samp}}\right)}{\left(\lambda^2 + 1\right) \cdot \cos\left(2\pi \frac{f}{f_{samp}}\right) - 2\lambda}\right) \qquad (3.9)$$

The phase runs from zero at low frequencies to -π at half the sampling rate. Since the Bark scale saturates at a value of approximately 25 Bark, and, according to Eq. (3.4), the warped frequency is proportional to the phase response of the all pass elements, the frequency to Bark mapping is given by

$$Z(f) = -\frac{Z_{max}}{\pi} \cdot \arctan\left(\frac{\left(\lambda^2 - 1\right) \cdot \sin\left(2\pi \frac{f}{f_{samp}}\right)}{\left(\lambda^2 + 1\right) \cdot \cos\left(2\pi \frac{f}{f_{samp}}\right) - 2\lambda}\right), \qquad (3.10)$$

where $Z_{max} \approx 25$.

In [SMI95], a formula is proposed for calculating the value of λ that yields the best possible approximation of the Bark scale:

$$\lambda_{opt} = 1.0211 \cdot \sqrt{\frac{2}{\pi} \cdot \arctan\left(0.076 \cdot f_{samp}\right)} - 0.19877 . \qquad (3.11)$$

The scaling factor Z_{max} can be found by calculating a linear regression. The value of λ calculated by Eq. (3.11) gives the best possible approximation in the low frequency range. In order to obtain the best possible approximation over the full scale in terms of a minimum squared error in the Bark-domain, λ has to be slightly smaller. For a sampling rate of 48 kHz, Eq. (3.11) yields a value of 0.822 for λ and Z_{max} becomes *23.5*. Optimisation of λ results in a value of 0.794 for λ and *24.4* for Z_{max}. The correlation between the resulting approximation for the Bark-scale and the commonly used approximation proposed by Zwicker and Terhardt [ZWI80] is above 0.998 (see Figure 3.5).

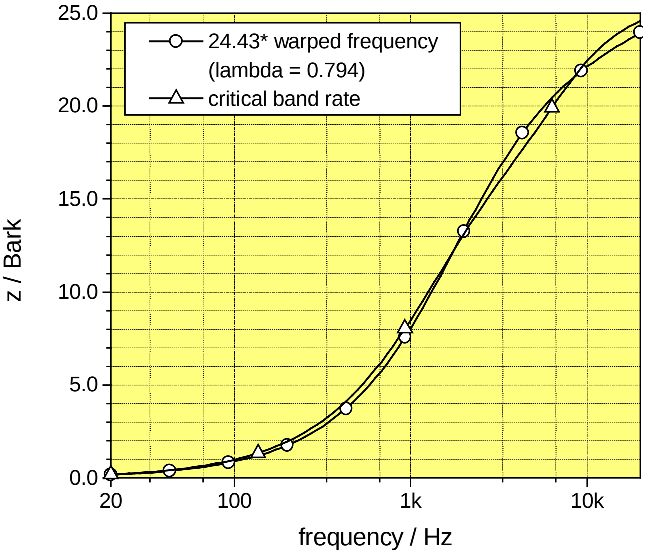


Fig. 3.5: Approximation for the Bark-scale by the phase response of a first order low pass compared to the commonly used approximation given by Zwicker and Terhardt [ZWI80].

The advantages of the concept of frequency warping are its computational efficiency and the fact that it is easy to implement and theoretically well defined. The disadvantages are the restriction of the frequency to pitch mapping to functions of the form Eq. (3.10) and the restriction of the number of filter channels to powers of two (of course both restrictions apply only when high computational efficiency is required).

3.3.6 FTT (Fourier-Time-Transform)

The *Fourier-Time-Transform* (FTT) has been introduced by Terhardt [TER85] as an auditory equivalent to the Fourier transform commonly used in signal processing. This (postulated) correspondence is also the reason for the somewhat misleading name of this filter bank. The Fourier-Time-Transform replaces the windowing functions of the DFT, which usually are symmetric and time-constrained, by single-sided exponentials. The FTT is realised by multiplying the input signal with a sine-wave of the respective centre frequency and filtering the result by a first order low pass filter. The multiplication shifts the spectral characteristic of the low pass filter to the centre frequency of the filter channel. The centre frequencies can be chosen arbitrarily and are normally equally distributed over the Bark scale as defined by Zwicker [ZWI67] . The bandwidths of the filters are given by the cut-off frequency of the low pass and are chosen accordingly.

The FTT concept can be generalised by replacing the first order low pass by an arbitrary low pass. In this case, it is identical to the *Gammatone* concept (see Section 3.3.7).

An (approximately) inverse transform to the FTT was proposed by Mummert [MUM91].

3.3.7 Gammatone Filter Banks

The *Gammatone filter bank* corresponds to an extension of the FTT to filter orders higher than one. The name "*gammatone*" results from the similarity between the function

$$g(t,a) = e^{-t} \cdot t^{a-1}, \tag{3.12}$$

which describes the impulse response of these filters, with the integrand in the definition of the *gamma-function*

$$\Gamma(x) = \int_0^\infty e^{-t} \cdot t^{x-1} dt \qquad \text{([BRO89], pp. 331).} \tag{3.13}$$

From this definition, the name *Gammatone filter bank* holds only when the low pass consists of a cascade of first order low pass filters. Like in the FTT concept, often a more general definition is used, allowing arbitrary low pass filters. Such generalised *Gammatone filter banks* are identical to generalised *FTTs* described in the previous paragraph.

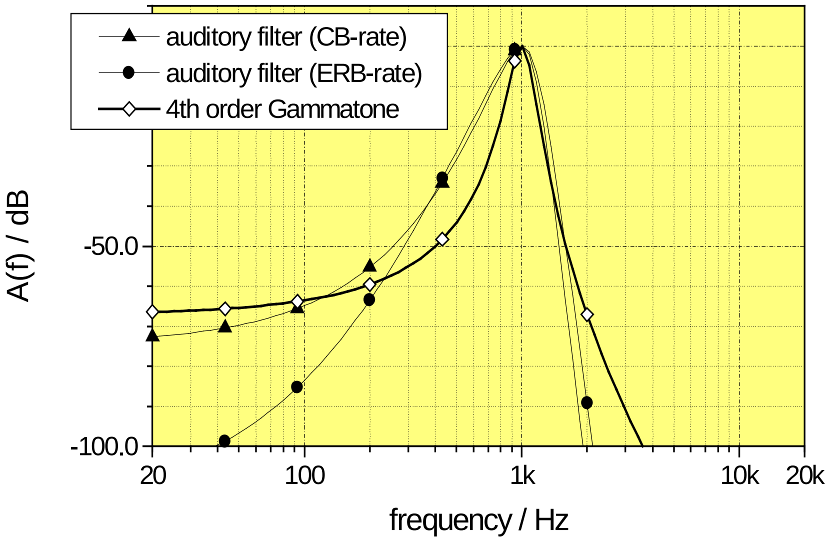


Fig. 3.6:** Fourth order* **Gammatone** ***function compared to an auditory filter shape for intermediate signal levels for a narrow band signal at 1 kHz. The auditory filter shape is approximated by a two sided exponential, related to the Bark-scale and related to the ERB-scale, respectively.

In many publications, a fourth order *Gammatone filter bank* is considered as being a good approximation for the auditory filter functions. However, this assumption apparently only holds for a limited dynamic range of approximately 50 dB (see Figure 3.6). The differences on the ascending slope in Figure 3.6 are tolerable because for higher levels the excitation flattens off and the stop-band attenuation of the filter will always be sufficient in the range above the absolute threshold. The descending slope starts to flatten off for attenuations above 50 dB.

In order to keep the stop-band attenuation higher than the dynamic range of the auditory system (ca. 100 dB), filter orders between 6 and 8 seem to be appropriate (Figure 3.7). Of course this only holds under the assumption that the auditory filter shapes can (at least at the descending slope) be represented by exponentials in the Bark- or ERB-domain as postulated by Zwicker [ZWI67] and Terhardt [TER79]. As these data are not very certain, the fourth order filters as shown in Figure 3.6 may already be sufficient.

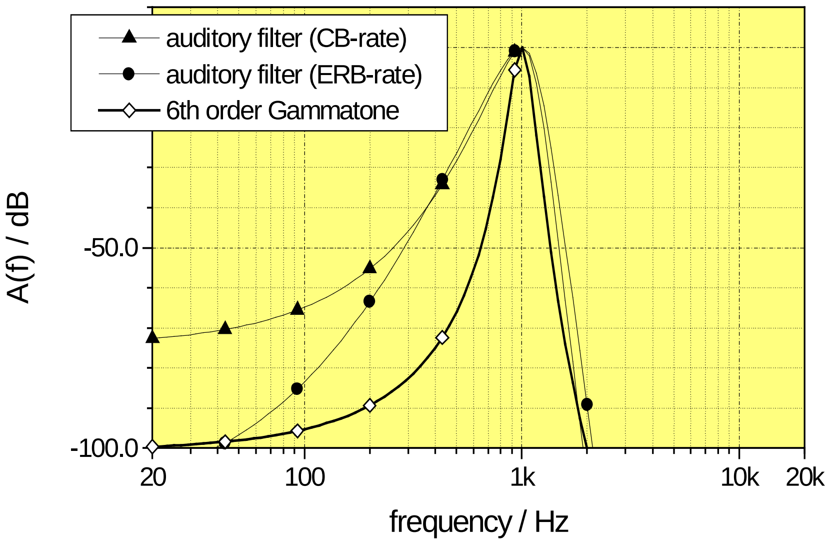


Fig. 3.7: Sixth order* Gammatone *function compared to an auditory filter shape for intermediate signal levels for a narrow band signal at 1 kHz. The auditory filter shape is approximated by a two sided exponential, related to the Bark-scale and related to the ERB-scale, respectively.

3.4 Wavelet-Transforms

Wavelets are originally defined by the property that the impulse responses of all filters look identical except for a scaling factor applied to the time axis. According to this definition, the centre frequencies of a wavelet filter bank would always be distributed logarithmically and thus only fit for frequencies above 500 Hz to an auditory frequency scale. The so-called wavelet-packet-transforms extend the definition of the wavelet concept and allow to approximate an auditory frequency scale in the full frequency range. As compared to filter banks in general, wavelet-concepts restrict the filter characteristics and filter distributions to combinations that provide orthogonality and allow perfect reconstruction. Both is advantageous for data reduction algorithms but not necessary for signal analysis. On the other hand, these restrictions limit the accuracy within which the shapes and distribution of auditory filters can be approximated. For this reason, wavelet concepts are not considered in detail in the present work.

3.5 Design of a Non-Linear Filter Bank

In this section, a new filter bank will be introduced which uses a frequency-sampling structure [OPP75] to implement time-varying FIR filters. It combines a high degree of flexibility with an accurate model of auditory excitations and low computational complexity.

3.5.1 Modelling Level-Dependent Excitation Patterns

From filter theory it is clear that the level dependence of the frequency response of auditory filters will be reflected in their temporal behaviour. In an auditory model, this correspondence between spectral and temporal resolution is retained through all linear processing steps, i.e. until the rectification. When the filter slopes are modelled by post-processing after the rectification process, this relation is lost. Furthermore, there may be interactions between different signal components which are not within the same critical band but have overlapping excitation slopes. Such an interaction might already occur at an early stage, where the temporal structure of the signals is still of influence. Modelling the auditory filter slopes within the filter bank is thus regarded as an advantage. Due to the level dependence of the filter slopes, such a filter bank would be non-linear (the relation $f(a+b)=f(a)+f(b)$ would not be valid anymore). Four different approaches for the modelling of level-dependent filter slopes were considered.

a) First Approach: Backward Adaptation of Filter Coefficients

The most straightforward way of modelling level-dependent excitation patterns would be to change the coefficients of the band pass filters depending on the levels at the filter outputs. The filter characteristics can be derived from the level-dependent shapes of excitation patterns given in [TER79]. However, in the case of level-dependent slopes, the transformation from excitation patterns to auditory filter shapes is not as simple as in the case of constant slopes (where the excitation patterns simply are inverted). The slopes cannot be expressed as a function of the signal level anymore (which is not known) but have to be expressed as a function of the output level of the filter itself. When using the excitation slopes from [TER79], the corresponding auditory filter shapes become:

$$A_{auditoy\ filter}(z)=\begin{cases}-S_1\,\dfrac{dB}{Bark}\cdot(z-z_0) & \mid z>z_0\\[2ex] S_2\,\dfrac{dB}{Bark}\cdot\dfrac{(z-z_0)}{1-0{,}2\cdot(z-z_0)}-L\cdot\left(\dfrac{0{,}2\cdot(z-z_0)}{1-0{,}2\cdot(z-z_0)}\right) & \mid z<z_0\end{cases}. \qquad (3.14)$$

The corresponding tuning curves (which are mirror images of the auditory filter shapes) are shown in Figure 3.8, together with the excitation corresponding to a 90 dB pure tone at 8 Bark.

In this approach the filter slopes depend on the output of the auditory filter itself, which seems much more realistic than the traditional approach where the slopes depend on a level that is calculated within an arbitrarily chosen bandwidth (the filter bandwidth, the width of one FFT line or one critical bandwidth).

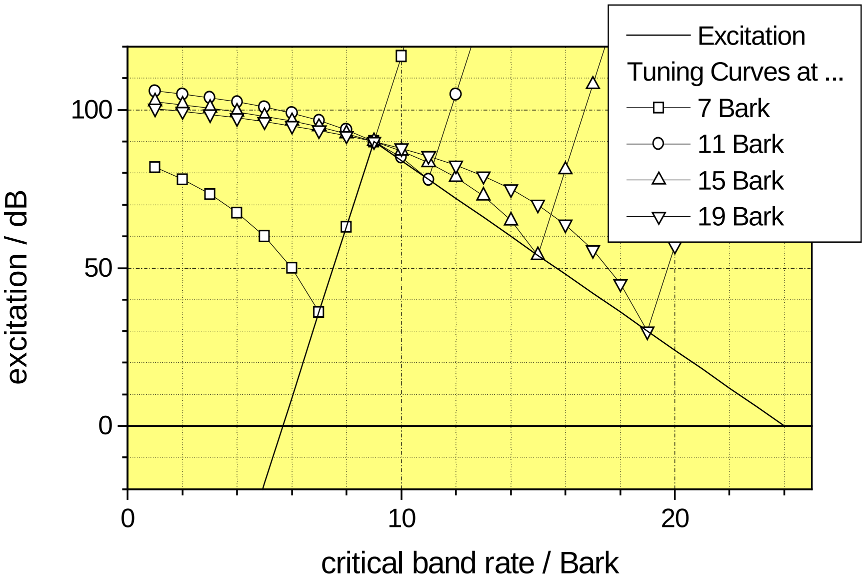


Fig. 3.8: Level-dependent tuning curves and their relation to an excitation pattern.

The computational effort of this approach is rather high because the coefficients of the band pass filters have to be re-calculated after each sample. Due to the feedback of the output level into the filter coefficients, possible problems with the stability of the filters have to be taken into account.

b) Second Approach: Switching between Filters

One rather simple approach of modelling level-dependent excitations without incorporating non-linear filters would be to run several sets of filters with different frequency responses in parallel. Each set consists of filters with the same centre frequency but fitted in with different signal levels. The filter shapes could be approximated by Eq. (3.14). Within each set the output levels of the filters are compared with the level the individual filters were designed for, and the output of the filter showing the minimum deviation will be used.

This approach is easier to realise than the first one, but the computational effort is also very high because multiple filters would have to be calculated for each frequency band.

c) Third Approach: Summation of Low Pass and Band Pass Signals

It is also possible to model level-dependent excitations by running a filter bank which solely consists of low pass filters instead of band pass filters. At low signal levels the difference between two neighbouring filters is taken, which will result in a band pass characteristic. At higher signal levels, only a fraction of the neighbouring filter output is subtracted and the result approaches the original low pass shape. From a signal processing point of view, it is more convenient to replace the weighted difference between two low pass filters by a weighted summation of a low pass and a band pass filter, which yields the same result. The task in designing a filter bank according to this concept would be to find appropriate filter shapes that yield the desired excitation patterns after a weighted summation.

This does not seem possible using a linear summation, but when allowing a summation of logarithmic values, appropriate filter shapes can be found: using a band pass filter defined by

$$A_{BP}(z) = \begin{cases} A_0 - 31\ \dfrac{dB}{Bark} \cdot (z - z_0) & \big|\, z > z_0 \\ A_0 + 24\ \dfrac{dB}{Bark} \cdot \dfrac{(z - z_0)}{1 + 0{,}2 \cdot (z - z_0)} & \big|\, z < z_0 \end{cases} \qquad (3.15)$$

and a low pass filter defined by

$$A_{LP}(z) = \begin{cases} A_0 - 31\ \dfrac{dB}{Bark} \cdot (z - z_0) & \big|\, z > z_0 \\ A_0 & \big|\, z < z_0 \end{cases} \qquad (3.16)$$

a summation of both filter outputs according to

$$A_{total} = (1 - a) \cdot A_{BP} + a \cdot A_{LP} \qquad (3.17)$$

with

$$a = \frac{A_{BP}}{120\ \mathrm{dB} - (A_{LP} - A_{BP})} \qquad (3.18)$$

yields exactly the level dependence given by Terhardt [TER79] for the excitation of one pure tone:

$$A_{total}(z)\Big|_{z \geq z_0} = A_0 - 31\ \frac{dB}{Bark} \cdot (z - z_0) \qquad (3.19)$$

$$A_{total}(z)\Big|_{z < z_0} = A_0 + \left(24\ \frac{dB}{Bark} - \ 0.2\ \cdot A_0 \cdot \frac{1}{Bark}\right) \cdot (z - z_0)\,. \qquad (3.20)$$

For more complex signals, the non-linear summation (levels instead of energies or amplitudes) of the filter outputs results in an amplification of excitations caused by more than one signal component. As a similar effect occurs in hearing, this might be tolerable or even an advantage. However, it seems that the amplification of excitations caused by multiple signal components is clearly stronger than in the corresponding psychoacoustical measurements.

d) Fourth Approach: Convolution of Complex Patterns

As already mentioned, the main shortcoming of modelling the excitation slopes by a convolution with a spreading function is the fact that the spreading changes the spectral characteristics of the filter bands without altering the temporal characteristics accordingly. Assume that the spreading would be carried out on the complex output values of a DFT. The convolution in the frequency domain corresponds to a time domain multiplication of the DFT window by the inverse Fourier transform of the spreading function. The impulse response of a pair of FIR filters corresponding to one complex spectral coefficient of the DFT is given by the product of the window function with a cosine or sine wave of the frequency corresponding to the given point of the spectrum. As a result, the impulse responses of the individual filters (if we regard the DFT as a filter bank) are changed by the spreading function and the correspondence between frequency response and impulse response is preserved by the spreading operation. The same happens also when the DFT is replaced by any other filter bank, as long as the overlap of the filters is sufficiently large and as long as the phase responses of the individual filters are identical. The latter requirement can easily be met by using linear-phase filters. Due to the exponential shape of the excitation patterns, the convolution can be carried out recursively by two first order IIR filter algorithms (this way of implementing a convolution with an exponential spreading function has already been used in [PAI92] together with an FFT-based ear model). If the level dependence of the upper excitation slope is modelled, at least the lower slope can be calculated this way.

In combination with a filter bank that provides linear phase together with a perceptually adapted frequency resolution and high computational efficiency, this concept is superior to the concepts discussed above. This also holds with respect to the computational efficiency, as well as with respect to the flexibility of the approach. A filter structure perfectly suited to be used together with this concept will be described in the following subsection.

3.5.2 FIR Filters using Recursive Algorithms

A sine wave can be recursively computed by calculating $a_{n+1}=a_n\cdot cos(\varphi) - b_n\cdot sin(\varphi)$ and $b_{n+1}=a_n\cdot sin(\varphi) + b_n\cdot cos(\varphi)$. This can be regarded as a first order IIR filter with a complex coefficient $e^{j\varphi}$ that has an infinitely long sine wave as its impulse response (i. e. it is not stable and its frequency response has a pole at the frequency of the generated sine wave). By adding a delayed version of the input signal after N samples and providing an appropriate phase shift by another complex multiplication, the impulse response can be cancelled out after N samples. Under the assumption that all calculations are carried out with absolute accuracy the resulting filter is now theoretically stable (see Section 4.3.8 for practically achieved stability). The impulse response is a time limited sine wave and the filter thus has a band pass characteristic.

When z^{-n} is a delay of n samples, Ω is the normalised frequency $2\pi\cdot f/f_{samp}$, Ω_{centre} is the normalised centre frequency $2\pi\cdot f_{centre}/f_{samp}$, and z equals $e^{j\Omega}$, the filter can be described in the z-domain by

$$y = x + e^{j\varphi_2} x \cdot z^{-N} + e^{j\varphi_1} y \cdot z^{-1} \tag{3.21}$$

and the frequency response of the filter is given by

$$A(z) = \frac{y}{x} = \frac{1 + e^{j\varphi_2} \cdot z^{-N}}{1 - e^{j\varphi_1} \cdot z^{-1}} \tag{3.22}$$

The amplitude response of this filter is

$$A(\Omega)=\frac{\sin\left(\frac{1}{2}(\varphi_2-N\Omega)\right)}{\sin\left(\frac{1}{2}(\varphi_1-\Omega)\right)} \tag{3.23}$$

Stability is achieved when the pole at $\Omega=\varphi_1$ is cancelled out, i. e. $\varphi_2=N\cdot\varphi_1$. Furthermore, it is obvious that φ_1 defines the centre frequency of the band pass. For this reason, φ_1. is renamed to Ω_{centre}.

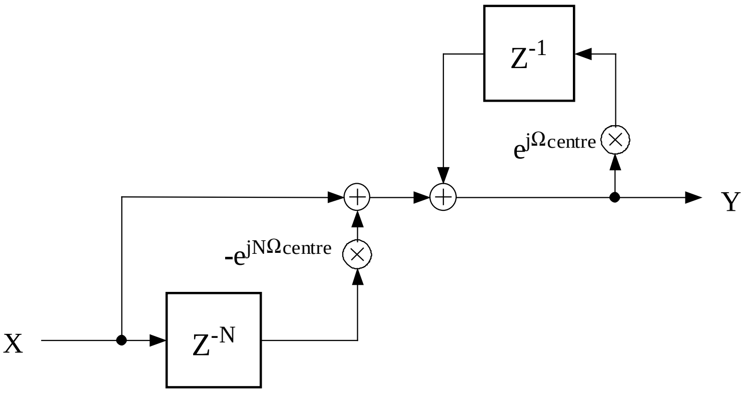


Fig. 3.9: Structure of a recursive filter realising a complex linear phase band pass.

The amplitude response for this case can be expressed as

$$A(\Omega)=\frac{\sin\left(\frac{N}{2}(\Omega_{centre}-\Omega)\right)}{\sin\left(\frac{1}{2}(\Omega_{centre}-\Omega)\right)}=N\cdot\frac{si\left(\frac{N}{2}(\Omega_{centre}-\Omega)\right)}{si\left(\frac{1}{2}(\Omega_{centre}-\Omega)\right)} \tag{3.24}$$

and the phase response is

$$\varphi(\Omega)=\frac{N-1}{2}\cdot(\Omega_{centre}-\Omega). \tag{3.25}$$

The filter has no poles but *N-1* zeros. Hence, it behaves like an N^{th} order FIR filter, even though its structure belongs to an IIR filter and its computational cost is correspondingly low. The filter even provides linear phase (Eq. 3.25). Its amplitude response is shown in Figure 3.10. The filter structure given in Fig. 3.9 can also be looked upon in a traditional way: the first element represents a N^{th} order FIR filter where all coefficients except the first and the last one are zero. Its pole-zero configuration consists of N zeros which are regularly spaced on the unit circle. The second element represents a recursive filter with one pole (this is possible because the coefficients in this example are complex) on the unit circle. With the given parameters, the pole is perfectly compensated by one of the zeros of the first filter, and the resulting filter thus yields the band pass characteristic shown in Figure 3.10. When the complex multiplications are separated into real and imaginary parts, the filter can be realised by the structure given in Figure 3.11. Such a filter structure has been proposed in [LIU93] for the computation of short-time FT in applications where the use of FFT algorithms is not possible due to implementational considerations.

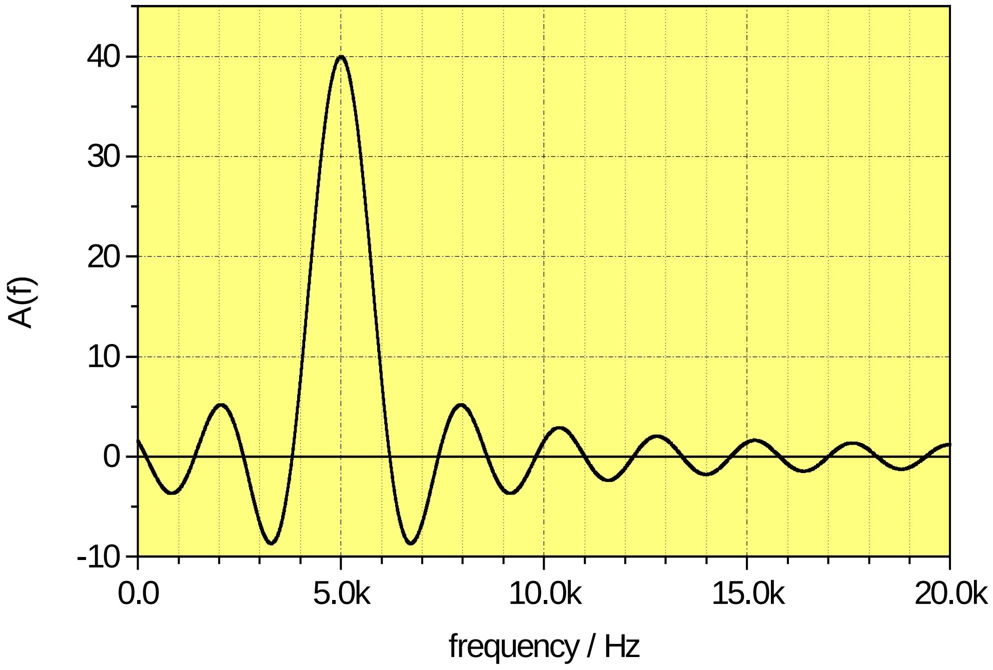


Fig. 3.10: Amplitude response of one filter as defined by Eq. (3.21) with a centre frequency of 5 kHz and a bandwidth of 1200 Hz.

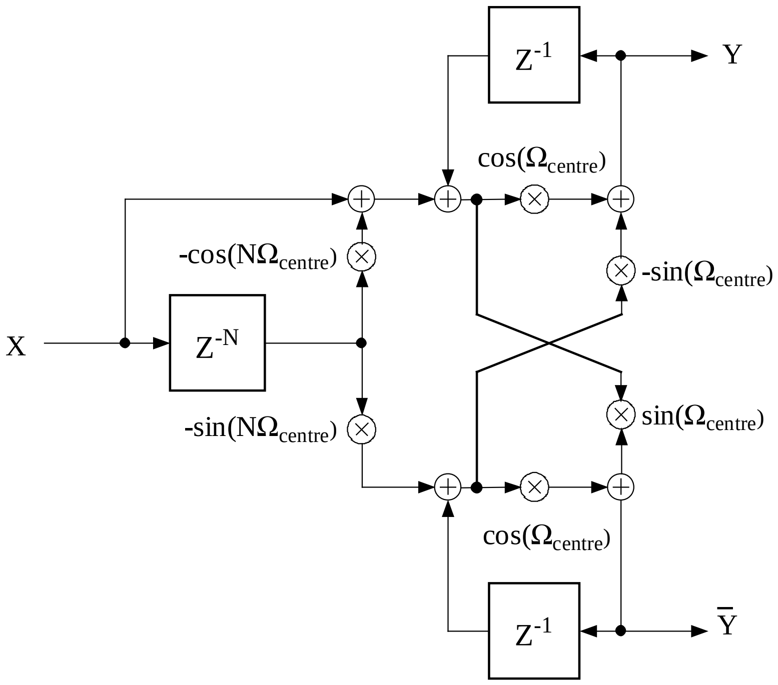


Fig. 3.11: Structure of a recursive filter realising a pair of linear phase band pass filters with 90° phase shift.

The sampling theorem for band pass signals allows to sample the output of a band pass filter at a sampling rate down to twice its bandwidth[4], which is typically much lower than the sampling rate of the input signals. It would be convenient if the output of the band pass were only computed each time the output signal is sampled. With a purely transversal filter structure this is always possible. With a recursive filter this is, in general, not possible, because the preceding output values are needed in order to compute a new output value. However, it would be possible if only each M^{th} preceding output value was used (where $M = f_{samp,input} / f_{samp,output}$). This can be achieved by recursively replacing the term $y \cdot z^{-1}$ in Eq. (3.21) with the result of Eq. (3.21) delayed by z^{-1}

$$y = x + e^{jN\Omega_{centre}} x \cdot z^{-N} + e^{j\Omega_{centre}} y \cdot z^{-1} \tag{3.26}$$

$$y = x + e^{jN\Omega_{centre}} x \cdot z^{-N} + e^{j\Omega_{centre}} \cdot \left(x + e^{jN\Omega_{centre}} x \cdot z^{-N-1} + e^{j\Omega_{centre}} y \cdot z^{-2} \right) \tag{3.27}$$

...

$$y = x \cdot \sum_{m=0}^{M-1} e^{jm\Omega_{centre}} \cdot \left(1 - e^{jN\Omega_{centre}} \cdot z^{-N} \right) \cdot z^{-m} + y \cdot e^{jM\Omega_{centre}} \cdot z^{-M} \tag{3.28}$$

The filter defined by Eq. (3.28) only depends on the input values and one output value delayed by M samples. Equation (3.28) also includes some symmetries which can be used for an additional reduction of the computational effort:

$$\begin{aligned} y = {} & x \cdot \sum_{m=0}^{M-1} e^{jm\Omega_{centre}} \cdot z^{-m} \\ & - x \cdot \sum_{m=0}^{M-1} e^{jm\Omega_{centre}} \cdot e^{jN\Omega_{centre}} \cdot z^{-N-m} + y \cdot e^{jM\Omega_{centre}} \cdot z^{-M} \end{aligned} \tag{3.29}$$

$$\begin{aligned} y = {} & x \cdot \sum_{m=0}^{M-1} e^{jm\Omega_{centre}} \cdot z^{-m} \\ & - x \cdot \sum_{m=0}^{M-1} e^{j(N+M-1-m)\Omega_{centre}} \cdot z^{-N-M+1+m} + y \cdot e^{jM\Omega_{centre}} \cdot z^{-M} \end{aligned} \tag{3.30}$$

---

[4] In case the cut-off frequencies of the band pass filter are integer multiples of its bandwidth.

$$y = e^{j\frac{N+M-1}{2}\Omega_{centre}} \cdot x \cdot \sum_{m=0}^{M-1}\left[e^{j\left(m-\frac{N+M-1}{2}\right)\Omega_{centre}} \cdot z^{-m} - e^{j\left(\frac{N+M-1}{2}-m\right)\Omega_{centre}} \cdot z^{-N-M+1+m}\right] + y \cdot e^{jM\Omega_{centre}} \cdot z^{-M} . \qquad (3.31)$$

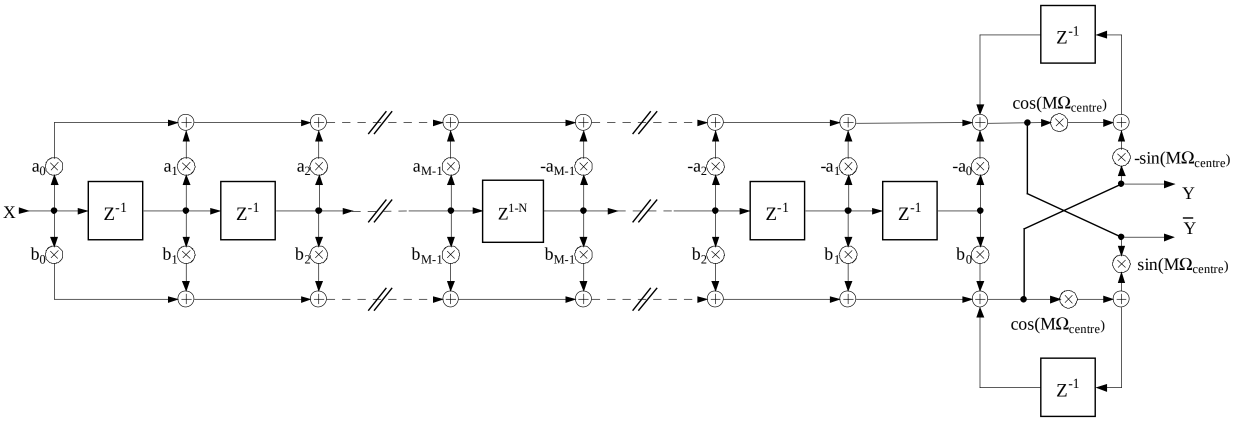
Omitting the first factor only results in a shift of the phase response. Hence, the filter defined by

$$y = x \cdot \sum_{m=0}^{M-1}\left[e^{j\left(m-\frac{N+M-1}{2}\right)\Omega_{centre}} \cdot z^{-m} - e^{j\left(\frac{N+M-1}{2}-m\right)\Omega_{centre}} \cdot z^{-N-M+1+m}\right] + y \cdot e^{jM\Omega_{centre}} \cdot z^{-M} \qquad (3.32)$$

has the same amplitude response as the original filter. The corresponding filter structure is shown in Figure 3.12. The phase response of the filter (Eq. 3.25) becomes

$$\varphi(\Omega) = -\frac{M}{2} \cdot \Omega_{centre} - \frac{N-1}{2} \cdot \Omega . \qquad (3.33)$$

For the filter defined by Eq. (3.32) the complex multiplication by *y* requires four multiplications and two summations. Each complex multiplication by *x* only requires two multiplications and two summations because the x values have no imaginary part.



$$a_m = \cos\left[\left(m - \frac{N+M-1}{2}\right) \cdot \Omega_{centre}\right] \qquad b_m = \sin\left[\left(m - \frac{N+M-1}{2}\right) \cdot \Omega_{centre}\right]$$

Fig. 3.12: Filter structure for efficient computation of subsampled band pass values.

Due to the symmetry of the coefficients, half of these multiplications can be omitted when calculating the sum and difference of each pair of delayed x-values first. Altogether, the filter requires $2 \cdot M+4$ multiplications and $4 \cdot M+2$ summations per output value or $2+4/M$ multiplications and $4+2/M$ summations per input value. The original filter structure required 6 multiplications and 6 summations per input value or $6 \cdot M$ multiplications and $6 \cdot M$ summations per output value.

The band pass characteristic of the filter defined by Eq. (3.32) is not yet very sharp. The amplitudes of the secondary lobes only descend with $1/(\Omega - \Omega_{centre})$. By running $K+1$ filters with slightly staggered centre frequencies in parallel, the decay of the secondary lobes can be enhanced. The highest attenuation of the secondary lobes is achieved with

$$A(\Omega) = \sum_{k=0}^{K} w_k \cdot A_k(\Omega) \tag{3.34}$$

where

$$\Omega_{centre,k} = \Omega_{centre} + \left(k - \frac{K}{2}\right) \cdot \frac{2\pi}{N} \tag{3.35}$$

and

$$w_k = \frac{2\pi}{N} \cdot 2^{-K} \cdot \binom{K}{k} \tag{3.36}$$

The secondary lobes of this filter descend with $1/(\Omega - \Omega_{centre})^{K+1}$. The impulse response of the filter has the shape

$$a_K(n) = \sin^K\left(\frac{\pi}{N} n\right) \cdot \cos\left(\frac{\Omega_{centre}}{f_{samp}} \cdot n\right) \qquad \Bigg| \qquad 0 \le n < N \tag{3.37}$$

for the real part and

$$a_K(n) = \sin^K\left(\frac{\pi}{N} n\right) \cdot \sin\left(\frac{\Omega_{centre}}{f_{samp}} \cdot n\right) \qquad \Bigg| \qquad 0 \le n < N \tag{3.38}$$

for the imaginary part. With $K=2$ this filter characteristic corresponds to the characteristic of one frequency line of a DFT using a raised cosine window. For higher values of K the amplitude response approximates a Gaussian shape. However, the described algorithm becomes inefficient when K gets too large. As long as the filter bands are not too wide, $K=2$ already yields sufficiently steep filter slopes at a comparably low computational effort (Figure 3.13).

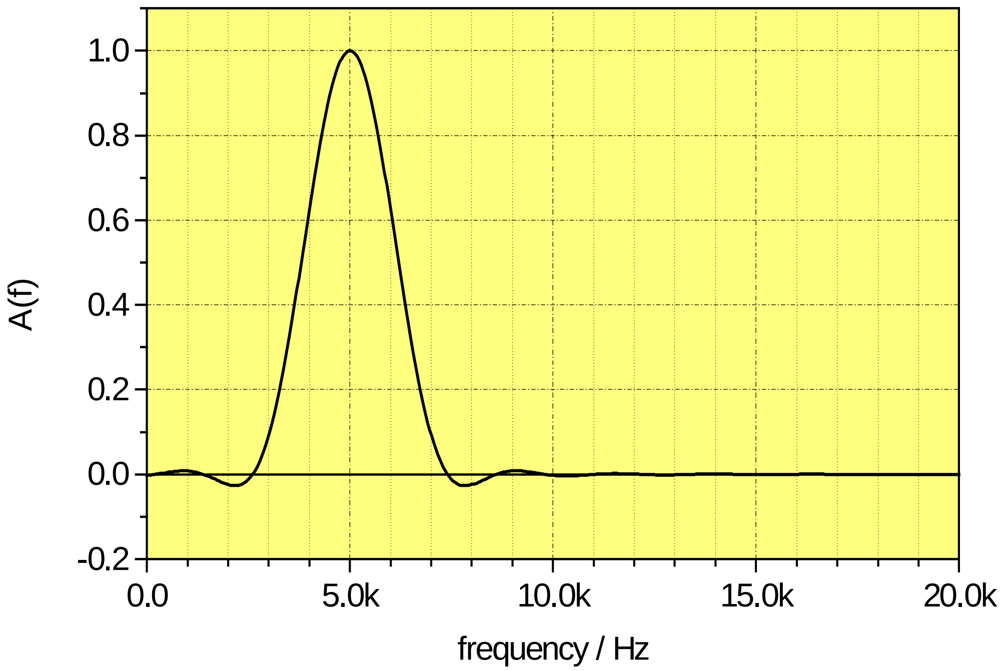


Fig. 3.13: Amplitude response of a second order FDC-filter with a centre frequency of 5 kHz and a bandwidth of 1200 Hz.

The weighted summation of $K+1$ staggered band pass filters running in parallel can be considered as a frequency domain convolution of the transfer function of the band pass with the Fourier transform of a time window. Consequently, we will call such a filter a K^{th} order *FDC (frequency domain convolution) filter*. In general, the concept of implementing FIR filters in such a parallel structure is known as frequency-sampling structure [OPP75].

Due to their high computational efficiency and the linear phase, such filters are perfectly fitted to be used together with the concept of convolving complex patterns, which was described in the previous section. As this again is a frequency domain convolution of transfer functions, we will call the complete filter bank an FDC filter bank.

The main shortcoming of the proposed filter structure is the fact that the filters are not necessarily stable when using floating-point arithmetic. As the filters work at the border of stability, the rounding errors occurring with floating-point arithmetic might cause the filters to become unstable. In practice, the filters produce a kind of band pass noise after switching off the input signals. Fortunately, this output noise remains below the threshold of audibility for the duration of the usual test signals (20 to 40 seconds). However, it might become a problem when using this filter bank on continuous signals. Therefore, the potential instability of the filter will be discussed in more detail later on (Section 4).

3.6 Computational Efficiency

For the example of an auditory filter bank with a spectral resolution of 0.5 critical bands and an assumed subsampling of the band pass outputs by a factor of eight, the computational efficiency of the FDC-filter bank was compared with other filter bank algorithms.

Naturally, the computational effort depends on the implementation. Table 3.1 shows the assumed number of multiplications and summations for the basic operations (where the estimation for the FFT is based on the simple algorithm given in [ENG85]). The resulting computational effort for the computation of each complex output value is given in Table 3.2.

Type of operation	**Number of multiplications (M) and summations (A)**
complex FFT, N tabs:	(4M+4A) ·ld(N)
first order IIR low pass:	1M+1A
first order IIR all pass:	2M+2A
second order IIR filter:	4M+4A
N^{th} order FIR filter:	N·(1M+1A)
complex multiplication (phase shift):	4M+2A

Tab. 3.1:
Number of multiplications and summations per input sample for the basic operations needed in filter bank algorithms.

The number of filter bands is 48 for all filter banks except the FAMlet filter bank, which requires a power of two for an efficient computation (because of the incorporated FFT). Under the assumption that 48 filter bands are sufficient, the relative computational effort of the FAMlet filter bank is therefore multiplied by a factor of 64/48 when compared to other filter banks. The filter bank algorithms marked with an asterisk (*) cannot make advantage of the subsampling of the output values by factor 8 as assumed above. Hence, their performance relative to the other filter banks will be increased when no subsampling is allowed and decreased when a higher subsampling factor is allowed.

In order to have a fair comparison, the FDC filter bank considered here does not model level-dependent excitations (because none of the other filter banks does). It therefore consist of second order FDC filters for the band pass decomposition and two first order IIR filters for the modelling of the auditory filter slopes. All estimations are made for filters with a complex output signal (i. e. filter pairs with 90° relative phase shift). The advantages of having complex filter outputs will be explained in the description of the new measurement method (Section 4).

Filter bank type, number of filter bands and description of the required operations	**Number of multiplications (M) and summations (A)**	
	M	**A**
Sixth order Gammatone filter bank, 48 filter bands. For each filter band and each sample: one complex multiplication for the input signal and six first order low pass filters for both real part and imaginary part.	16	14
Second order FDC-filter bank, 48 filter bands. For each filter band : three FDC-filters. For each output value: four first order IIR filters.	8	13.25
FAMlet, 64 filter bands (the next power of two which is larger than or equal 48) For each input value: 64 first order all pass filters. For 64 output values: one complex FFT with 128 tabs each eighth sample.	9	9
FIR straightforward, 512 tabs, 48 filter bands. For each filter band: two 512th order FIR filters each eighth sample.	128	128
FIR [GOE68], 512 tabs, 48 filter bands. For each filter band: one 512th order FIR filter each sample.	512	512 (*)
Twentieth order IIR filter, 48 filter bands. For each filter band and each sample: two cascades of ten second order IIR filters.	80	80 (*)
Tenth order IIR filter, 48 filter bands. For each filter band and each sample: two cascades of five second order IIR filters.	40	40 (*)
FIR + FFC, 8192/512 tabs For 8192-511 samples: one complex FFT with 8192 tabs.	55.5	55.5 (*)

Tab. 3.2:
Multiplications and summations per filter band and input sample.

For the given parameters, the computational efficiency of Gammatone filter bank, FAMlet filter bank and FDC filter bank is in the same order of magnitude. FIR filters in combination with an FFC algorithm and tenth order IIR filters require an approximately four times higher computation effort. Higher order IIR filter banks and FIR filters in a straightforward implementation are less efficient in this context.

The computational efficiency of all these filter banks can be somewhat improved by using a tree-structured low pass cascade as a pre-filter in order to achieve lower sampling rates at the inputs of the band pass filters. This concept is used together with an FFC algorithm in the measurement scheme described in [SPO96].

- **Memory Consumption**

Besides the number of multiplications and summations, another aspect of computational efficiency is the amount of memory required for storing the filter coefficients and the buffering of input and output values.

With respect to the storing of filter coefficients, filter banks based on FIR filters have the largest memory consumption and the Gammatone filter bank has clearly the lowest. For *FAMlet* filter banks and the *FDC* filter bank, the number of filter coefficients is slightly larger than for the *Gammatone* filter bank but the difference is not really relevant. With respect to the buffering of in- and output values, the fast-forward-convolution has an extremely high memory consumption, whereas for the *Gammatone* filter bank, the *FAMlet* filter bank and the *FDC* filter bank the memory consumption again is sufficiently low to be neglected.

3.7 Flexibility

Among the filter bank concepts with the highest computational efficiency, the FAMlet concept is restricted to certain frequency-to-pitch mapping functions, and Gammatone-filter banks are restricted to a certain filter shape. Even though for the FDC-filter bank the choice of the filter shape is not completely free either, at least the three most important parameters of auditory filters, frequency-to-pitch mapping, bandwidth, and slope rate can be chosen independently. Moreover, it is the only concept that integrates the modelling of level-dependent excitation slopes into the filter bank.

3.8 Conclusions

As it combines the highest flexibility with a low computational complexity and provides an elegant method of modelling level-dependent auditory filter slopes, the *FDC*-filter bank is chosen as the default time-frequency decomposition for the new measurement method.

4. DIX - A New Perceptual Measurement Method

This section describes the development and first implementation of a new perceptual measurement method. The original idea behind the new method was to compute a multidimensional representation of the audible differences between processed and original signal, and map these parameters to a global quality index which quantifies the disturbance of the perceived errors. For this reason, the method was called *Disturbance Index* (*DIX*).

The dimensions of audible differences can be defined similar to the basic psychoacoustical parameters proposed by Zwicker [ZWI90], loudness, roughness, fluctuation strength, and sharpness. The error measure corresponding to loudness can be defined as the partial loudness of the signal components that do not belong to the original signal. For sharpness, roughness and fluctuation strength, the equivalent error measures are not that obvious. Neither a sharpness difference nor the sharpness of the difference signal would really make sense, and something like a “partial sharpness” has never been defined, and would have no relation to psychoacoustical experiments either. The same holds for error measures corresponding to roughness or fluctuation strength. What can be taken as a minimum definition is that sharpness is related to the spectral envelope, and roughness and fluctuation strength are related to the temporal envelope of a signal. Hence, in addition to the partial loudness of the distortion, the measurement method should provide a measure for changes in the spectral envelope, and a measure for changes in the temporal envelopes (Figure 4.1).

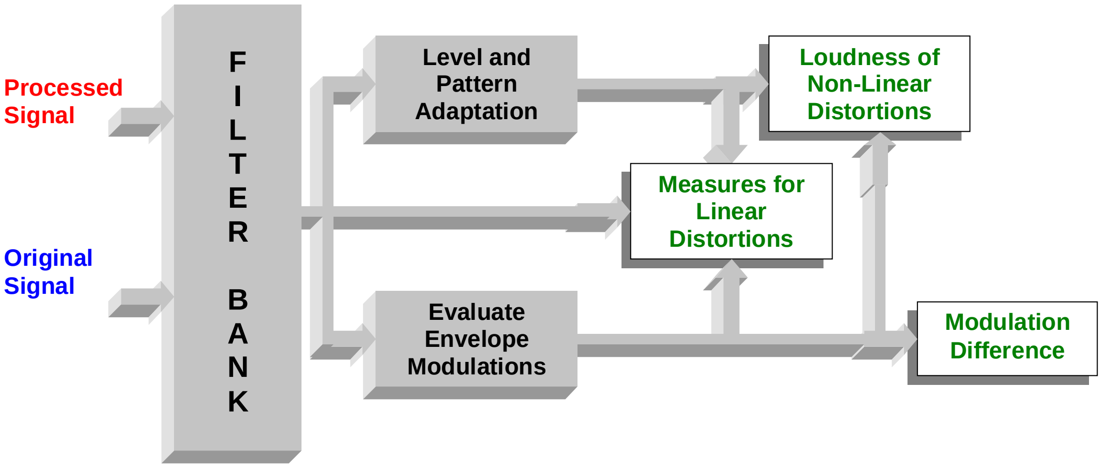

Processed Signal
Original Signal
FILTER BANK
Level and Pattern Adaptation
Evaluate Envelope Modulations
Loudness of Non-Linear Distortions
Measures for Linear Distortions
Modulation Difference

Fig. 4.1: Computation of quality indicators in DIX.

A multidimensional quality measure is difficult to verify against listening test data which only consist of ratings for the overall quality of the processed signal. Mapping functions which combine numerous model parameters into one overall quality indicator have a high number of degrees of freedom, and are hence not reliable unless a huge set of test data is used to establish the mapping. If the test data set has to be very large, an experimental optimisation of the model will become impractical because of the large computational effort. For this reason, earlier implementations of the model employed only one-dimensional mappings. As a consequence, only those model output variables could be verified that explain a major part of the distortions. Using the accessible test data sets, only the partial loudness of additive distortions could be validated, but not the measures for linear distortions and modulation differences. As the possibility to validate an error measure is a requirement for optimisation, the latter measures could not be optimised either. Later on, the increased computational efficiency of the model together with higher computational power made it possible to apply multidimensional mappings. These mapping functions will be described in Section 6.

4.1 Objectives and Requirements

Even though commonly used psychoacoustical models can serve as a starting point for the development of a perceptual measurement method, it has turned out that in order to achieve satisfactory results, many parts of a measurement method have to be optimised by "trial and error" (like, for example, in [BEE92] and [BEE93]). Most psychoacoustical models are derived by listening experiments using comparably simple test signals, and the models often fail when applied to very complex signals like, for example, the audio excerpts used in codec comparison tests. Therefore, the general structure of the new perceptual model should correspond to established psychoacoustical models, but not the parameter settings. In order to allow for an extensive experimental optimisation, the new model should provide a high degree of flexibility. On the other hand, the need for experimental optimisation also requires to keep the number of degrees of freedom as low as possible, in order to reduce the danger of adapting the model to the test data set without actually improving it.

In existing methods, the optimum resolution in the time and frequency domain could not be investigated without side-effects, because, due to the application of FFT-based algorithms, the product of spectral and temporal resolution was worse (or at least not better) than the spectral and temporal resolution of the human auditory system. For this reason, the optimum resolution found with such models is probably not the optimum spectral and optimum temporal resolution, but only the best possible trade-off between both. In order to optimise temporal and spectral resolution independently, the product of spectral and temporal resolution of the new model should be clearly better than in the auditory system (that is, it should be as good as possible).

As a result of experimental optimisation, some parts of the auditory models incorporated in existing methods are not in line with known psychoacoustical data, and in some methods parts of the psychoacoustical model are even left out [BEE94]. If possible, the new model should avoid such contradictions to established psychoacoustical models, even though the performance on real-world audio data should have the preference before the performance on simple test material as used in psychoacoustical experiments.

4.2 Overview

At the input of the ear model, both the processed and the original signal are adjusted to the assumed playback level and sent through a high pass filter in order to remove DC and subsonic components of the signals. The signals are then decomposed into band pass signals by linear phase filters which are equally distributed over a perceptual pitch scale. A frequency dependent weighting function is applied to the band pass signals, which models the spectral characteristics of the outer and middle ear. The level-dependent spectral resolution of the auditory filters is modelled by a frequency domain convolution of the outputs with a level-dependent spreading function (Figure 4.2).

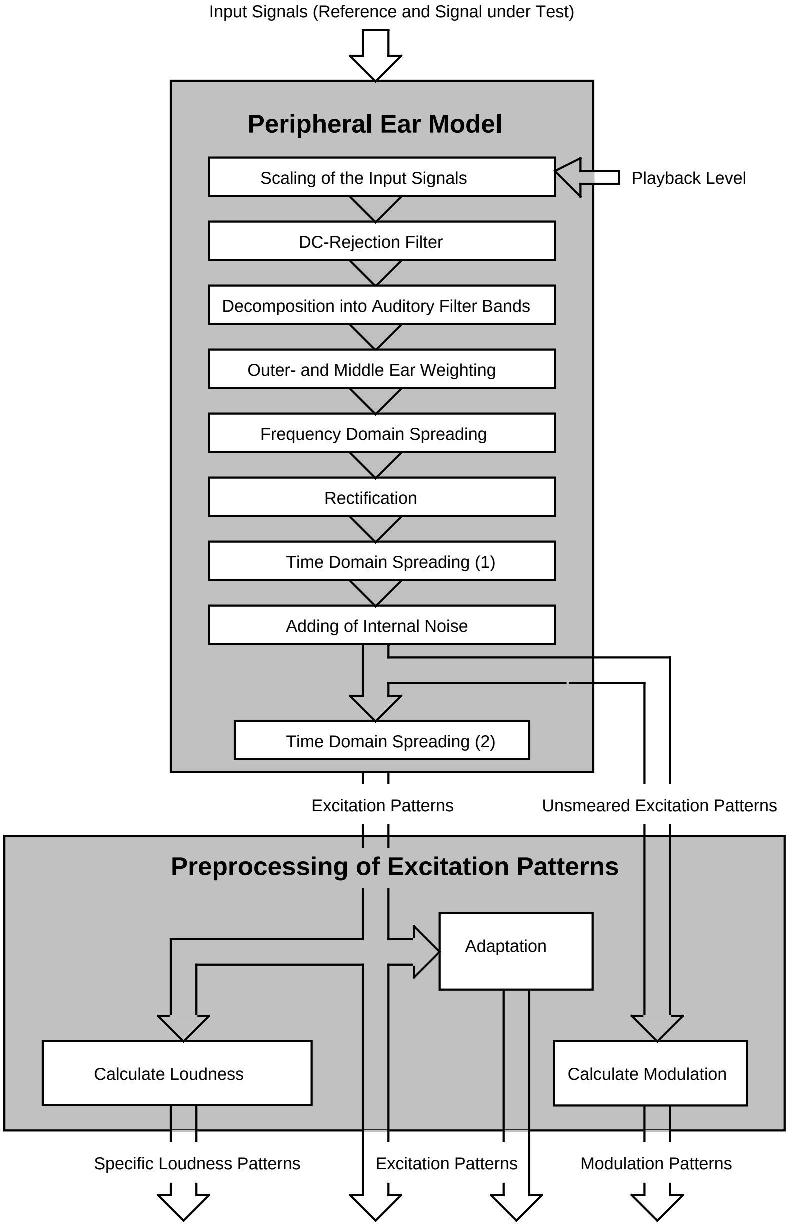

Input Signals (Reference and Signal under Test)
Peripheral Ear Model
Scaling of the Input Signals
Playback Level
DC-Rejection Filter
Decomposition into Auditory Filter Bands
Outer- and Middle Ear Weighting
Frequency Domain Spreading
Rectification
Time Domain Spreading (1)
Adding of Internal Noise
Time Domain Spreading (2)
Excitation Patterns
Unsmeared Excitation Patterns
Preprocessing of Excitation Patterns
Adaptation
Calculate Loudness
Calculate Modulation
Specific Loudness Patterns
Excitation Patterns
Modulation Patterns

Fig. 4.2: Structure of the perceptual model of DIX.

The envelopes of the signals are calculated using the Hilbert-transform of the band pass signals (”rectification”) and a time domain convolution with a window function is applied in order to model backward masking. Then, a frequency dependent offset is added, which takes internal noise in the auditory system into account and models the threshold in quiet. Finally, forward masking is modelled by another time domain convolution using an exponential spreading function (Figure 4.2).

The thus obtained excitation patterns are used to compute specific loudness patterns, and the patterns obtained before the final time domain spreading (”*unsmeared excitation patterns*”) are used to calculate *modulation patterns*. These, together with the excitation patterns themselves, are the basis for the computation of the model values. In order to separate the influence of the steady-state frequency response of the device under test from other distortions, the excitation patterns of processed and original signal are also spectrally adapted to each other (”adaptation”). Modulation patterns and specific loudness patterns are calculated from both the adapted and the non-adapted excitation patterns. However, the adaptation has no significant influence on the latter and is therefore not included in Figure 4.2 for these patterns.

4.3 Peripheral Ear Model

As the performance of the filter bank is one of the most important parts of the perceptual model, special care has been taken to find an algorithm which is well adapted to the human auditory system and provides both a high degree of flexibility, and high computational efficiency. Concerning these requirements, the FDC filter bank introduced in Section 3.5 was considered as the best possible solution.

4.3.1 Structure of the Filter Bank

The FDC-filter bank as depicted in Figure 4.3 consists of a variable number (N) of parallel recursive FIR band pass filters (RFIR) for each auditory filter band (Figure 4.3a and b). Each filter element is implemeted using the the structure given in Figure 3.12. In order to achieve a reasonably good stop-band attenuation, the outputs of the N filter elements belonging to each filter band are weighted according to Eq. (3.36) and added up (Figure 4.3c). Another weighted summation is carried out among the different auditory filter bands to achieve exponential filter slopes in the roll-off of the filters (Figure 4.3d). This is formally equivalent to the convolution with a spreading function that is carried out in FFT-based models, but yields a very different result, as this spreading operation is carried out before any non-linear operations (like rectification) and therefore preserves the relation between spectral and temporal characteristics (impulse response) of the filters. Hence, the output signals of the filter bank after this spreading operation are identical to the output signals of hypothetical filters directly realising the exponential slopes of auditory filters. However, this operation only yields the desired result when the phases of the individual filters are equal. As the FDC filters are of linear phase, equal phases are easily achieved by delaying each filter output by half the difference between the length of the impulse response of the current filter and the length of the impulse response of the filter with the lowest bandwidth.

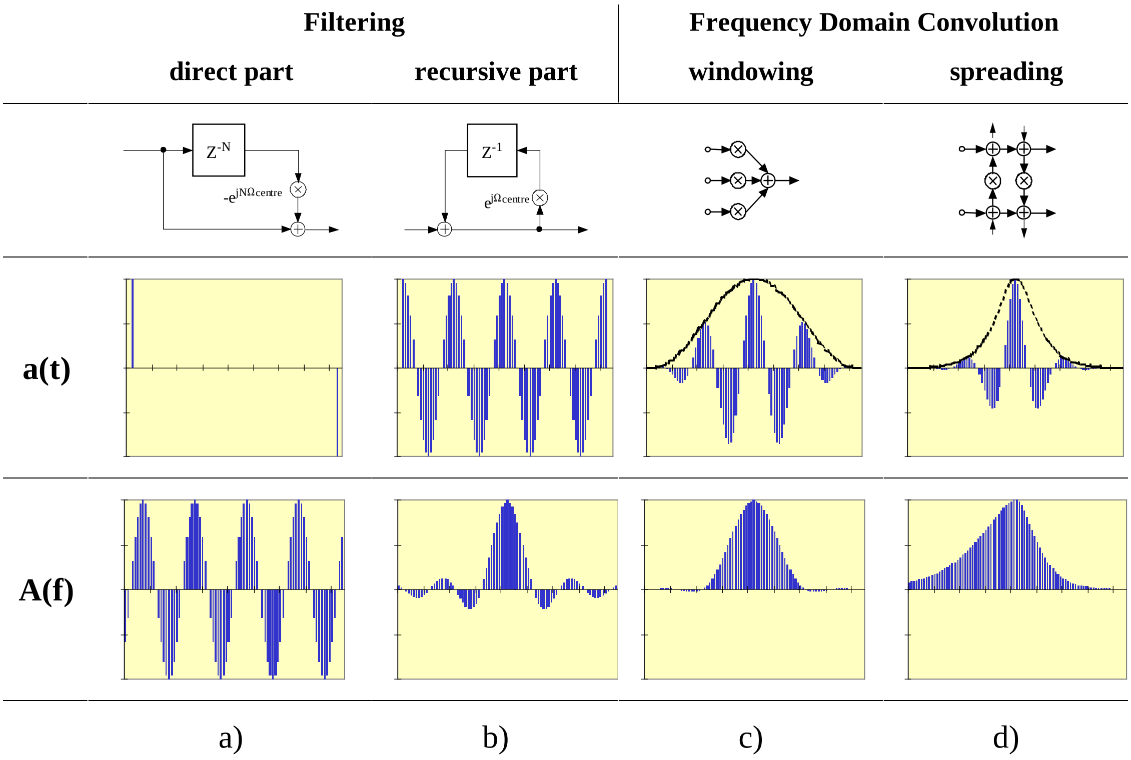


Fig. 4.3: Processing steps within the FDC filter bank. The middle row shows the pulse response at the different stages, and the lower row shows the corresponding frequency responses.

The slope rates of the filters can be defined independently for the ascending and the descending slope. It can be chosen level and frequency dependent as in the approximation by Terhardt [TER79], but, optionally, also constant slope rates can be modelled. Minimum and maximum slope rate as well as the amount of level and frequency dependence can be chosen arbitrarily:

$$S\left(f_{centre}, L\right) = \max\left(S_{\min}, S_{\max} + \frac{a}{f_{centre}\,/\,Hz} - b \cdot L\,/\,dB\right)\frac{dB}{Bark}, \qquad (4.1)$$

where L is the local energy level in the current filter band, S_{min} is the minimum slope rate, S_{max} is the maximum slope rate at intermediate and high centre frequencies, a determines the frequency dependence and b determines the level dependence.

In addition, the worst case curve proposed in [BRA87] (see Section 2.2.3) was also implemented, and could optionally be chosen. By default, a level-dependent upper slope (when related to excitation patterns) was chosen, using the slope rates and level dependences that were originally proposed by Terhardt [TER79]. As it is not very likely that the auditory filter slopes change instantaneously with changing sound pressure level, the slope rate of the level-dependent slope is smoothed in time, using a first order low pass filter.

The centre frequencies of the filter bands are distributed according to the approximation of the critical band scale given by Schroeder [SCH79]:

$$f\,/\,kHz \approx \sinh\left(\frac{z\,/\,Bark}{7}\right). \qquad (2.23)$$

Nevertheless, most other known approaches for auditory frequency scales are also implemented, including the more precise approximation of the critical band scale proposed in [ZWI80], the ERB-scale as modelled in [STU93], and the SPINC-scale [TER92] (see Section 2.2.3).

The overlap between adjacent filters can be chosen arbitrarily as well. In the beginning of this work, it was set to a value which would allow perfect reconstruction in case the filters were equally spaced. This value can also be interpreted as the number of filters required in order to preserve all information that would be included in a continuous frequency representation (this is explained in more detail in Section 4.3.3). After some experiments (see Section 4.7), this value was reduced such that adjacent filters overlap at their -6 dB points. For the chosen filter shapes (second order FDC filters), the latter approach yields half the number of filter bands of the original approach.

The number of filter bands, which, together with the overlap, defines the filter bandwidths, was subject to experimental optimisation and was varied between less 20 and 250 bands. The upper bound was given by the memory restrictions of the first (MS-DOS based) implementation, but turned out to be sufficiently high above the optimum number of filter bands.

4.3.2 Pre-Filtering and Scaling

a) Setting of Playback Level

For the correct modelling of the threshold in quiet and the level-dependent auditory filter slopes, the model must be adapted to the playback level of the test signal. From the assumed sound pressure level L_{MAX} of a full scale sine tone, a scaling factor *fac* for the input signal is calculated by

$$fac = \frac{10^{L_{\max}/20}}{A_{full\ scale}}, \qquad (4.2)$$

in order to enforce that an amplitude of one corresponds to a sound pressure level of zero decibels. In the test data used, the values of L_{MAX} are in the range of 85 to 100 dB SPL. Therefore, in case the exact playback level is not known, L_{MAX} is set to 92 dB SPL, which also is close to the dynamic range of the 16 bit PCM format of the test data.

b) DC-Rejection Filter

As the filter bank turned out to be too sensitive to subsonics in some of the test signals, a DC-rejection filter is applied before the input signals are fed into the filter bank. A fourth order Butterworth high pass filter with a cut-off frequency of 20 Hz is used. The filter is realised as a cascade of two second order IIR filters

4.3.3 Sampling in the Time and Frequency Domain

The number of filter bands and the possible subsampling of the band pass signals was estimated by assuming a continuous time-frequency representation and estimating the required sampling rates in both dimensions.

a) Sampling in the Time Domain

The spectrum of the temporal envelope of a band pass signal has a low pass characteristic, where the cut-off frequency is determined by the width of the band pass. Therefore, the envelope of an ideal band pass signal can be perfectly reconstructed from its sampled representation as long as the sampling rate is higher than twice its bandwidth. When the temporal envelopes are calculated using the Hilbert transform of the band pass signal (e. g. by computing another band pass with a phase response shifted by 90°), the band pass signals can directly be sampled with the sampling rate required by their temporal envelopes and the band pass signals do not have to be reconstructed again[5].

b) Sampling in the Frequency Domain

One advantage of directly modelling the shapes and distribution of auditory filters by a filter bank is given by the property that the output values of the individual filters can be looked at as a sampled representation of a continuous excitation pattern. Just like in the case of sampling a continuous time signal, when sampling a continuous frequency representation, the pattern can be perfectly reconstructed from a limited number of samples (which in this case means: filter bands). The only requirement for this are identical filter shapes and a limited length of the Fourier transform of the filter shape. The first requirement is obviously not fulfilled when looking at the filter shapes as a function of frequency. However, when looking at the filter shapes as a function of the pitch units on the auditory frequency scale, the filter shapes should ideally be identical.

In the filter bank described above, this is at least approximately fulfilled. As the final filter shapes are derived from a convolution of the original filter shapes by the spreading functions that define the upper and lower slope rate of the auditory filters, their Fourier transform is given by the product of the Fourier transforms of the original filter shapes and the Fourier transforms of the spreading functions. As the latter are theoretically infinitely long (the Fourier transform of an exponential is given by a rational function), the limitation of the length of the Fourier transform of the final filter shapes must be given by the Fourier transform of the original filter shapes. Even though the filter shapes are distorted by the frequency warping of the auditory frequency scale, they approximately have the same shape on the auditory frequency scale as on a linear frequency scale, because the filter bands are sufficiently narrow to approximate the frequency warping by a linear mapping for the range around the centre frequency of each filter. As the filters have a finite impulse response of the length $T(f_{centre})$ and the frequency warping is assumed to be approximately linear within the pass band of each filter, the Fourier transform of a filter shape on the auditory frequency scale is given by the (time-reversed) impulse response of the

---

[5] This strategy of subsampling band pass filtered signals without having to reconstruct them for further processing corresponds to the usual procedure when using a DFT.

filter, scaled by a constant factor. This scaling factor is given by the derivation of the frequency warping function

$$m = \frac{1}{\frac{dz}{df}} = \frac{df}{dz}. \tag{4.3}$$

The frequency domain sampling, which equals the maximum distance between adjacent filters in the critical band domain, Δz_{max}, is then given by

$$\Delta z_{\max} = \frac{1}{T(f_{centre}) \cdot m(f_{centre})} = \frac{\frac{dz}{df}(f_{centre})}{T(f_{centre})}, \tag{4.4}$$

which should not depend on the centre frequencies of the filters. When the filters are equally distributed over the auditory frequency scale and have a constant overlap, the filter bandwidths are given by

$$bw = f\left[z(f_{centre}) + \frac{bw'}{2}\right] - f\left[z(f_{centre}) - \frac{bw'}{2}\right] \approx \frac{bw'}{\frac{dz}{df}(f_{centre})}, \tag{4.5}$$

where *bw'* is the filter bandwidth on the auditory frequency scale. The length of the impulse response *T* of a band pass filter is proportional to the inverse of the bandwidth:

$$T(bw) = \frac{c}{bw}, \tag{4.6}$$

Inserting this into Eq. (4.4), the maximum distance between adjacent filters becomes

$$\Delta z_{\max} = \Delta z \cdot \frac{\frac{dz}{df}(f_{centre})}{c \cdot \frac{dz}{df}(f_{centre})} = \frac{bw'}{c}, \tag{4.7}$$

and the maximum required number of filter bands Z_{max} is given by

$$Z_{\max} = \frac{z(20kHz) - z(0)}{\Delta z_{\max}} = \frac{c}{bw'} \cdot [z(20kHz) - z(0)], \tag{4.8}$$

which only depends on the width of the filter bands, *bw'*, and on the product of bandwidth and length of the impulse response, *c*. An increase of the number of filter bands beyond this value yields (theoretically) no additional information.

- **Example 1 (analytical solution)**

  In an FDC filter bank, the minimum filter order is mainly determined by the required side lobe attenuation. When the filter bandwidth is 3/5 of a critical bandwidth (which corresponds to one ERB), second order filters show a sufficient side lobe attenuation for a dynamic range of approximately 80 dB (Figure 4.5). In the case of second order FDC filters, the product of bandwidth and impulse response length is two, and Eq. (4.8) yields a maximum required number of 80 filter bands.

- **Example 2 (numerical solution, using a DFT)**

  The discrete Fourier transform of a filter shape represented over a range from -20 to 40 Bark on the critical band scale consists of 200 non-zero values and numerous leading and trailing zeros. An excitation pattern obtained from a filter bank in which all filters have this filter shape and are equally distributed over the given range of the critical band scale, is then completely represented by a total of 200 filter outputs. As only the range between zero and 24 Bark is of interest, a number of $200 \cdot 24/60 = 80$ filters is sufficient. The number of accessible filter outputs can then, if necessary, be multiplied by inserting virtual filters (which do not actually have to be calculated) with an output value of zero and smoothing the resulting pattern using an ideal low pass. When compared to the processing of sampled time signals, this procedure corresponds to a sampling rate conversion.

4.3.4 Rectification

With respect to the signal processing in the human auditory system, a half-wave rectification would be the most realistic rectification strategy. However, this would require a high effort in the post-processing of the rectified signals, to ensure that the excitation caused by a sine tone yields a perfectly flat temporal envelope. Even though a flat envelope is probably not present in the neural excitations, a sine tone is always perceived as a constant sound event. Together with a half-wave rectification this can only be achieved using either low pass filters of a high order or a peak detection algorithm. Both would introduce an increased complexity into the model which is not justified by a corresponding improvement in the expected performance of the model. Moreover, this strategy would add several degrees of freedom to the model, which makes a reliable experimental optimisation of the complete model more difficult.

A more convenient way of rectifying the filter outputs is to adopt the rectification strategy used in FFT-based approaches. This can be done by calculating the Hilbert transform of the filter outputs, which represents the imaginary part of the filtered signal, and computing the temporal envelope by adding the squared values of each filter output and its Hilbert transform. As a filter realising a Hilbert transform can only be approximated with a high computational expense, it is more convenient and also more precise to include the Hilbert-transform into the band pass filters. This is done by replacing each band pass filter by a pair of band pass filters with a 90° shift between their phase responses. The FDC-filter described in the previous section already supplies such filter pairs.

The advantages of this approach are the possible subsampling of the filter outputs (cf. Section 4.3.3), and the property of perfectly flat temporal envelopes for steady-state signals without a need for large time constants.

4.3.5 Time Domain Smearing

Whereas simultaneous masking is already taken into account in an earlier stage of the filter bank, temporal masking has yet to be modelled by low pass filtering the signal envelopes after the rectification. As the temporal resolution of the filter bank is still extremely high at this stage, there are almost no restrictions on the temporal masking curves to be modelled. Nevertheless, the low pass filters used to model temporal masking should not be too complex, because this would increase both the computational effort of the model and the number of degrees of freedom in the parameter settings. The low pass filters consist of two stages, a raised cosine shaped FIR low pass, and a first order IIR low pass. The former filter mainly is responsible for the ascending slope of the complete filter, and the latter one mainly influences the descending slope. The ascending slope models backward masking, and the descending slope models forward masking. The time constant of the IIR low pass depends on the centre frequency of the corresponding auditory filter and is given by

$$\tau(f_{centre}) = \tau_0 + \frac{100Hz}{f_{centre}} \cdot (\tau_{100} - \tau_{\min}) \tag{4.9}$$

The length of the FIR low pass is equal for all filter bands. Both the length of the FIR low pass and the limiting time constants of the IIR low pass were subject to experimental optimisation. As the FIR low pass has a reasonably good stop-band attenuation, the signal can then be subsampled according to its bandwidth.

4.3.6 Subsampling

In order to limit the computational effort, the sampling rate at which the time-frequency patterns are processed is reduced at the output of the FDC filter bank and after the first stage of the low pass filters used for time-smearing. The local sampling rates at the different stages of processing are given in Figure 4.4. At a sampling rate f_{samp} of 48 kHz, reasonable model performance was achieved using values between 8 and 32 for the sampling rate reduction after the filter bank, r_{filter} , and values between one to twelve for the sampling rate reduction after the first time-smearing low pass, $r_{time\text{-}smearing}$, which determines the modelled amount of backward masking.

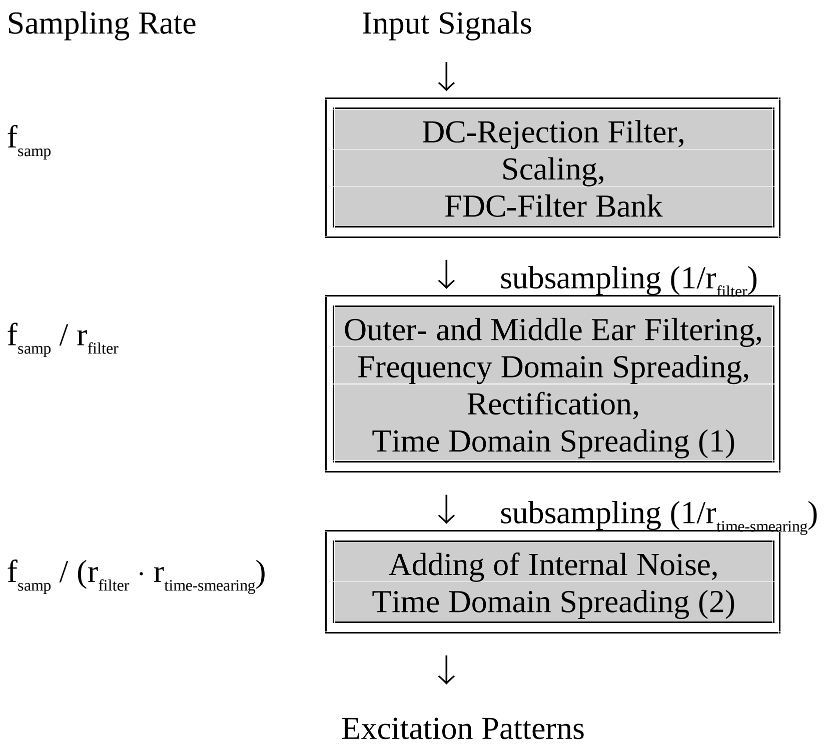

Sampling Rate
Input Signals
f_{samp}
DC-Rejection Filter, Scaling, FDC-Filter Bank
subsampling ($1/r_{filter}$)
f_{samp} / r_{filter}
Outer- and Middle Ear Filtering, Frequency Domain Spreading, Rectification, Time Domain Spreading (1)
subsampling ($1/r_{time\text{-}smearing}$)
$f_{samp} / (r_{filter} \cdot r_{time\text{-}smearing})$
Adding of Internal Noise, Time Domain Spreading (2)
Excitation Patterns

Fig. 4.4: Subsampling in DIX.

4.3.7 Threshold in Quiet

The threshold in quiet is modelled in two stages: in the first stage the input signal is filtered using an FIR filter that represents the parts of the threshold in quiet that are usually assigned to the outer and middle ear transfer function, and in a later stage a frequency dependent offset is added to the excitation patterns, representing internal noise. Similarly to other perceptual models, both parts are derived from the approximation of the threshold in quiet given in [TER79] (cf. Eq. 2.47)

$$\begin{aligned} threshold\,/\,dB = {} & 3.64 \cdot \left(f\,/\,kHz \right)^{-0.8} - 6.5 \cdot e^{-0.6 \cdot \left(f\,/\,kHz - 3.3 \right)^2} \\ & + 10^{-3} \cdot \left(f\,/\,kHz \right)^4 \end{aligned} . \tag{4.10}$$

The middle ear transfer function is modelled by the equation

$$\begin{aligned} A_{middle\ ear}\,/\,dB = {} & -w \cdot 3.64 \cdot \left(f\,/\,kHz \right)^{-0.8} + 6.5 \cdot e^{-0.6 \cdot \left(f\,/\,kHz - 3.3 \right)^2} \\ & - 10^{-3} \cdot \left(f\,/\,kHz \right)^4 \end{aligned} , \tag{4.11}$$

where *w* determines the distribution of the low frequency roll-off of the threshold in quiet between internal noise and middle ear transfer function. This weighting factor was set to *0.6*. The internal noise function is modelled by the remaining part of Eq. (4.10)

$$E_{internal\ noise} / dB = (1 - w) \cdot 3.65 \cdot (f / kHz)^{-0.8}. \qquad (4.12)$$

In order to save computational effort, the internal noise is added after the first stage of the time-smearing, where the sampling rate is lower than at the output of the FDC filter bank. The patterns obtained after this operation, but before the second stage of the time-smearing low pass filter, are used for the calculation of modulation patterns and are referred to as "unsmeared excitation patterns".

4.3.8 Characteristics of the Filter Bank

This section illustrates the properties of the filter bank. The filters are of linear phase, which is known to differ from the properties of the auditory filters, but has the advantage of preserving the temporal shape of the signals and yielding a rather good temporal resolution (when defined as the equivalent rectangular bandwidth of the impulse responses). The spectral and temporal characteristics of the filter bank are shown for the examples of pulses and sine tones as input signals.

a) Side Lobe Attenuation and Filter Order

The appropriate filter order of an FDC filter bank is mainly determined by the required side lobe attenuation. The amplitude of the side lobes should in no place exceed the amplitude of the corresponding auditory filter shape. As the amplitudes of the side lobes of an M^{th} order FDC filter decrease with the *(M+1)th* power of the distance to the main lobe

$$A_{side\ lobe} = \frac{a}{\left| \frac{f - f_{centre}}{bw} \right|^{M+1}} \quad , \qquad (4.13)$$

this requirement can be written as

$$\left. \frac{a}{\left| \frac{f - f_{centre}}{bw} \right|^{M+1}} \overset{!}{\leq} 10^{-\frac{S}{20} \cdot [z(f) - z(f_{centre})]} \right|_{\text{for all values of } f} \quad , \qquad (4.14)$$

which cannot be solved analytically, but can easily be solved numerically. Apparently, this requirement is easily fulfilled when the filter bandwidths are low, but becomes more critical when the filter bandwidths are high.

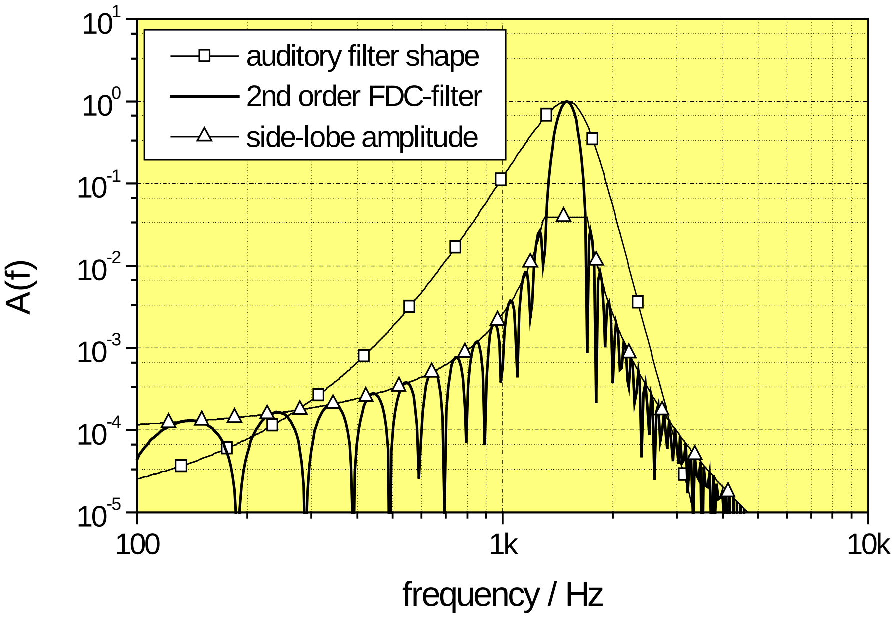


Fig. 4.5: Side lobes of a second order FDC-filter compared to an auditory filter shape.

When the filter bandwidth is 3/5 of a critical bandwidth (which corresponds to one ERB), second order filters show a sufficient side lobe attenuation for a dynamic range of approximately 80 dB (Figure 4.5). In practice, the side lobe attenuation is slightly enhanced by the spreading operation carried out to achieve exponential filter slopes.

b) Response to Sine Tones

Figure 4.6 shows the excitation pattern at the onset of a 1 kHz sine tone. The filter bank in this example consists of 40 second order FDC filters, which overlap at their -6 dB points (this was also the final configuration of the model after the experimental optimisation). The position of the maximum excitation corresponds to the frequency of the sine tone. The increase of the excitation in the left hand part of the figure reflects the internal noise function. The front part of the figure gives an impression of the response of the filter bank to steady-state signals, whereas the parts in the back show the response to the onset of the tone, and mainly reflect the temporal resolution of the filter bank and the shape of the low pass filters that model temporal masking.

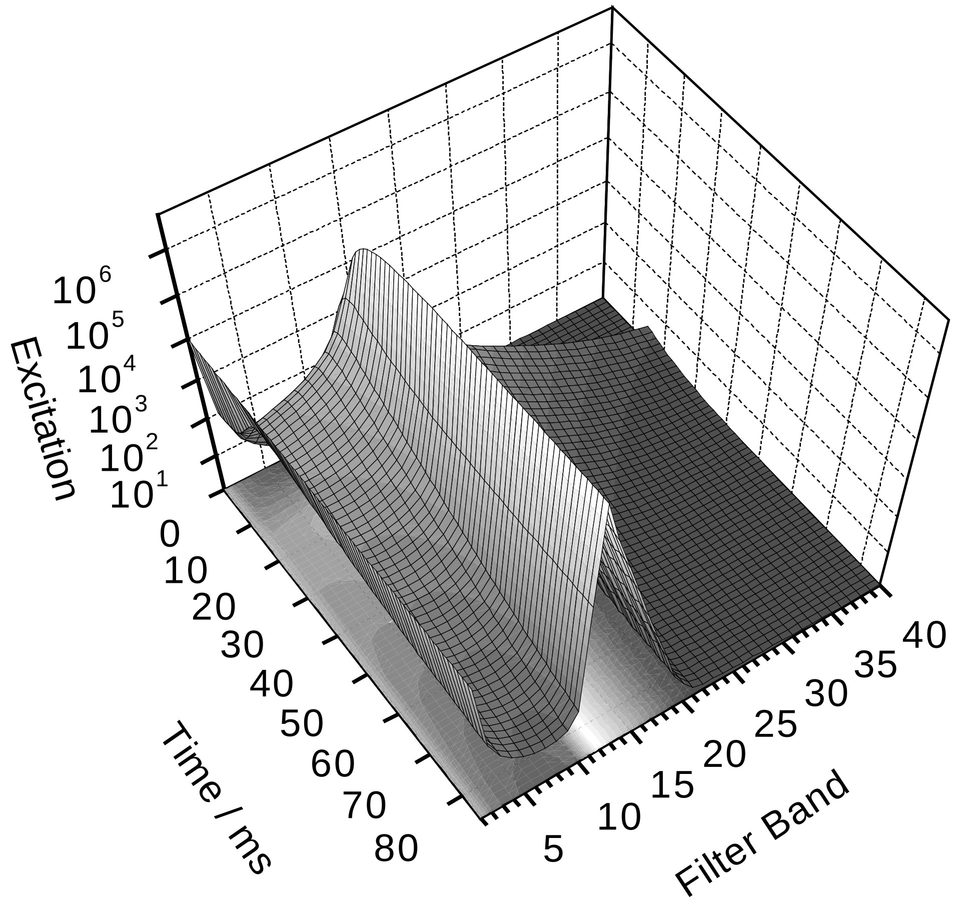


Fig. 4.6: Excitation pattern at the onset of a 1 kHz sine tone.

c) Response to Pulses

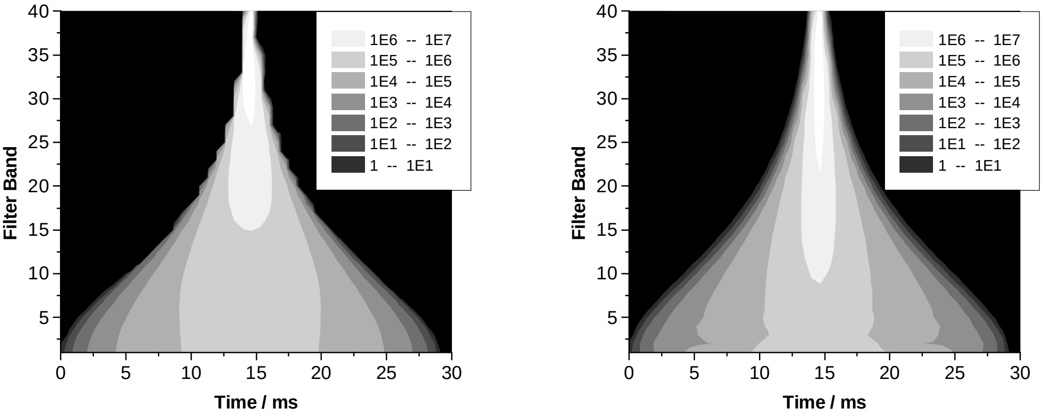


Fig. 4.7: Impulse response of the FDC filter bank without both spectral smearing and temporal smearing.

Fig. 4.8: Impulse response of the FDC filter bank without temporal smearing, but with spectral smearing.

Figures 4.7 and 4.8 show the amplitudes of the impulse responses of the filter bank when no time domain smearing is applied. In Figure 4.7, the modelling of the exponential filter slopes is omitted as well. The figures show that the temporal resolution in the upper filter bands is much higher than in the lower filter bands (which is one of the main advantages of filter banks).

Furthermore, the figures corroborate that the spectral smearing carried out for the modelling of exponential auditory filter slopes preserves the relation between spectral and temporal resolution. Obviously, the reduction of the spectral resolution when modelling the exponential filter slopes concentrates the energy of the impulse responses to a considerably reduced time period. When representing the impulse response of a single filter band on a linear scale, this becomes even more evident (Figure 4.9). The figure also demonstrates that the impulse responses change in correspondence to the level dependence of the auditory filter shapes.

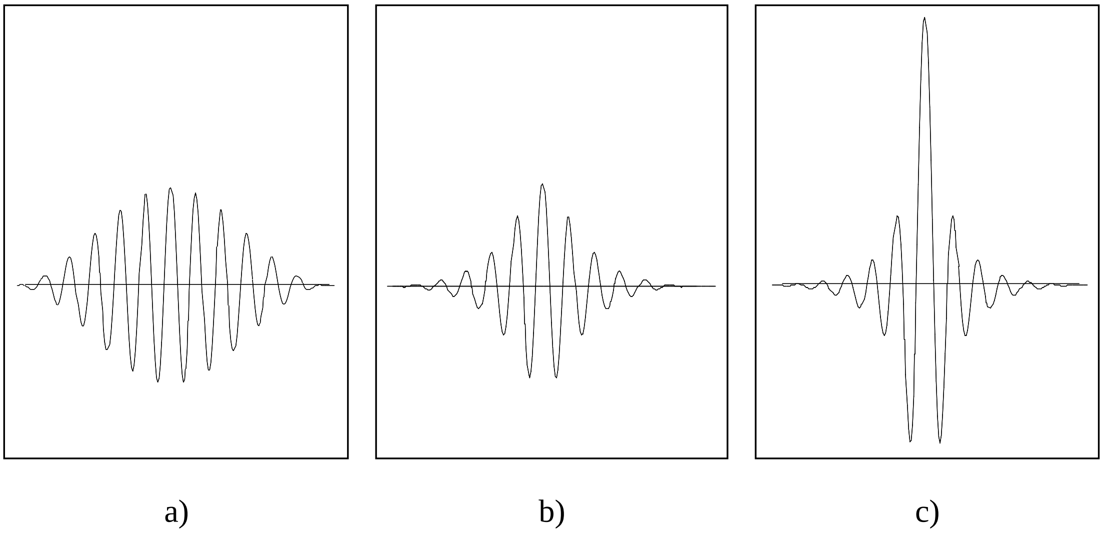


Fig. 4.9: Impulse responses of one filter channel (a) prior to and (b, c) after the spreading operation for two pulses of different energy (b: low energy, c: high energy). The shape of the pulse response after the complex spreading operation closely corresponds to the post-masking curves shown in [ZWI67].

d) Stability

As already mentioned in the description of the FDC filter algorithm, stability of the filters is only guaranteed as long as no rounding errors occur. This is the case when using integer arithmetic, but not when using floating point arithmetic.

When implemented using integer arithmetic, the filters would be truly linear and stability could be checked either by testing whether the pulse response has a finite energy, or, in the z-domain, by checking the position of the poles and zeros. If the filters became unstable because of the quantisation of the filter coefficients, stability could be enforced by adjusting the filter coefficients accordingly. However, the use of integer arithmetic is inconvenient because the word length of the integer would have to be larger than the 32 bit format provided by most current processors to achieve a sufficient accuracy.

When implemented using floating point arithmetic, the filters are not perfectly linear anymore because of the rounding errors occurring in the last digit of the floating point values. Checking the stability of the filters becomes more complicated because stability now depends on the input signals.

Under the assumption that rounding errors introduce a kind of uncertainty in the position of the poles and zeros of the filter the cancellation between the pole and one of the zeros of the filter cannot be perfect anymore. The pole may therefore remain, and, due to the uncertainty of its position, it may be on or outside the unit circle. Hence, the filter is not perfectly stable anymore. Stability can only be achieved by moving the pole sufficiently far into the unit circle in order to remain inside the border of the unit circle even with the worst possible rounding error. This can be easily achieved without changing the filter structure by simply reducing the absolute values of the filter coefficients of the recursive part of the filters. To retain the original filter characteristics, the positions of the zeros would have to be changed accordingly. As this destroys the symmetry of the filter coefficients, the computational effort increases. For this reason, the amount of errors produced by the potential instability of the filters was investigated.

When an analogue band pass filter becomes unstable, it normally starts to oscillate at its centre frequency, and the amplitude of the oscillation increases until it is limited by clipping. Because of the symmetry of the coefficients, the FDC filter depicted in Figure 3.11 cannot produce a considerable error when the input is constant. The magnitude of the error signal will therefore approximately remain constant when no input signal is present and increase during the input signal. A rough estimation of the error can be made by assuming that for each output value the error magnitude is given by the floating point accuracy (approx. $6 \cdot 10^{-8}$ when using single precision) multiplied by the signal energy. Under the assumption that the errors occurring in successive samples are statistically independent, the error energies add up, and the total error magnitude is proportional to the duration of the signal. Since the filter also includes subtractions, the relative error energy might be somewhat larger than expected from the floating point precision. It is also possible that the assumption of energy summation does not hold and, among successive samples, amplitudes add up, which would further increase the error energy.

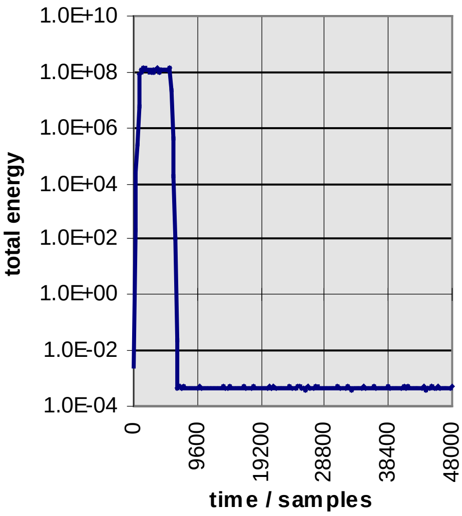


Fig. 4.10: Response to a 100 ms noise burst followed by silence.

The filter bank was tested against noise bursts, sine bursts, and single pulses. The settings of the filter bank were identical to those used for measurement except that the level dependence of the filter slopes was switched off (because in the case of short pulses, very high signal amplitudes had to be used).

The first experiment was to investigate the development of the error magnitude after switching off a signal. Figure 4.10 shows the total energy of all filter bands after a

100 milliseconds noise burst. It shows that the error signal is sufficiently far below the maximum signal energy when related to the dynamic range of the auditory system. It also shows that the error remains approximately constant after the signal is switched off.

Next, the dependence of the error magnitude on the length of the signal was tested. For this, the energies at the filter bank outputs were measured after switching off white-noise bursts of different duration.

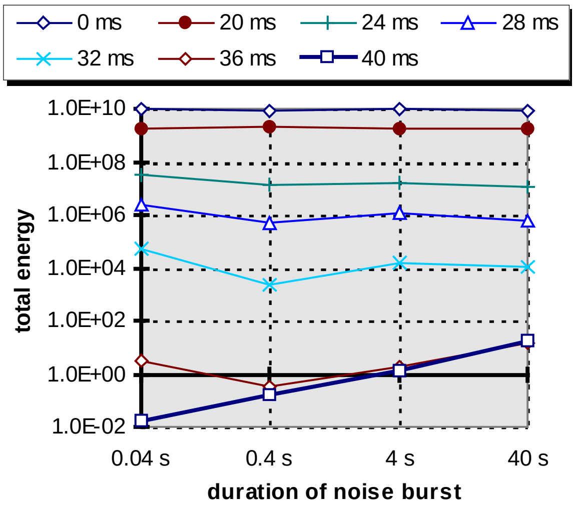


Fig. 4.11: Trailing noise produced by the filter bank as a function of the duration of a 100 dB noise burst. Parameter is the elapsed time after switching off the noise burst.

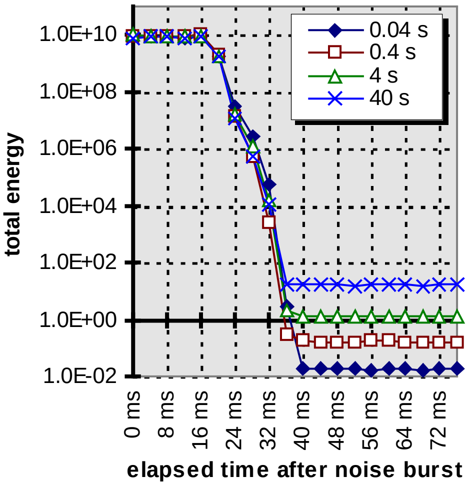


Fig. 4.12: Trailing noise produced by the filter bank as a function of the elapsed time after switching off a 100 dB noise burst. Parameter is the duration of the noise burst.

Figure 4.11 shows the dependence of the error magnitude on the length of the input signal. Curves are shown for zero to 40 milliseconds after the noise burst. After 40 milliseconds, the signal remains constant. This curve represents the error signal. The slope of this curve is almost exactly one, which corroborates the assumption of energy summation. The other curves do, of course, not represent an error but simply the decay of the filter bank response to the noise burst.

In Figure 4.12 the same data is shown as a function of the time elapsed after the noise burst and with the duration of the noise burst as parameter. Again, the decay of the filter bank response to the noise is followed by an almost constant component representing the errors. Even after the 40 seconds noise burst, the error is almost 90 dB below the maximum. When using a sine burst instead of a noise burst as test signal, the results are in the same order of magnitude, but the curves are not as smooth as those obtained with the noise bursts.

As 40 seconds is in the range of the longest test signals used for codec comparison tests, the potential instability of the filter bank will usually have no significant influence on the measurement results. For test items considerably longer than that, the produced errors can reach an order of magnitude where they possibly affect the measurement results.

When using double precision floating point arithmetic, the error amplitude is reduced by a factor of 10^6 and, as the error energy is proportional to the item length, the duration up to which the filter bank remains stable is increased by a factor of 10^{12}. This means, the filter bank can then be considered as perfectly stable when related to the dynamic range (and lifetime) of the auditory system.

4.4 Alignment and Adaptation

4.4.1 Introduction

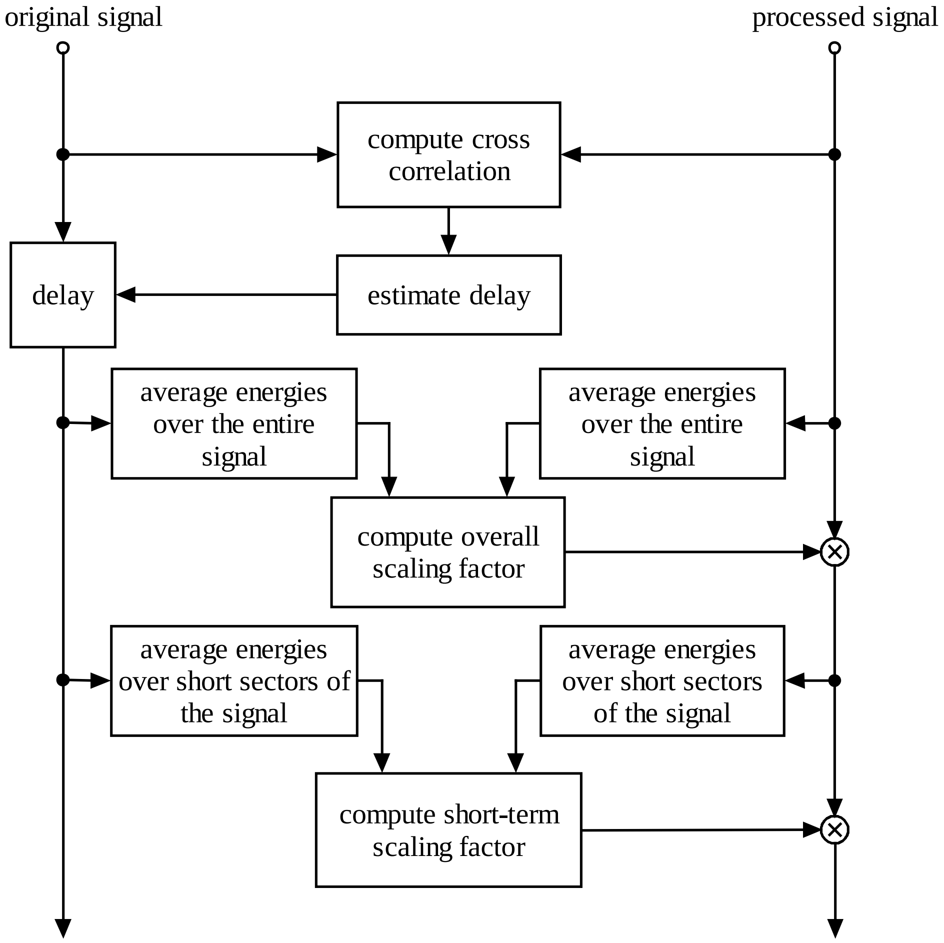

original signal
processed signal
compute cross correlation
delay
estimate delay
average energies over the entire signal
average energies over the entire signal
compute overall scaling factor
average energies over short sectors of the signal
average energies over short sectors of the signal
compute short-term scaling factor

Fig. 4.13: Level adaptation and time alignment within a typical perceptual measurement method.

Not all kinds of differences between test signal and the original signal are perceived as errors by a normal listener. This holds especially for signal delays and for a constant amplification or attenuation. Slow changes of the amplification may not be perceived as an error as well, or are at least less annoying than additive distortions. Hence, delays and level differences are usually compensated for before the perceptual model is run (Figure 4.13).

Apart from those signal differences which are not audible (or at least not annoying) at all, there are also signal differences which are audible, but less annoying compared to other kinds of distortions. This holds especially for changes in the spectral envelope (linear distortions) which are caused by a non-uniform frequency response of the device under test. The compensation for such linear distortions is performed by adapting the spectral envelopes of original and processed signal to each other. As linear distortions are not completely inaudible, but only less annoying, both the excitation patterns before and after the pattern adaptation must be evaluated separately. Therefore, the model splits the internal signal representation into two parts, one including and one excluding linear distortions. This can be regarded as modelling an aspect of perceptual streaming.

When a pattern adaptation is carried out, care has to be taken to prevent the adaptation algorithm from compensating also for errors which originate from non-linear distortions, like, for example, additive noise.

4.4.2 Time Alignment

Time alignment is not an integral part of the perceptual model. The input signals have to be time aligned before the measurement algorithm is applied. For most test signals, this is easily achieved by computing the cross-correlation function of the temporal envelopes of original and processed signal and determining the position of the maximum of the correlation function, which corresponds to the delay of the processed signal. However, for some test signals, the delay estimation is more difficult.

- When the signals include long sequences of periodical signal forms, the cross correlation function has more than one maximum and the appropriate delay cannot be found during these parts of the signals.
- When the distortions in the processed signal are very large, the maximum of the cross correlation function becomes less significant and its position cannot be determined exactly.
- When large parts of the signals solely consist of low frequency components, the maximum of the cross correlation function becomes very flat. Additive noise in the processed signals can then introduce a large amount of uncertainty into the delay estimation.

Therefore, the delay estimation was complemented by a visual control of the estimated delay, and the delay was either adjusted manually or was determined using manually selected sections of the signal which allowed a more reliable delay estimation (for example, regions around the attack of a musical instrument).

4.4.3 Dynamical Level and Pattern Adaptation

As the auditory system processes sound events continuously, and also adapts itself continuously to the signal characteristics, the adaptation algorithm in a perceptual model should not require a priori knowledge about the entire signal, but rather do a continuous processing as well. For this reason, a steady-state level and pattern adaptation is not modelled, and all adaptations are performed dynamically.

The dynamical adaptation of the levels and spectral distributions of original and processed signal is based on the local energies of the corresponding excitation patterns, smoothed in time by first order low pass filters. The time constants depend on the centre frequencies of the filters and are chosen as

$$\tau(f_{centre}) = \tau_0 + \frac{100Hz}{f_{centre}} \cdot (\tau_{100} - \tau_0) \ , \tag{4.15}$$

where the limiting time constants, τ_0 and τ_{100}, were set to five times the value of the time constants used in the modelling of forward masking.

When adapting original and processed signal to each other the adaptation must solely be based on signal components that are common to both signals, but has to be independent of components that only exist in either of the signals. If this is not considered, the adaptation may suppress errors which actually should be measured, or may even introduce errors in regions which actually were error free. A typical example for the latter case is given by the following situation: if the processed signal is band-limited, an adaptation of the overall levels will amplify the energy of the processed signal in the pass band of the device under test. This might then be interpreted as additive noise in a frequency region where no errors are present.

A reliable decision about which signal components of the processed signal belong to the original signal, and vice versa, would require a complete model of perceptual streaming. This is, of course, not possible, because it would not only require a detailed model of all signal recognition effects, but also a database including all “known” signals that can be recognised as separate events (which means, a catalogue of the life-time listening experience of the test listener). However, a simplified model of perceptual streaming can be established by regarding either the complete original signal or the complete processed signal as one auditory event, and assuming that the time-frequency patterns of the remaining auditory events are orthogonal to the time-frequency patterns of the original event. Orthogonality is defined by the relation

$$\int_{-\infty}^{\infty} A(x) \cdot e(x) \, \mathrm{dx} = 0 \, , \tag{4.16}$$

where *x* can be either the time or the frequency variable. If the signal *B(x)* consists of a fraction of the signal *A(x)*, and one part, *e(x)*, which is orthogonal to *A(x)*,

$$B(x) = m \cdot A(x) + e(x), \tag{4.17}$$

the weighting of the part of *A(x)* included in *B(x)* is given by

$$m = \frac{\int_{-\infty}^{\infty} A(x) \cdot B(x) \, \mathrm{dx}}{\int_{-\infty}^{\infty} [A(x)]^2 \, \mathrm{dx}} \, . \tag{4.18}$$

This is similar to the calculation of a cross correlation coefficient, but, unlike cross correlation, the orthogonality relation is not symmetrical. Therefore, a choice has to be made whether it is more appropriate to look for components of the original signal that are present in the processed signal, or to look for components of the processed signal that are present in the original signal. This choice depends on the character of the expected differences between original and processed signal.

Of course, the adaptation should not make signal components audible which originally are inaudible. To avoid this, the adaptation should never amplify signal components, but only attenuate them. Hence, it is not possible to consistently either adapt the processed signal to the original signal or vice versa. Instead, the stronger signal always has to be adapted to the weaker signal.

a) Level Adaptation

When adapting the short-term signal levels of original and processed signal to each other, the error source most likely is a band-limitation in the processed signal. In this case, the adaptation factor should only be determined by the energy ratio between the signal components within the pass band of the device under test. This factor can be derived from the orthogonality relation when the original signal is divided into one part that is preserved in the processed signal and one part that is not preserved in the processed signal. It is also assumed that the level adaptation is based on amplitudes rather than energies. When Eq. (4.18) is rewritten for discrete spectral patterns where k denotes the filter band and n denotes the input sample, the ratio between the preserved signal components is then given by

$$R_{level}(n) = \left(\frac{\sum\limits_{k=0}^{Z-1} \sqrt{E_{proc}(k,n) \cdot E_{orig}(k,n)}}{\sum\limits_{k=0}^{Z-1} E_{proc}(k,n)} \right)^2 \tag{4.19}$$

If the correction factor $R_{level}(n)$ is larger than one, the original signal will be divided by the correction factor, otherwise the processed signal will be multiplied by the correction factor:

$$E_{L,orig}(k,n) = E_{orig}(k,n)/R_{level}(n) \qquad \bigg| \qquad R_{level}(n) > 1\ , \tag{4.20}$$

$$E_{L,proc}(k,n) = E_{proc}(k,n) \cdot R_{level}(n) \qquad \bigg| \qquad R_{level}(n) \le 1\ , \tag{4.21}$$

where the index L denotes the excitations after the level adaptation, and the indices *orig* and *proc* denote the excitations for the original and processed signal, respectively.

b) Pattern Adaptation

When the spectral envelopes are to be adapted between original and processed signal, the distinction between common signal components and additive or missing signal components is made in the time domain. Here, the main problem is to distinguish within the processed signal between weighted components of the original signal and additive distortions. For this reason, the orthogonality relation is used in the opposite direction as in the level adaptation. As the adaptation should be able to change over time, the orthogonality relation given in Eq. (4.18) is not evaluated for the full signal length, but for a sliding time window. For reasons of computational efficiency, the time window is modelled by a first order low pass. The discrete representation of Eq. (4.18) can therefore be written as

$$R_{pattern}(k,n) = \frac{\sum_{i=0}^{n} a^i \cdot E_{L,proc}(k,n-i) \cdot E_{L,orig}(k,n-i)}{\sum_{i=0}^{n} a^i \cdot \left[E_{L,orig}(k,n-i)\right]^2} \qquad (4.22)$$

The values for *a* are derived from the time constants of the low pass, which are identical to those used in the level adaptation (Eq. 4.15).

If $R_{pattern}(k, n)$ is larger than one, the correction factor for the processed signal is set to $R_{pattern}(k, n)^{-1}$ and the correction factor for the original signal is set to one. In the opposite case, the correction factor for the original signal is set to $R_{pattern}(k, n)$ and the correction factor for the processed signal is set to one:

$$\begin{aligned} &R_{proc}(k,n) = \frac{1}{R_{pattern}(k,n)}, \quad R_{orig}(k,n) = 1 && \Big| R_{pattern}(k,n) \geq 1, \\ &R_{orig}(k,n) = R_{pattern}(k,n), \quad R_{proc}(k,n) = 1 && \Big| R_{pattern}(k,n) < 1. \end{aligned} \qquad (4.23)$$

The correction factors are averaged over *M* successive filter bands (Eq. 4.24) and smoothed over time, using the same time constants as before. The value of *M* is chosen such that the frequency window has a width of approximately two critical bands.

$$R_{P,proc}(k,n) = a \cdot R_{P,proc}(k,n-1) + (1-a) \cdot \frac{1}{M} \cdot \sum_{i=-M_1}^{M_2} R_{proc}(k-i,n)$$

$$R_{P,orig}(k,n) = a \cdot R_{P,orig}(k,n-1) + (1-a) \cdot \frac{1}{M} \cdot \sum_{i=-M_1}^{M_2} R_{orig}(k-i,n) \qquad (4.24)$$

$$\left| \begin{aligned} &M_1 = M_2 = \frac{M-1}{2} && \big| M \text{ odd} \\ &M_1 = \frac{M}{2} - 1, \quad M_2 = \frac{M}{2} && \big| M \text{ even} \end{aligned} \right.$$

Since at the limits of the frequency scale the frequency window would exceed the range of filter bands, the width of the frequency window is reduced accordingly:

$$M_1 = \min(M_1, k), \qquad M_2 = \min(M_2, z-k-1), \qquad M = M_1 + M_2 + 1. \quad (4.25)$$

The level adapted input patterns are weighted with the corresponding correction factors $R_{P,proc/orig}(k, n)$ in order to obtain the spectrally adapted patterns:

$$E_{P,Ref}(k,n) = E_{L,Ref}(k,n) \cdot R_{P,orig}(k,n) \qquad (4.26)$$

$$E_{P,Test}(k,n) = E_{L,Test}(k,n) \cdot R_{P,proc}(k,n)\,. \qquad (4.27)$$

The adaptation is directly applied to the outputs of the filter bank, but can optionally also be applied to the excitation patterns instead.

4.5 Evaluation of Envelope Modulations

When the temporal envelopes of the output signals of each auditory filter are evaluated, many effects of auditory perception can be modelled in a more logical way than in purely frequency domain based approaches. The structure of the temporal envelopes is taken into account by a *modulation measure* derived from a simple model of the detection of additive components in a steady-state signal which is explained below. The spectral distribution of the modulation measure is called the *modulation pattern*.

The asymmetry of masking between pure tones and noise-like signals can be explained when assuming that the auditory system enhances the detection of new components in a signal by suppressing the original signal. A very simple model of such a process is to predict the expected value of the excitation by its preceding values, and subtract the estimate from the actual value. The prediction can be realised by a linear filter which, in general, has a low pass characteristic. The suppression algorithm can thus be considered as a high pass filter. When the prediction is even more simplified by assuming that the estimated excitation is directly given by its preceding value, the high pass is replaced by a differentiator. Consequently, the masking threshold in such a model is proportional to the absolute value of the temporal derivation of the excitation:

$$M(f,t) \sim \left| \frac{dE(f,t)}{dt} \right|. \qquad (4.28)$$

As the masking threshold is normally expressed as the product of a threshold factor and the steady-state value of the excitation,

$$M(f,t) = s \cdot E(f,t), \qquad (4.29)$$

it is convenient to rewrite Eq. (4.28) in a similar form:

$$M(f,t) = \frac{c_0 \cdot \left|\frac{dE(f,t)}{dt}\right|}{E(f,t)} \cdot E(f,t), \tag{4.30}$$

where c_0 is a constant factor. In order to limit the threshold factor to a finite value, a small offset is added to the denominator. Moreover, the suppression of the masker cannot be perfect, and the threshold factor can therefore never be zero. This is taken into account by adding another offset, c_2:

$$M(f,t) = \left(c_2 + c_0 \cdot \frac{\left|\frac{dE(f,t)}{dt}\right|}{c_1 + E(f,t)} \right) \cdot E(f,t). \tag{4.31}$$

The term in parentheses determines the threshold factor. The right hand part of the term in parentheses is a measure for the amount of modulation in the excitation patterns, and is proportional to the threshold factor. Among the three free parameters in Eq. (4.31), c_0 and c_2 determine the mapping between modulation measure and threshold factor, and c_1 limits the modulation measure at very low excitation levels. The threshold factor calculated this way takes not only the masking asymmetry between tonal and noise-like signals into account, but also roughly follows the frequency dependence of the threshold factor for the masking of noise by pure tones, because the bandwidth of a critical band noise increases at high centre frequencies, and consequently its temporal derivation can assume larger values. The same effect also contributes to the modelling of the additivity of masking, as multi-component maskers have a wider spectrum than the individual components.

When assuming that the suppression procedure takes place at an early stage of auditory processing, i. e., before temporal integration takes place, the temporal integration has also to be applied to the suppressed excitation patterns. For this reason, the term

$$\left|\frac{dE(f,t)}{dt}\right| , \tag{4.32}$$

which approximates the suppressed excitation patterns in Eq. (4.31), is calculated from the *unsmeared excitation patterns* instead of the final excitation patterns and is smoothed over time by the same low pass filter as used for the modelling of forward masking. Furthermore, it is assumed that the suppression procedure from which the modulation measure is derived is based on loudness rather than excitations. Therefore, the excitations are replaced by a simplified loudness which is approximated by a power function of the excitations with an exponent of *0.3*. Summarising this, the calculation of the modulation measure is carried out in the following steps:

From the *unsmeared excitation patterns*, a simplified loudness is calculated by raising the excitation to a power of *0.3*. These values, $E'(f, t)$, and the absolute values of their temporal derivation are smeared out over time, using the same filter as in the modelling of forward masking:

$$E_\Delta(f,t) = \int_{t'=-\infty}^{t} \left|\frac{dE'(f,t')}{dt}\right| \cdot e^{\frac{t'-t}{\tau(f)}} dt', \tag{4.33}$$

and

$$\overline{E}(f,t) = \int_{t'=-\infty}^{t} E'(f,t') \cdot e^{\frac{t'-t}{\tau(f)}} dt'. \tag{4.34}$$

As in the modelling of forward masking, the time constants $\tau(f)$ depend on the centre frequency of each filter band and are given by

$$\tau(f_{centre}) = \tau_0 + \frac{100Hz}{f_{centre}} \cdot (\tau_{100} - \tau_0)\ . \tag{4.35}$$

The limiting time constants are also in the same range as in the modelling of forward masking. From the values E_Δ and $\overline{E}$ the modulation measure is calculated by normalising the temporal derivation of the envelope by its magnitude:

$$Mod(f,t) = \frac{E_\Delta(f,t)}{1 + \frac{1}{c_1} \cdot \overline{E}(f,t)}\ . \tag{4.36}$$

As indicated above, the modulation measure, $Mod(f, t)$, is mainly used in order to determine the threshold factor (see Section 4.6.1). Together with the simplified loudness, $\overline{E}(f, t)$, it is also used to calculate a separate measure for errors in the envelope modulation (see Section 4.6.3).

4.6 Model Output Values

The model yields output values for each processed sample. The following descriptions always refer to the momentary values of the model parameters. The averaging of the momentary model output values over time is described in a separate subsection.

4.6.1 Measures for Non-Linear Distortions

a) Partial Loudness of Additive Distortions

The most important attribute of a distortion is its loudness. In the quality range addressed by the measurement method, the distortion is usually close to the masking threshold, and therefore partially masked by the original signal. A reliable method for the calculation of the partial loudness of complex sounds should therefore be a good starting point for a perceptual measurement method. However, neither the approach given in [SCH79] nor the method recently proposed in [MOO97] yield results that can be correlated to subjectively perceived quality.

Whereas former approaches for the calculation of partial loudness were based on the results of simple psychoacoustical experiments, the partial loudness calculation used in DIX was designed to yield a consistent transition between models for auditory perception near threshold and the loudness of the signals in the absence of a masker. The partial loudness should satisfy the following criteria:

- In the absence of a masker or in situations where the level of the distortion is far above the masker level, the partial loudness should converge to the well-established loudness calculation proposed in [ZWI67].
- Near masked threshold, it should be possible to map the partial loudness to a detection probability. As derived in Section 2.4.6, this is achieved when the specific partial loudness converges to the ratio between distortion and masker. This is a common property of the main output variables of most established perceptual measurement methods, and can therefore be considered as proven by practical experience.
- The threshold factor should be calculated from the characteristics of the temporal envelopes of original (masker) and processed signal (masker plus maskee). Hence, it might be convenient to split the threshold factor in Zwicker's loudness formula into two parts, one depending on the properties of the masker alone, and one depending on the properties of masker and maskee together.

Zwicker's loudness formula (Eq. 2.45) can be rewritten as

$$N' = k \cdot \left(\frac{1}{s} \cdot \frac{E_{thres}}{E_0}\right)^{\gamma} \cdot \left[\left(1 + \frac{s \cdot (E - E_{thres})}{E_{thres}}\right)^{\gamma} - 1\right]. \tag{4.37}$$

The excitation at the threshold in quiet, E_{thres}, apparently plays two different roles in this equation. In the numerator of the second fraction, E_{thres} plays the role of the masker (the internal noise in the auditory system), and the difference $E\text{-}E_{thres}$ corresponds to the difference between the excitation produced by masker plus maskee and the excitation produced by the masker alone. In the denominator of the second fraction and in the numerator of the first fraction, E_{thres} models the absolute threshold in terms of limited detector sensitivity. For the case of partial masking, the excitations in the numerator of the second fraction can therefore be replaced by the excitations of the processed signal and the original signal

$$N' = k \cdot \left(\frac{1}{s} \cdot \frac{E_{thres}}{E_0}\right)^{\gamma} \cdot \left[\left(1 + \frac{s \cdot E_{proc} - s \cdot E_{orig}}{E_{thres}}\right)^{\gamma} - 1\right]. \tag{4.38}$$

As the envelope modulations which are used to determine the threshold factor (see Section 4.6.3) are calculated separately for processed and original signal, different values of s are used, depending upon whether it is used together with the original or with the processed signal. The threshold factor in the first term is used to compensate for the weighting of the loudness by the threshold factor of the second term. It is therefore more appropriate to use the threshold factor for the processed signal here than to use the threshold factor of the original signal. However, this remains one of the weak points when trying to relate this formula to Zwicker's specific loudness formula.

$$N' = k \cdot \left(\frac{1}{s_{proc}} \cdot \frac{E_{thres}}{E_0} \right)^{\gamma} \cdot \left[\left(1 + \frac{s_{proc} \cdot E_{proc} - s_{orig} \cdot E_{orig}}{E_{thres}} \right)^{\gamma} - 1 \right] \tag{4.39}$$

Equation (4.39) does not yet model partial masking. Partial masking is achieved by introducing an additional term in the denominator of the second fraction that increases when approaching masked threshold and disappears when the distortions are far above masked threshold.

An expression of the form $(1+a)^x$ can be approximated by $(1+ax)$ for small values of a. For small distortions, the partial loudness formula is therefore proportional to the second fraction in Eq. (4.39). On the other hand, in Section 2.4.6 it was shown that for small distortions the specific partial loudness has to be proportional to the ratio between error signal and original signal. As the numerator of the second fraction in Eq. (4.39) corresponds to the error signal, the denominator should be proportional to the excitation produced by the original signal. A very simple expression that meets this requirement is

$$E_{thres}' = E_{thres} + s_{orig} \cdot E_{orig} \cdot e^{\left(-\alpha \cdot \frac{E_{proc} - E_{orig}}{E_{orig}} \right)} . \tag{4.40}$$

Moreover, the specific partial loudness should never assume negative values. Therefore, the weighted excitation difference has to be limited to positive values. The complete formula for partial loudness calculation is then given by

$$N' = k \cdot \left(\frac{1}{s_{proc}} \cdot \frac{E_{thres}}{E_0} \right)^{\gamma} \cdot \left[\left(1 + \frac{\max\left(s_{proc} \cdot E_{proc} - s_{orig} \cdot E_{orig} ,\ 0 \right)}{E_{thres} + s_{orig} \cdot E_{orig} \cdot e^{-\alpha \cdot \frac{E_{proc} - E_{orig}}{E_{orig}}}} \right)^{\gamma} - 1 \right] . \tag{4.41}$$

The partial loudness depends upon three free parameters, the factor α, which determines the amount of partial masking, and the two coefficients used in the linear mapping of the modulation measure to the threshold factor. This mapping is expressed as

$$s_{proc} = m_s \cdot Mod_{proc}(f,t) + c_s \qquad and$$

$$s_{orig} = m_s \cdot Mod_{orig}(f,t) + c_s , \tag{4.42}$$

where m_s is in the range of 0.2 seconds and c_s is in the range of 1.0. All other constants are either pure scaling constants (like E_0 and k), or are already determined (like the exponent γ, which, according to [ZWI67], is set to 0.23).

The partial loudness measure is usually calculated from the excitation patterns obtained after the pattern adaptation. In this case, it is a measure of the impact of additive non-linear distortions and is called *Partial Loudness of Additive Distortions*.

The main shortcoming of the partial loudness measure is that it does not work for all combinations of modulation differences and excitation differences. If in any filter band the processed signal is stronger modulated than the original signal, but the local excitation is lower, both differences may compensate for each other in the partial loudness measure because of the term $s_{proc} \cdot E_{proc} - s_{orig} \cdot E_{orig}$ in Eq. (4.41).

The same might also happen in the opposite constellation. However, both cases do not happen as long as the assumption of additive distortions holds. An introduced signal component usually increases the modulation of the entire signal. It may lower the modulation only if it either is much larger than the original signal or a similar component is already included in the original signal. In the first case, the distortion is very large, and the probability that it might be compensated due to its lowered weighting is rather low. In the latter case, the error is not perceived as an additive distortion anymore, and a partial loudness measure is not the appropriate measure for this kind of error.

Another constellation where the partial loudness measure defined in Eq. (4.41) will not respond to differences between processed and original signal is when both modulation and excitation decrease. When this happens in larger regions of the time-frequency plane, the error is a linear distortion rather than an additive distortion. This case should be prevented by the level and pattern adaptation. However, in small regions of the time-frequency plane, this happens also in the case of additive distortions. It can be measured by interchanging the roles of processed and original signal in Eq. (4.41). The partial noise loudness derived for this case is called Partial Loudness of Missing Components. As this measure does essentially the same thing as the *Partial Loudness of Additive Distortions*, the mapping to basic audio quality should be the same for both measures except for a weighting factor. Consequently, they can be combined into one single quality measure by a weighted summation.

4.6.2 Measures for Linear Distortions

a) Loudness Pattern Correlation

A very simple measure that summarises all differences between the spectral envelopes of processed and original signal into one single value is the cross correlation between their internal representations. Except for the frequency warping due to the non-uniform distribution of the filter bands, this measure is not psychoacoustically motivated. When the excitation patterns are replaced by specific loudness patterns, this measure becomes slightly closer related to psychoacoustics. For this purpose, specific loudness patterns are calculated by applying Zwickers loudness formula (Eq. 2.45) to the excitations in each filter band. As the processed signal normally is still very similar to the original signal, the values of the cross correlation coefficient are normally close to one. Therefore, instead of the correlation coefficient itself, its difference to one is used as output parameter.

b) Partial Loudness of Linear Distortions

A better measure for linear distortions can be derived when the algorithm for the calculation of partial loudness is modified to yield the partial loudness of the components of the original signal that are lost in the processed signal. This is achieved by applying the algorithm described in Section 4.6.1 to the excitation patterns of the original signal before and after the pattern adaptation. The excitation pattern before the adaptation is used in the place of the processed signal, and the excitation pattern after the adaptation is used in the place of the original signal.

4.6.3 Measures for Changes in the Temporal Structure

One advantage of filter banks is their property of preserving the temporal envelopes within the auditory filter bands. In the auditory system, the temporal envelopes at the auditory filter bands are decomposed into modulation spectra, which, like the auditory filters, are represented over a non-uniform frequency scale. Modelling this is computational expensive, but, as the temporal envelopes at the outputs of the auditory filters are processed at a reduced sampling rate, the computational complexity is still low enough to allow to incorporate such a process in a perceptual model. However, the evaluation of modulation spectra turned out to be impractical because it results in multidimensional signal representations (energies as a function of centre frequency, modulation frequency and time) that cannot be mapped in a reasonable way to a single value. Therefore, more simple measures were used for the influence of envelope modulations on the perceived audio quality.

a) Envelope Correlation

Like in the evaluation of linear distortions, a simple cross correlation also appears to be the most simple way to summarise changes in the temporal envelopes of the signals into one single value. In each filter band, a cross correlation coefficient between the temporal envelopes of the excitation patterns or specific loudness patterns of original and processed signal is calculated within a sliding time window. The time window is modelled by a first order low pass with a time constant of 250 milliseconds. The distance measure is given by the difference of the correlation coefficient to one. It is averaged linearly over all frequency bands to yield the momentary envelope deviation measure

$$EnvDev = \sum_{\text{all filter bands}} 1 - \frac{\int_{-\infty}^{0} e^{\frac{t}{\tau}} \cdot E_{orig} \cdot E_{proc} dt}{\sqrt{\int_{-\infty}^{0} e^{\frac{t}{\tau}} \cdot E_{orig} \cdot E_{orig} dt \cdot \int_{-\infty}^{0} e^{\frac{t}{\tau}} \cdot E_{proc} \cdot E_{proc} dt}} . \tag{4.43}$$

b) Modulation Difference

The modulation measure that is used for the calculation of the threshold factor can also be used to derive a very simple measure for changes in the temporal envelopes. As the modulation measure is theoretically justified by its relation to the threshold factor, a difference measure based on this value is closer related to the properties of the auditory system than the above described envelope correlation. The local modulation difference measure is the absolute difference between the local modulation measures of original and processed signal, normalised by the local modulation measure of the original signal

$$ModDiff(f,t) = \frac{\left|Mod_{proc}(f,t) - Mod_{orig}(f,t)\right|}{Mod_{orig}(f,t)} \quad . \tag{4.44}$$

To limit the value of the modulation difference in the case when the original signal is not modulated at all, a small offset is added to the denominator. In order to take the fact into account that introduced modulations are likely perceived as additive distortions and therefore more annoying than left out modulations, a weighting factor is applied which depends on whether the processed signal is stronger or less strong modulated than the original signal

$$ModDiff(f,t) = w \cdot \frac{\left|Mod_{proc}(f,t) - Mod_{orig}(f,t)\right|}{offset + Mod_{orig}(f,t)} \quad , \tag{4.45}$$

where

$$\left| \begin{array}{ll} w = 1.0 & \left| Mod_{proc}(f,t) > Mod_{orig}(f,t) \right. \\ w = negWt & \left| Mod_{proc}(f,t) < Mod_{orig}(f,t) \right. \end{array} \right. . \tag{4.46}$$

The optimum weight for the lost modulations is between 0.1 and 1.0. The offset is between 0.01 and 1.0. Typically, when one of both parameters is set to a high value, close to one, the other parameter should be set to a small value. Both together determine the asymmetry of the modulation difference measure. When the offset is set to a small value, the amount of asymmetry increases when the test signal includes only small modulations. When the weighting factor is small, the amount of asymmetry depends less on the absolute amount of modulation.

As in test items with time-varying loudness the modulation differences in louder sequences of the signal contribute more to the perceived quality of the entire signal than in the more silent parts, a temporal weighting is applied to the momentary values of the modulation difference. The weighting factor is determined by the ratio between temporally smoothed excitation and threshold in quiet. It is calculated for each filter band, and averaged over filter bands before it is applied to the momentary modulation difference:

$$TempWt(n) = \sum_{k=0}^{Z-1} \frac{1}{1 + levWt \cdot \frac{E_{Thres}(k)}{\overline{E}(k,n)}} \quad . \tag{4.47}$$

The free parameter *levWt* determines the influence of the weighting. Best results are achieved with values between 1.0 and 100. If it is zero, no weighting is applied. The weighting is relative, i. e. the weighted average values are normalised by the average of the weighting factor.

4.6.4 Temporal Averaging

The momentary values of the model parameters are averaged linearly, squared or by their square root. Compared to the linear average, the squared average assigns a higher weight on the temporal maxima of the distortion, whereas the averaging of square roots tends to suppress the temporal maxima.

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

$$AvgX = \frac{1}{N} \cdot \sum_{n=0}^{N-1} X[n]. \tag{4.48}$$

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

$$RmsX = \sqrt{\frac{1}{N} \cdot \sum_{n=0}^{N-1} X[n]^2}\,. \tag{4.49}$$

- The square root average value (prefix *"Msr"*) is calculated by

$$MsrX = \left(\frac{1}{N} \cdot \sum_{n=0}^{N-1} \sqrt{X}[n] \right)^2. \tag{4.50}$$

In these equations, X stands for the model output variable and N is the number of time samples for which momentary values of X have been calculated.

Which averaging strategy works bests, depends strongly on the listening test data used for verification. In the perception of speech signals the perceived loudness is determined by the temporal maxima rather than by the average level [ZWI77]. Therefore, it was expected that also in the case of temporally varying audio quality, the maximum distortions are more important than their long term average. When the influence of different averaging strategies on the performance of the model was first investigated, this assumption was not corroborated. In general, the results of older listening tests are best reproduced when averaging linearly or by square roots. Nevertheless, the more recent listening tests are better reproduced when using the squared average. This might be a result of the increased experience of the test listeners that allows them to concentrate more on the moments where the maximum errors occur, but more likely it reflects the fact that none of the averaging strategies fully models the perception of sounds with temporally varying quality.

One effort to overcome the inconsistency in the temporal averaging strategies was the introduction of a windowed average value (prefix *"Win"*), which is calculated by

$$WinX = \sqrt[4]{\frac{1}{N-L+1} \cdot \sum_{n=L-1}^{N-1} \left(\frac{1}{L} \cdot \sum_{i=0}^{L-1} \sqrt{X[n-i]} \right)^4} \ , \qquad (4.51)$$

where L is the length of the sliding time window in time samples. The window length is approximately 100 milliseconds. The windowed average suppresses temporal maxima in the short term, but enhances temporal maxima in the long term.

4.6.5 Selection of Valid Sequences of the Audio Signal

Apart from inaudible or less annoying artefacts, which are handled by the level and pattern adaptation procedure, there are also artefacts that are both audible and annoying, but nevertheless do not influence subjective quality gradings, because listeners do not assign these artefacts to the device under test. This holds for errors at the onset of the audio signal and for some single events (for example one click within a 20 seconds test item).

A single error event within a long test signal may sometimes be assigned to the playback device instead of the device under test and has therefore only little influence on subjective quality gradings. In other cases, single error events are clearly recognised as an artefact of the device under test, and cause the test listeners to downgrade the whole test item. Handling this is rather difficult, if not impossible, because the circumstances that make a listener know that the error event is caused by the device under test cannot be predicted by a measurement method[6]. Moreover, the measurement method is supposed to detect audible errors, and, as a single click is an audible error, it should be measured, even if does not affect subjective gradings. Consequently, the detection of single events was neither suppressed nor enhanced, but test items that include single error events were not considered as relevant when optimising and testing the measurement method.

Unfortunately, one item including a clearly audible single click, which has not been recognised by the test listeners, has been included in the data set used in the first comparative test between perceptual measurement methods in the ITU. Consequently, for this particular test, it was necessary to suppress the detection of such errors. This has been done by comparing the square root average, which is less sensitive to single events, with the linear average. When the ratio exceeds a certain threshold (around 6), the overall error measure is probably determined by a single event, and the square root average is used instead of the linear or squared average.

Errors at the onset of an audio excerpt are often not recognised or not considered by test listeners. This is partly because a listener needs a short moment to "tune in" to the test signal before being able to distinguish between parts of the original signal and introduced artefacts. If the test signals are switched on without fading, a short part in the beginning of the test signal might be missing in the processed signal, and most listeners would not even notice this (it is the same signal, though one is a little bit

---

[6] Measurement methods with a low temporal resolution normally do not have to deal with that problem, because they do not detect single clicks anyway. However, avoiding this problem by lowering the temporal resolution is not an appropriate solution, because this clearly affects the overall performance of the model, and, as single clicks are audible, they should also be detected by the measurement method.

shorter). Truncations of the beginning or end of a test signal are also often errors of the audio files used when feeding the test signals into the measurement method, and may not have been present in the listening tests. Such events at the signal onset are suppressed by discarding the first 500 milliseconds of a test item, and discarding all artefacts that are measured before another 50 milliseconds after the loudness of the test signal has reached a certain threshold (normally: 0.1 sone) for the first time. The latter not only takes the fact into account that a listener needs to tune in to a signal, but is also required in order to allow the level and pattern adaptation of the measurement method to reach a steady state.

4.7 Optimisation of the Model

Most parts of the peripheral ear model of DIX were held very flexible in order to allow for an experimental optimisation. Subjects of optimisation were the distribution of the centre frequencies of the filter bands, the spectral and temporal resolution, the modelling of simultaneous and temporal masking, and the spectral alignment between processed and original signal. The performance criterion was the accuracy (in terms of a cross correlation) by which the results of several listening tests for the quality evaluation of audio codecs were reproduced (more detailed information on these test data is given in Section 6.2).

4.7.1 Auditory Frequency Scales

All established approximations for auditory frequency scales have been implemented in DIX, and the influence of the choice of the frequency scale on the performance of the model has been investigated. To eliminate the influence of the absolute width of the frequency units, all scales have been used with the same number of filter bands covering the audible frequency range.

It turned out that the approximation of the critical band scale proposed in [SCH79] yields the best results. It was not only superior to other frequency scales, like ERB or SPINC scale, but worked also better than the more accurate approximations of the critical band scale proposed in [TER79]. This is rather unexpected because the approximation given in [SCH79] was explicitly designed for frequencies below 5 kHz, which is sufficient for speech coding, but not for the evaluation of high quality audio codecs.

The most obvious difference between the approximation given in [SCH79] and the approximation given in [TER79] is, that the latter yields higher filter bandwidths in the upper frequency range. This might be an end-of-scale effect that does not really reflect the frequency resolution of auditory filter bands. With this explanation, the approximation given in [SCH79] would really be the better representation of auditory frequency resolution. This is somewhat corroborated by the similarity between this scale and the ERB-scale in the upper frequency range. However, it is also possible that the wider filter bands resulting from [TER79] simply cause practical problems in the measurement method.

4.7.2 Sampling in the Time and Frequency Domain

a) Sampling in the Frequency Domain

When investigating the required number of filter bands, the bandwidth of the filters was always adjusted accordingly to keep the overlap between adjacent filters constant. When the overlap was chosen according to Section 4.3.3, the optimum number of filter bands was around 80, which corresponds to a filter bandwidth of 0.6 Bark or one ERB. Using a higher number of filter bands did not yield any improvement, but also no deterioration. Reducing the number of filter bands down to 50 clearly affected the performance of the model on some sets of listening test data, but normally still yielded reasonable results. Later on, it turned out, that a reduced overlap where the distance between adjacent filters was doubled without changing their bandwidth and characteristics did not significantly affect the results. For this reason, the final version of the filter bank consists of 40 filter bands that overlap at their 6 dB points.

b) Sampling in the Time Domain

With the filter bandwidth of 0.6 Bark, which was found to be optimal, the bandwidth of the upper filter bands is approximately 1700 Hz, which means that, at least theoretically, a sampling rate of 3400 Hz or a subsampling by a factor of 14 is tolerable[7]. As the filters do not provide a perfect band pass characteristic, the sampling rate should remain somewhat higher. When trying different subsampling factors, it turned out that even lower sampling rates could be used. The subsampling factor could be increased up to a value of 32 without affecting the performance of the model. This is a great advantage when running the model on larger test data sets, because the subsampling significantly reduces the computational effort. Nevertheless, it has to be kept in mind that aliasing might occur with some test signals. Such test signals must include high frequency components that are amplitude modulated, and the modulating signal must include frequency components between 750 Hz and a little more than 1700 Hz. Moreover, noise-like signals cannot cause severe errors, because even when aliasing occurs, the aliasing components are similar to the baseband components. Therefore, such critical test signals are not very likely to occur, and are apparently not present in the available test data sets.

4.7.3 Simultaneous Masking

a) Slope Rates

The upper slope of the spreading function was modelled either by constant slope rates, by the worst case curve proposed in [BRA87], or by modelling the level dependence as described in [TER79]. When using constant slope rates, a clear optimum was found at a value around 12 dB/Bark. Both the worst case curve and the modelling of level-dependent slopes yielded a slight, but not really significant improvement.

---

[7] The maximum modulation frequency of the envelopes of a band pass signal is only half its bandwidth. But, as the squared envelope is processed, the bandwidth is doubled again, so that the same sampling rate still applies.

The final choice was, to use the level-dependent model, because it is closer to theory than the other approaches.

Changing the lower slope rate of the spreading function had only a minor influence on the performance of the model. Higher slope rates tended to give slightly better results on problem items. Therefore, the highest value for the lower slope rate found in literature, 31 dB/Bark, was used.

b) Shape of the Spreading Function

Spreading functions with a more smooth shape, like the approximation given in [SCH79] were tested, but did not improve the performance of the model. Apparently, the shapes of the band pass filters used in the filter bank already ensure sufficiently smooth shapes of the resulting auditory filters.

4.7.4 Temporal Masking

a) Backward Masking

The impulse response length of the FIR low pass that models backward masking is of major influence on the performance of the model. Best results were achieved with a length of approximately 400 samples, which corresponds to 8 milliseconds. As half of this length belong to the ascending slope, which is responsible for backward masking, and the 6 dB width again is half of this length, the value of 8 milliseconds corresponds to an effective length of 2 milliseconds for backward masking. This is in good correspondence with psychoacoustical data.

When reducing the impulse response length of the filter to less than 200 samples, the performance of the model starts to decrease very fast. When the filter is left out completely, the correlation between model predictions and subjective quality scores breaks down completely. When the impulse response length becomes larger than the optimum, the performance of the model decreases more slowly.

b) Forward Masking

The time constants of the IIR low pass that models forward masking turned out to be less critical. For minimum time constant τ_0, which determines forward masking at high and intermediate frequencies, the best results were achieved with a value of 8 milliseconds. However, using half or twice this value still yielded reasonable results. The other time constant τ_{100}, which determines forward masking at low frequencies, had almost no influence on the performance of the model when varied between 8 and 100 milliseconds. Obviously, the accessible data sets do not include any test items where time-varying distortions at low frequencies play a significant role.

4.7.5 Dynamical Level and Pattern Adaptation

The time constants used in the level and pattern adaptation had an astonishingly small influence on the results of the measurement method. Theoretically, they should be rather large because the auditory system also needs some time to adapt to the characteristics of an audio signal. Therefore, the default value of five times the time constants used for forward masking has been retained.

Whether or not the level and pattern adaptation is switched on, does not play a significant role when looking at any single listening test data set, but remarkably increases the performance of the model when merging data from different listening tests. When assuming that the main effect of the level and pattern adaptation is the streaming between linear and non-linear distortions, this might be explained by the different codec generations used in the different listening tests. Possibly, within one generation of codecs, linear and non-linear distortions are highly correlated because, when the bit rate is not sufficient for transparent coding, the codecs introduce both band-limiting and quantisation noise at the same time. Among different codec generations, the trade-off between band-limiting and quantisation noise may change. Hence, both effects have to be evaluated separately when comparing different generations of codecs.

4.8 Summary

DIX is a perceptual measurement method that provides an accurate ear model, and a high degree of flexibility at a reasonable computational effort. It offers an optimum product of time and frequency resolution, and therefore allows to investigate the effects of different time and frequency resolutions almost without interactions between both. It is capable of modelling level-dependent auditory filter slopes, and models several auditory phenomena by analysing the temporal envelope at the outputs of the auditory filters. Its main output is the partial loudness of additive distortions, but it also provides measures for linear distortions and changes in the modulation structure.

In the experimental optimisation of the model it turned out that the critical band scale is most suitable for the definition of the distribution of the filter bands. However, the optimum width of the filters is only 3/5 of the critical bandwidth, which, in the intermediate and upper frequency range, approximately corresponds to the equivalent rectangular bands measured by Patterson [PAT76]. Using these filter bandwidths, approximately 40 filter bands are sufficiently representing the auditory excitation patterns. The additional widening of the critical bands at very high frequencies, as assumed by Zwicker [ZWI67] and Terhardt [TER79] was not corroborated when investigating the performance of the perceptual model. A simplified approximation of the critical band scale that ignores this additional widening of the critical bands [SCH79] yielded better results than the original critical band scale. When using a raised-cosine time-window to limit the temporal resolution of the ear model, a window length of 8 milliseconds was optimum. This corresponds to a temporal resolution of 4 milliseconds and a time constant of approximately 2 milliseconds for backward masking. When modelling forward masking by an additional IIR low pass, a time constant of 8 milliseconds was optimum for the intermediate and upper filter bands. The time constants for the lower filter bands had no significant influence on the performance of the model, and could therefore not be determined.

5. PEAQ - The New Standard for Objective Measurement of Perceived Audio Quality

5.1 Introduction: Standardisation within the ITU

In 1994 the *International Telecommunication Union* (ITU) established a new task group, ITU-R TG 10/4, that was to prepare a recommendation for the objective measurement of perceived audio quality. In 1995 a call for proposals was released, and six perceptual measurement methods were considered in the further work of TG 10/4. The six proposals were (in order of appearance)

- NMR by the *Fraunhofer Institute for Integrated Circuits* (FhG-IIS), Erlangen, Germany
- PAQM by the *Royal PTT Nederland* (KPN), Leidschendam, Netherlands
- PERCEVAL by the *Communications Research Centre* (CRC), Ottawa, Canada
- POM by the *Centre Commun d'Etudes de Télédiffusion et Télécommunications* (CCETT), Rennes, France
- DIX by the *Technical University of Berlin* (TUB), Berlin, Germany
- “Toolbox” by the *Institut für Rundfunktechnik* (IRT), Munich, Germany

In 1996 the ITU conducted a comparative test among these methods in order to select one of the methods to become the final recommendation. The outcome of this test was that none of the proposed methods was yet reliable enough to become an international standard (the detailed results of the test are given later on in Section 7). Consequently, the model proponents were urged to establish a joint model that combines features of all original proposals in order to achieve the best possible performance. This joint model was named PEAQ (Perceptual Evaluation of Audio Quality). It should be validated approximately one year later, and the model that performed best in the first comparative test (PAQM) should serve as a reference. If the PEAQ performed clearly better than the reference model, it could be recommended for standardisation.

PEAQ is split into two parts: one filterbank-based model, and one FFT-based model. DIX served as a starting point for the filterbank-based model. If possible all features of the six original models should be integrated in both, filterbank-based and FFT-based model. Therefore, the peripheral ear model of DIX was re-implemented in a way that allowed an easy integration of features of other models, and the model parameters of DIX were re-implemented in a way that they could be integrated into a FFT-based model.

5.2 Combining Output Values of Different Models

As both, masked threshold concept and comparison of internal representations, have their own specific advantages and disadvantages, it is sensible to combine quality indicators based on each concept. Among the *model output variables* (MOVs) based on a comparison of internal representations, the model output variables originating from DIX performed clearly best with the filterbank-based ear model as well as with the FFT-based ear model. These MOVs where therefore used in all combinations of quality indicators. All model output variables based on the masked threshold concept originated from NMR. Unfortunately, these MOVs are computed from an error spectrum, which is derived from the linear spectra at the output of the FFT. Therefore, the FFT-based ear model has to be included if the masked threshold concept is incorporated in a combined model. In addition, some model output variables originating from PERCEVAL and OASE turned out to be useful. The most useful model output variable from PERCEVAL is based on the comparison of linear spectra (compare Section 2.3.3) and consists of a cepstrum-like measure that yields information about the harmonic structure of additive distortions.

PEAQ includes two different setups, one called the ”*Basic Version*” (BV) and one called the ”*Advanced Version*” (AV). The advanced version provides the maximum possible accuracy in the prediction of subjectively perceived basic audio quality. The basic version aims for applications, where high computational efficiency is more important than absolute accuracy.

5.2.1 Advanced Version

The advanced version of PEAQ uses the concept of comparing internal representations in combination with a filterbank-based ear model, and the masked threshold concept in combination with the FFT based ear model. The model variables based on the filter bank are measuring the loudness of non-linear distortions, the amount of linear distortions, and distortions of the temporal envelope. The model variables based on the FFT include a noise-to-mask ratio and a cepstrum-like measure for the occurrence and variability of fundamental frequencies in the error signal. A total of five MOVs is used for the prediction of the perceived basic audio quality. Three out of these five MOVs originate from DIX.

5.2.2 Basic Version

The basic version of PEAQ uses the FFT based ear model both with the concept of comparing internal representations and the masked threshold concept. In order to compensate for the somewhat poorer performance of the FFT-based ear model as compared to the filter bank, it uses additional model variables that help to recover some of the information that is lost due to the limited temporal resolution of the FFT. Moreover, it uses a higher spectral resolution, which also helps to recover some of the lost temporal information. Basically, the model output variables used in the basic version cover the same range of perception cues as the model output variables used in the advanced version but for some aspects of the distortion it uses more than one model output variable. A total of eleven MOVs is used for the prediction of the perceived basic audio quality. Four out of these eleven MOVs originate from DIX.

5.3 Description of the Combined Model

The following description of PEAQ is not intended as a guidance for re-implementing this method, but is only supposed to explain its general structure and features. An implementational description is given in the second annex of the ITU-R recommendation [ITU98].

5.3.1 Outline

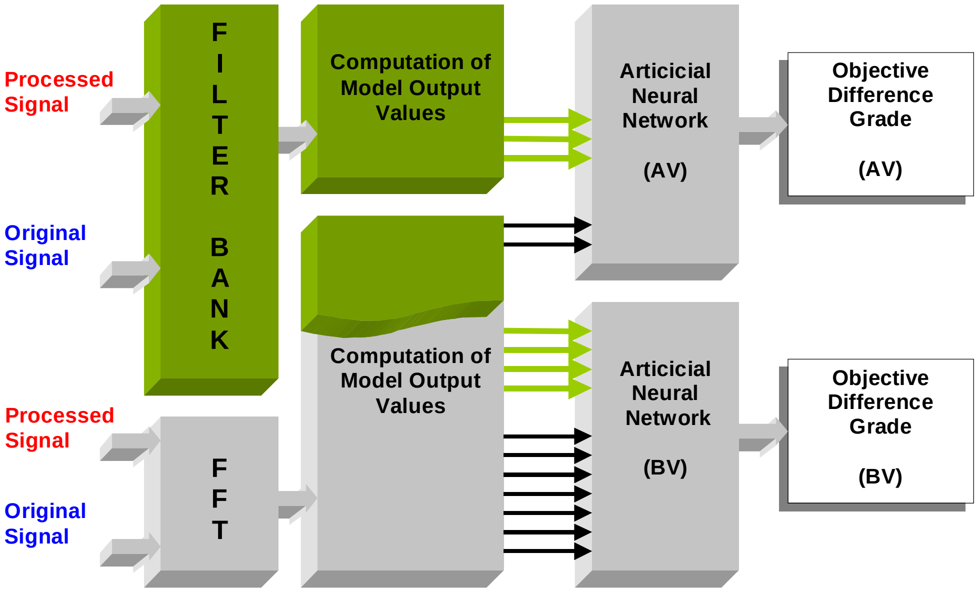

Processed Signal
Original Signal
Processed Signal
Original Signal
FILTER BANK
FFT
Computation of Model Output Values
Computation of Model Output Values
Articicial Neural Network (AV)
Articicial Neural Network (BV)
Objective Difference Grade (AV)
Objective Difference Grade (BV)

Fig. 5.1: Generic block diagram of the measurement scheme (the green areas and arrows indicate the parts originating from DIX).

The global structure of PEAQ can be summarised into a peripheral ear model, several intermediate steps (here referred as “pre-processing of excitation patterns”), the calculation of (mostly) psychoacoustically based model output variables (“MOVs”) and a mapping from a set of model output variables to a single value, representing the basic audio quality of the processed signal. It includes two peripheral ear models, one based on a FFT and one based on a filter bank. Except for the calculation of the error signal (which is only used with the FFT-based part of the ear model) the general structure is the same for both peripheral ear models.

The inputs for the MOV calculation are:

- The excitation patterns for both processed and original signal.
- The spectrally adapted excitation patterns for both processed and original signal.
- The specific loudness patterns for both processed and original signal.
- The modulation patterns for both processed and original signal.
- The error signal calculated as the spectral difference between processed and original signal (only for the FFT-based ear model).

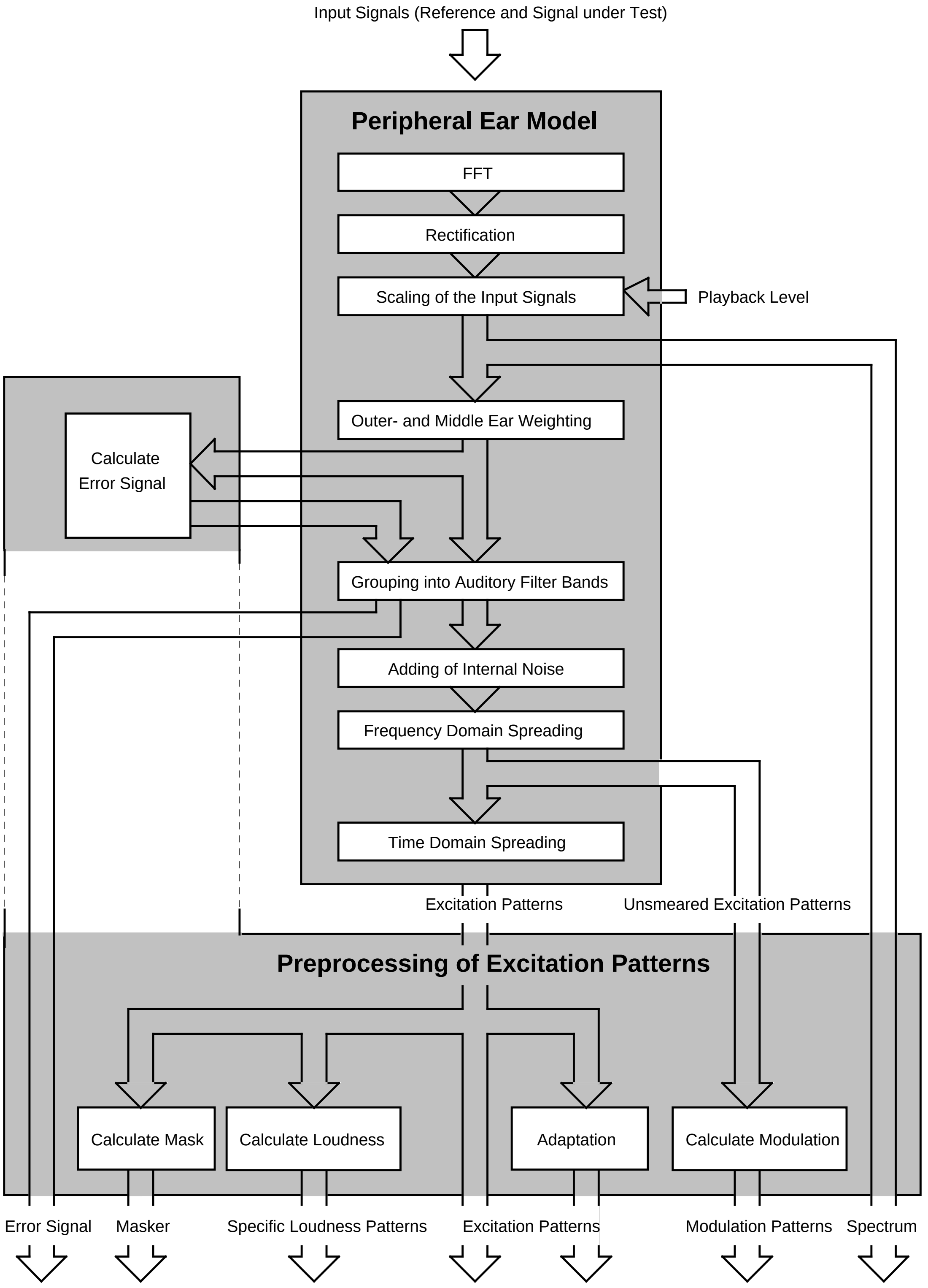


Fig. 5.2: Peripheral ear model and pre-processing of excitation patterns for the FFT-based part of the model.

In the case of stereo signals all computations are performed independently and in the same manner for the left and right channel.

In all given equations, the subscript ”*Ref*” denotes all patterns calculated from the original signal, the subscript ”*Test*” denotes all patterns calculated from the processed signal. The index ”*k*” denotes the discrete frequency variable (i. e. the frequency band), and ”*n*” denotes the discrete time variable (i. e. either the frame counter or the sample counter). If the values for *k* or *n* are not explicitly defined, the computations are to be carried out for all possible values of *k* and *n*. All other abbreviations are explained at the place they occur.

In the names of the model output variables, the index ”*A*” denotes the variables that are calculated from the filterbank-based part of the ear model and the index ”*B*” denotes the variables that are calculated from the FFT-based part of the ear model.

5.3.2 FFT-Based Ear Model

The FFT-based part of the ear model was mainly implemented by FhG and is thus not part of this work. However, for the understanding of PEAQ, a short description of the FFT-based part of the model is given here.

a) Overview

The FFT based ear model processes the input signals in frames of 2048 samples (about 0.042 seconds) with an overlap of 50%. Each frame is scaled to the playback level and transformed to the frequency domain using a short term FFT with a Hann window. A weighting function is applied to the spectral coefficients, which models the outer and middle ear frequency response. The transformation to the pitch representation is done by grouping the weighted spectral coefficients into critical bands. A frequency dependent offset is added to simulate internal noise in the auditory system. A level-dependent spreading function is used to model auditory filter shapes. Forward masking is taken into account by a time domain spreading.

From the obtained excitation patterns, specific loudness patterns and masking patterns are calculated. From the patterns before the final time domain spreading (”unsmeared excitation patterns”), modulation patterns are calculated. The error patterns are calculated from the difference between the spectral patterns of processed and original signal at the output of the outer and middle ear filter. Like the excitation patterns, also the error patterns are mapped to the pitch scale by grouping adjacent spectral coefficients into critical bands.

b) FFT

The mapping from the time domain to the frequency domain is done by applying a Hann window

$$h_w(k) = \frac{1}{2}\sqrt{\frac{8}{3}}\left[1-\cos\left(2\pi\frac{k}{N-1}\right)\right] \quad \Bigg| \quad N = 2048 \qquad (5.1)$$

to the input signals and decomposing them into linear spectra by a short term Fourier transform. A scaling factor is calculated from the assumed sound pressure level *Lp* of a full scale sine wave by

$$fac = \frac{10^{\frac{Lp}{20}}}{norm}, \tag{5.2}$$

where the normalisation factor *norm* is the maximum absolute value of the spectral coefficients over 10 frames when the input signal is a sine wave of 1019.5 Hz and 0 dB full scale. If the sound pressure level is unknown it is recommended to set *Lp* to 92 dB SPL.

c) Threshold in Quiet

The threshold in quiet is modelled by a middle ear transfer function and an internal noise function. Both functions are modelled by the same approximations that are also used in DIX. The only difference is in the upper frequency roll-off, which has been shifted towards higher frequencies by changing the exponent of the last term in Eq. (2.47) from 4.0 to 3.6. As for normal hearing listeners this shift of the roll-off appears to be more realistic than the original approximation, this exponent has also been changed in the filterbank-based part of the model. In both, filterbank-based and FFT-based model of PEAQ, the middle ear transfer function is modelled by a weighting function that is applied after the time-frequency decomposition instead of the filter that has been used in DIX.

d) Grouping into Critical Bands

The simplified approximation of the critical band scale given in [SCH79], which worked best within the perceptual model of DIX, was also taken for the FFT-based part of PEAQ. The frequency borders of the filters range from 80 Hz to 18 000 Hz. The widths and spacing of the filter bands correspond to a resolution of *0.25* Bark for the basic version and *0.5* Bark for the advanced version. This corresponds to a number of *109* non-overlapping frequency bands for the basic version and *55* for the advanced version. The frequency to pitch mapping is performed by a summation of the energies of the spectral coefficients within each frequency band. If the centre frequency of a spectral coefficient is at the border between two frequency bands, its energy is distributed between both bands.

e) Frequency Domain Spreading

The exponential slopes of auditory filters are modelled by a frequency domain convolution with a level-dependent spreading function. The slopes of the spreading function are calculated in the same way as in DIX, except that the lower slope rate is set to 27 dB/Bark, whereas DIX modelled a slightly steeper slope. A temporal smoothing of the slope rates was not necessary because of the lower temporal resolution of the FFT-based ear model. In addition, the additivity of masking is modelled by compressing the energies within each filter band before the spreading, and decompressing the resulting values after the spreading. The compression is carried out by a power function with an exponent of 0.4. Moreover, energy preservation is forced by normalising the resulting excitation patterns by the energy of the spreading function.

f) Time Domain Spreading

Forward masking is taken into account by a first order low pass that is applied on the excitation patterns. The frequency dependence of the time constants was taken from DIX, but the time constant that determines forward masking in the lower frequency bands was reduced from 50 milliseconds to 30 milliseconds.

g) Masking Threshold

The masking threshold is calculated by weighting the excitation patterns with a frequency dependent threshold factor, *m(k)*, which approximately is proportional to the width of the critical bands.

$$m(k) = \begin{cases} 10^{-0.3} & \mid \mathrm{k} \le 12 \\ 10^{-0.025 \cdot k} & \mid \mathrm{k} > 12 \end{cases} \tag{5.3}$$

h) Calculation of the Error Signal

The error signal is calculated in the frequency domain by taking the absolute difference between the power spectra of processed and original signal after the outer and middle ear weighting. The error signal is grouped into frequency bands in the same manner as processed and original signal, but no time or frequency domain spreading is applied.

5.3.3 Filterbank-based Ear Model

The filterbank-based part of the ear model of PEAQ is identical to the peripheral ear model of DIX, which was described in Section 4.

5.3.4 Model Output Variables

The final version of PEAQ includes model output variables originating from DIX, NMR, OASE, and PERCEVAL. Model output variables from POM and "Toolbox" have also been implemented in PEAQ, but do not contribute to the estimation of perceived basic audio quality. Unfortunately, model output variables from PAQM have not been implemented in PEAQ. An overview on the model output variables is given in Table 5.1.

From DIX, the Partial Loudness of Additive Distortions (see Section 4.6.1), Partial Loudness of Linear Distortions (see Section 4.6.2), and Modulation Difference (see Section 4.6.3) are used in the final estimation of the perceived basic audio quality.

Model Output Variable (MOV)	**calculated in ... ear model**		**used in ... version**	
	FFT	**filter bank**	**basic**	**advanced**
WinModDiff1(B)	yes	no	yes	no
AvgModDiff1(B)	yes	no	yes	no
AvgModDiff2(B)	yes	no	yes	no
RmsModDiff(A)	no	yes	no	yes
RmsNoiseLoud(B)	yes	no	yes	no
RmsNoiseLoudAsym(A)	no	yes	no	yes
AvgLinDist(A)	no	yes	no	yes
BandwidthRef(B)	yes	no	yes	no
BandwidthTest(B)	yes	no	yes	no
Total NMR(B)	yes	no	yes	no
RelDistFrames(B)	yes	no	yes	no
Segmental NMR(B)	yes	no	no	yes
MFPD(B)	yes	no	yes	yes
ADB(B)	yes	no	yes	no
EHV(B)	yes	no	yes	yes

Tab. 5.1: Allocation of the model output variables to versions and ear models.

a) Partial Loudness

The settings of the free parameters for the measures based on partial loudness are given in Table 5.2. The calculation of these measures is performed as described in Sections 4.6.1 and 4.6.2. *AvgNoiseLoud(B)* is the linear average of the partial loudness of additive distortions and *RmsNoiseLoud(B)* is the squared average of the partial loudness of additive distortions, both calculated from the FFT-based ear model. All other partial loudness measures are calculated from the filterbank-based ear model. *AvgLinDist(A)* is the linear average of the partial loudness of linear distortions. *RmsNoiseLoud(A)* is the squared average of the partial loudness of additive distortions. For this measure, all momentary values of the partial noise loudness that are below a threshold of 0.1 sone are set to zero. *RmsMissingComponents(A)* is the squared average of the partial loudness of missing frequency components. *RmsNoiseLoudAsym(A)* is the weighted average between RmsNoiseLoud(A) and RmsMissingComponents(A), where the weight of RmsMissingComponents(A) is *0.5* and the weight of RmsNoiseLoud(A) is *1.0*. For all measures based on partial loudness, temporal averaging starts 50 milliseconds after the loudness of the processed signal has reached a threshold of 0.1 sone for the first time.

MOV (Xxx=Win/Avg/Rms)	α	m_s	c_s
XxxNoiseLoud(B)	1.5	0.15	0.5
XxxMissingComponents(A)	1.5	0.15	1
XxxNoiseLoud(A)	2.5	0.3	1
XxxLinDist(A)	1.5	0.15	1

Tab. 5.2: Setting of the free parameters for the partial noise loudness. The parameters are explained in Section 4.6.1.

b) Modulation Difference

The settings of the free parameters for the measures based on the modulation difference are given in Table 5.3. The calculation of these measures is performed as described in Section 4.6.3. *WinModDiff1(B)* is the windowed average of the modulation difference, *AvgModDiff1(B)* is the linear average of the modulation difference, and *AvgModDiff2(B)* is the linear average of the modulation difference with emphasis on introduced modulations and modulation changes where the reference contains little or no modulations. They are calculated from the FFT-based ear model. *RmsModDiff(A)* is the squared average of the modulation difference, calculated from the filterbank-based ear model.

MOV (Xxx=Win/Avg/Rms)	negWt	offset	levWt
XxxModDiff1(B)	1	1	100
XxxModDiff2(B)	0.1	0.01	100
XxxModDiff(A)	1	1	1

Tab. 5.3: Setting of the free parameters for the modulation difference. The parameters are explained in Section 4.6.3.

c) Bandwidth

These model output values were contributed from FhG and estimate the mean bandwidth of the original and processed signal in FFT lines. Local bandwidths are calculated for each frame by searching from both sides of the spectrum for the first line in which the energy exceeds a certain threshold. The local bandwidths are linearly averaged over time. A frame is only taken into account if the bandwidth of the original signal is larger than 346 lines, which corresponds to 8 kHz. Frames with low energy at the beginning and the end of the items are ignored (see Section 5.3.5 b)).

d) Noise-to-mask ratio

These model output values were contributed from FhG and are calculated from error signal and masking threshold. For all NMR-based model output variables, frames with low energy at the beginning and the end of the items are ignored (see Section 5.3.5 b)).

- **Total NMR**

The *Total NMR* is the linear average of the noise-to-mask ratio using

$$NMR_{tot} = 10 \cdot \log_{10}\left[\frac{1}{N} \cdot \sum_{n=0}^{N-1}\left(\frac{1}{Z} \cdot \sum_{k=0}^{Z-1}\frac{ErrorSignal(k,n)}{Mask(k,n)}\right)\right] \tag{5.4}$$

- **Segmental NMR**

The *Segmental NMR* is the logarithmic average of the noise-to-mask ratio using

$$NMR_{seg} = \frac{10}{N} \cdot \sum_{n=0}^{N-1}\log_{10}\left[\frac{1}{Z} \cdot \sum_{k=0}^{Z-1}\frac{ErrorSignal(k,n)}{Mask(k,n)}\right] \tag{5.5}$$

- **Relative Disturbed Frames**

The *Relative Disturbed Frames* (abbreviation: RelDistFrames) represent the number of frames where the noise-to-mask ratio

$$10 \cdot \log_{10}\left(\frac{ErrorSignal(k,n)}{Mask(k,n)}\right) \tag{5.6}$$

exceeds a value of -1.5 dB in at least one filter band, divided by the total number of frames of the item.

e) Detection Probability

The measures based on a detection probability originate from OASE and are implemented by FhG and FAU. The combination of left and right channel is carried out by taking the maximum value within each filter band and frame. The detection probability is derived by comparing the difference between the logarithmic excitations of processed and original signal, *e(k, n),* to the just noticeable difference (JND). The JND depends on the signal energies, and is approximated by a polynomial, which is fitted to data measured by Zwicker [ZWI90]. From these values, a local detection probability is calculated by

$$\mathrm{p}(\mathrm{k},\mathrm{n}) = 1 - 10^{-[\mathrm{a}(\mathrm{k},\mathrm{n}) \cdot \mathrm{e}(\mathrm{k},\mathrm{n})]^{\mathrm{b}}}, \tag{5.7}$$

where *b* determines the steepness of the detection function, and *a(k, n)* scales the detection probability to a value of *0.5* at the JND:

$$\mathrm{a}(\mathrm{k},\mathrm{n}) = \frac{10^{\frac{\log_{10}(\log_{10}(2.0))}{\mathrm{b}}}}{JND(\mathrm{k},\mathrm{n})}. \tag{5.8}$$

The momentary detection probability *P(n)* is calculated according to

$$\mathrm{P}(\mathrm{n}) = 1 - \prod_{\forall \mathrm{k}} [1 - \mathrm{p}(\mathrm{k},\mathrm{n})] \ . \tag{5.9}$$

In addition, the distance of the error to the threshold is expressed in multiples of the JND within each filter band, and added up over all filter bands:

$$\mathrm{Q}(\mathrm{n}) = \sum_{\forall \mathrm{k}} \frac{|\mathrm{e}(\mathrm{k},\, \mathrm{n})|}{JND(\mathrm{k},\, \mathrm{n})} \ . \tag{5.10}$$

- **Maximum filtered probability of detection (MFPD)**

The maximum filtered probability of detection is calculated by smoothing the detection probability over time by a first order low pass with a time constant of approximately 0.2 seconds. The temporal average is calculated continuously, and includes a non-linear operation: if the momentary value of the detection probability is smaller than the current value of its average, the current average value is preserved.

- **Average distorted block[8] (ADB)**

The average error of distorted frames is calculated by averaging the distance of the error to the threshold, *Q(n)*, over all frames where the detection probability exceeds a value of 0.5:

$$ADB = \log_{10}\left[\frac{\sum\limits_{\forall n} Q(n)}{\sum\limits_{\forall n} \begin{cases} 1 & | \ P(n) \geq 0.5 \\ 0 & | \ P(n) < 0.5 \end{cases}}\right] . \tag{5.11}$$

If the number of distorted frames is zero, ADB is set to zero.

f) Harmonic Structure of Error (EHV)

This model output variable was contributed from CRC, and is calculated from the difference between the logarithmic spectra of processed and original signal. Frames with low energy at the beginning and the end of the items are ignored (see Section 5.3.5 a) and b)). It estimates the temporal variation of the harmonic structure of the distortions over time. The magnitude of a harmonic complex is derived from a cepstrum-like measure by computing the difference between the logarithmic spectra of processed and original signal and calculating the spectrum of the autocorrelation function of this difference. The largest peak of this spectrum, except for the DC-component, defines the most prominent harmonic complex The average of its magnitude over time is the model output value *EHS (Error Harmonic Structure)*, and the standard deviation is the model output value *EHV (Error Harmonic Variation)*.

---

[8] The term "block" is equivalent to "frame" in this context.

5.3.5 Temporal and Spectral Averaging

a) Energy threshold

For the model output variable EHV, an energy threshold is applied. When the energy of the most recent half of a frame of 2048 points is less than 8000, the frame is ignored. Frames have a 50 percent overlap and only the half of the frame containing new data is evaluated. Application of this criterion prevents processing of frames with very little energy.

b) Data boundary

The model output variables EHV, BW, NMR, and RelDistFrames use a threshold criterion to avoid invalid frames in the beginning and in the end of a test excerpt. The beginning or end of data is defined as the first location, scanning from the start or end of the file, where the sum of the absolute values over five succeeding samples exceeds 200. Frames are subsequently ignored if the data is read from outside of this range.

5.4 Optimisation of the Model

When using features of DIX within the context of the combined model PEAQ, the optimum settings for some parts of the model differ from the original settings.

5.4.1 Forward Masking

For the purely filterbank-based version of PEAQ, the time constants used in the modelling of forward masking remained the same as in DIX. For the model version that combines filterbank-based and FFT-based model output values, the accuracy of the model could be slightly increased when using shorter time constants. The final model used 4 milliseconds for the intermediate and high frequency range, and around 20 milliseconds for the low frequency range. The performance of the filterbank-based model output values slightly dropped when using these settings, but in combination with model output values from the FFT-based part, the overall performance increased.

5.4.2 Dynamic Level and Pattern Alignment

As the structure of the FFT-based ear model does not allow to apply the level and pattern adaptation the way it was used in DIX, the adaptation was applied to the final excitation patterns instead of the filter bank outputs. With this structure, best results were achieved when lowering the time constants used in the adaptation. The final model used 8 milliseconds for the intermediate and high frequency range, and around 50 milliseconds for the low frequency range.

6. Estimation of Perceived Basic Audio Quality

The output values of a perceptual measurement method normally reflect certain aspects of the perceived distortion, but are in general not identical to the aspects of the distortion that are asked in subjective listening tests. In listening tests carried out for the comparison of audio codecs, the listeners are asked to give a rating for the *basic audio quality* of the complete test item. As different kinds of perceived errors contribute to the basic audio quality, it cannot be used to validate a perceptual model directly. Instead, the outputs of the perceptual model have to be mapped to a single value that corresponds to basic audio quality, and the model output values can only be validated together with the mapping functions.

6.1 Grading Scales for Subjective Audio Quality Assessment

6.1.1 The Five-Grade Impairment Scale

A test procedure for the subjective assessment of high quality audio codecs is defined in the ITU-R recommendation BS.1116 [ITU96]. The listening tests have to be carried out with the “hidden reference / double blind / triple stimulus” method. In this test method the listeners can switch between the original signal (reference), A, and two other signals, B and C. One of these two signals is the processed signal, and the other one is once again the original signal (hidden reference). Neither the test subject nor the supervisor knows, which of the signals B and C is the hidden reference, and which is the processed signal. The listeners have to decide, which signal is the hidden reference, and judge the overall quality (“basic audio quality”) of the other signal. The basic audio quality is measured on the “five-grade-impairment-scale”. This scale covers a continuous range from one to five, where five anchor points are defined by a verbal description of the perceived quality (Figure 6.1). It looks similar to the MOS-scale used in speech quality assessment, but covers a very different range of distortions and is derived with a different listening test procedure.

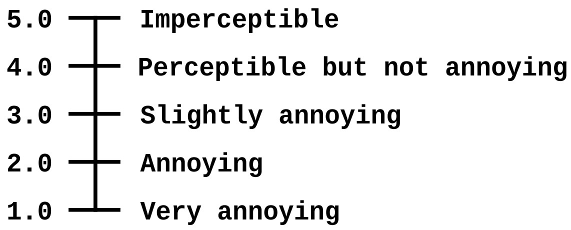


Fig. 6.1: The five-grade impairment scale according to ITU-R BS.1116.

6.1.2 SDG (“Subjective Difference Grade”)

The results of codec comparison tests are normally given in terms of *subjective difference grades* (SDGs). The SDG is the difference between the absolute quality grade assigned to the distorted signal, and the absolute quality grade assigned to the original signal, both measured on the five-grade impairment scale. The quality criterion is the similarity between the assessed signal and the original signal. Therefore, the listeners should always assign a grade of five to the original signal. If all listeners identify the hidden reference correctly, the SDG is, except for an offset of -5, identical to the grade on the five-grade impairment-scale assigned to the processed signal. The value of an SDG should therefore never be outside a range of minus four to zero. Nevertheless, the theoretical range of the SDG-scale is from minus four to four. Positive values do not mean a higher quality, but indicate that the listeners have problems to identify the hidden reference.

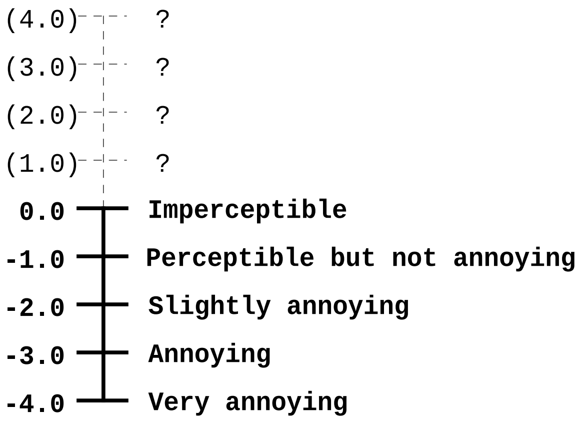


Fig. 6.2: The SDG-scale.

6.1.3 ODG (“Objective Difference Grade”)

If a measurement method yields an estimate of an SDG, this value is located on the same scale as the SDG, but is an objective measure instead of a subjective grading. Such a value is therefore called *objective difference grade* (ODG). The ODG is the only output of a measurement method that can directly be verified against listening test data derived from codec comparison tests according to ITU-R BS.1116.

6.2 Accessible Databases

Most of the data sets that were used for the optimisation and validation of the measurement method originate from codec tests performed by MPEG and by the ITU Radiocommunications study group. Two sets of listening tests have been carried out by the ITU-R task group 10/4 with the explicit task to check the performance of the measurement methods. These tests are referred to as database 2 and database 3. The codec tests that had been carried out before these tests are referred to as database 1. Additionally, listening test data from two listening tests carried out at the CRC have become accessible later on (EIA’95 and CRC’97). A summary of the audio data and types of codecs used in all these tests is given in an ITU-R document [ITU98x]. The subjective tests have been carried out either with loudspeaker or with headphone presentations. In some tests, the listeners were allowed to switch between headphones and loudspeakers. As in loudspeaker tests the results may be influenced by room acoustics and speaker characteristics, the results produced with headphone presentations are preferred if headphone and loudspeaker data are separately available.

6.2.1 Database 1

Database 1 consists of data from five listening tests, three carried out by MPEG (1990, 1991 and 1994) and two carried out by the ITU (1992 and 1993). The test material consists mostly of musical excerpts partly taken from the SQUAM disk, but also taken from normal radio program material or explicitly created for codec tests. Apart from musical excerpts, each listening test included one speech excerpt and the MPEG 1990 test also included one artificial item (bass synthesiser) that can hardly be regarded as a musical item.

6.2.2 Database 2

Database 2 was created within the ITU-R TG 10/4 for the comparison of perceptual measurement methods. Therefore, it includes many items that are especially critical for the measurement methods. It consists of 91 test items. One aim of this test was to check whether perceptual measurement methods designed for codec comparisons are also capable to access artefacts that are not originating from audio codecs. Many of these artefacts caused severe problems for most measurement methods. It also includes one coding artefact that is clearly underestimated by all measurement methods, and points to a basic problem within the psychoacoustical models.

6.2.3 Database 3

Database 3 was created within the ITU-R TG 10/4 for the final validation of the perceptual measurement method to be recommended. It again includes items that are critical for the measurement methods, but consists solely of coding artefacts. This database consists of 84 test items, and the distortions in at least two of the items are clearly underestimated by all measurement methods.

6.2.4 Other Databases

- **EIA Test 1995**

This test was carried out on codecs that were aimed for digital broadcasting. It includes 81 test items, which are mostly located in the upper half of the quality range.

- **CRC Codec-Test 1997**

This data set includes 136 test items. Seventeen codecs have been tested on eight audio excerpts. The confidence intervals of the subjective data are comparably low. The items equally cover the full quality range. This data set has not been accessible during the optimisation of the measurement methods involved in the standardisation process within ITU-R TG 10/4. Therefore, it was used as an additional validation database in the selection of the final model version to be recommended for standardisation. Compared to database 3, it has the advantages of a larger number of test items, lower confidence intervals, and of being a "real world" application.

6.3 Mapping from Model Output Values to Objective Difference Grades

The main problem when searching for the optimum mapping between model output values and listening test results is the decision, whether the mapping function really reflects the correspondence between the MOV and the SDG. With one-dimensional mappings, this can easily be decided by checking whether the mapping function is monotonically decreasing (provided that increasing values of the model output variable indicate an increasing distortion). With multidimensional mappings this is not that easy anymore. In general, the risk of getting an unreliable mapping increases with the number of degrees of freedom of such a mapping. When a mapping has too many degrees of freedom, there is a considerable probability of getting a high correlation between the mapped value and the target (here: the subjective grade) by chance, even if the input values have not really a significant relation to the target values. This can only be avoided by using a sufficiently large set of test data. From neural network training, it is known that the test data set should be at least ten times larger than the number of degrees of freedom.

6.3.1 Polynomial Mapping Functions

When using polynomial mapping functions, the polynomials must be at least of third order to allow for a reasonable mapping. Consequently, a mapping of three model output values to one quality grade has at least ten degrees of freedom. If also mixed polynomials are included (i. e., products between different model output values), the number of degrees of freedom is even larger. Without mixed polynomials, non-linear interactions (like minimum or maximum) between the model output values cannot be modelled.

The risk of ending up with an unreliable mapping also depends on the characteristics of the mapping functions. Polynomials are, in this respect, rather problematic functions, because they can approximate any possible shape. Due to this behaviour, the required test data set must be even larger than in neural network training. In practice, it turned out that multidimensional mappings were almost never reliable when using polynomial mapping functions. However, polynomial mappings have the advantage that the best possible mapping in the sense of minimum squared difference can be explicitly calculated. Therefore, they were used to get a quick overview of the possible performance of single model output variables.

6.3.2 One-Dimensional Mapping Functions

For one-dimensional mappings, third order polynomials normally provide a reasonable mapping. As even ordered polynomials cannot provide a monotonic mapping because of their even symmetry, the next possible alternative would be 5th order polynomials. For the normal size of the test data sets, which include between 32 and 136 test items, such polynomials already exceed the permissible number of degrees of freedom, and mappings by 5th order polynomials are often not monotonic anymore.

The performance and reliability of the mappings can be improved when using mapping functions that are better fitted to the expected relation between model output values and subjective quality gradings. Convenient mapping functions are sigmoid functions

$$sig(x) = \frac{w}{1 + e^{-(mx+c)}} + b \ , \tag{6.1}$$

and rational functions

$$f(x) = \frac{w}{1 + (mx + c)^2} + b \quad \left| \begin{array}{c} m \cdot w > 0 \\ mx + c > 0 \end{array} \right. . \tag{6.2}$$

These functions have the same number of degrees of freedom as third order polynomials, but normally provide a better fit between model values and subjective data. Moreover, they are always monotonic, and hence the risk of getting unreliable mappings is comparably low. The sigmoid function maps a two-sided infinite range [-∞, ... ,+∞] to a limited range [b, ... ,b+w]. The rational function maps a one-sided infinite range [-c/m, ... ,+∞] to a limited range [b, ... ,b+w]. As an error measure normally is limited at one side, the rational function appears to be the more appropriate mapping function. This was corroborated when applying these functions to map the partial loudness of additive distortions to subjective gradings. However, the performances of sigmoids and rational functions differed only marginally. The problem with both functions is that the optimum settings cannot be calculated analytically, but have to be found numerically.

6.3.3 Multidimensional Mapping Functions

A multidimensional mapping function should facilitate the modelling of all possible interactions between different perceptions, but minimise the risk of getting unreliable mappings. A mapping is considered as potentially unreliable, when the relation between any input parameter and the output value is non-monotonic. Possible interactions are a summation or the maximum between internal representations of different error measures. Products between different inputs have not been modelled because this would result in an unacceptably high number of degrees of freedom. A convenient mapping algorithm is given by the following three step approach (Figure 6.3):

- All inputs are mapped to an internal quality representation. The internal quality scale is assumed to be identical for all kinds of perception.
- The average and minimum value among the internal quality representation are calculated.
- A weighted summation between average and minimum quality is carried out, and the result is mapped to the 5-grade impairment scale.

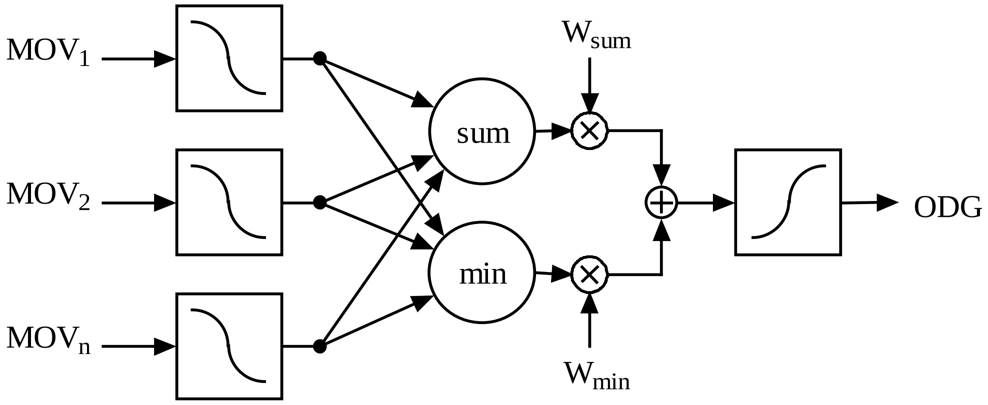


Fig. 6.3: Mapping from multiple MOVs to one ODG.

The mapping functions used in the first step are rational functions, and are restricted to a range where the mapping is monotonic and the slopes are negative (Eq. 6.3).

$$f(x) = \frac{w}{1+(mx+c)^2} \quad \left| \begin{array}{l} m \cdot w > 0 \\ mx + c > 0 \end{array} \right. . \tag{6.3}$$

In the third step, the same functions are used, but are restricted to positive slopes (Eq. 6.4).

$$f(x) = \frac{w}{1+(mx+c)^2} \quad \left| \begin{array}{l} m \cdot w < 0 \\ mx + c > 0 \end{array} \right. \tag{6.4}$$

This mapping algorithm has several advantages as compared to polynomial mappings or mappings using neural networks (see Section 6.4):

- The risk of getting unreliable mappings is minimised.
- In most cases, the number of degrees of freedom is reduced.
- There is always an unambiguous relation between the individual error measures and the resulting quality measure. When using neural networks, the relation between inputs and output value can normally not be expressed in a comprehensible form.
- The mapping algorithm can be realised with a lower computational complexity than the training of a neural network.

Nevertheless, it turned out that neural network mappings often perform better when the number of inputs is larger than four. This is probably explained by the rather simple gradient search this algorithm performs to find the optimum mapping parameters. When the number of inputs is high, the probability to end up in a local minimum increases. Another explanation is the ability of a neural network to model more complex relations among the inputs. The most successful strategy when searching for mappings between different model parameters and subjective quality gradings was to use the above algorithm to determine useful combinations of model output variables, and train a neural network on the derived combinations.

6.4 Application of an Artificial Neural Network

In the combined ITU model, PEAQ, an artificial neural network was used to map the model output variables to an estimate of the perceived basic audio quality. The training algorithm was supplied by the Canadian *Communication Research Centre* (*CRC*), and uses a backward propagation structure. Only network setups with one hidden layer were used.

The activation function of the neural network is an asymmetric sigmoid

$$sig(x) = \frac{1}{1+e^{-x}} \ . \qquad (6.5)$$

The network uses I inputs and J nodes in the hidden layer. The mapping is defined by a set of input scaling factors $a_{min}(i)$, $a_{max}(i)$, a set of input weights $w_x(i)$, a set of output weights $w_y(j)$ and a pair of output scaling factors b_{min} and b_{max}. In the first step, the input values $x(i)$ are scaled to ensure a similar range for all inputs.

$$x_{scaled}(i) = \frac{x(i) - a_{min}(i)}{a_{max}(i) - a_{min}(i)} \qquad (6.6)$$

For the training data set, the scaled input values range from zero to one. In each node of the hidden layer, the scaled input values are weighted and added up. After adding a bias, the result is sent through the activation function.

$$y(j) = sig\left(w_x(I, j) + \sum_{i=0}^{I-1} w_x(i, j) \cdot x_{scaled}(i) \right) \qquad (6.7)$$

In the output layer, the outputs of the hidden layer are weighted and added up. Again, after adding a bias, the result is sent through the activation function.

$$y = sig\left(w_y(J) + \sum_{j=0}^{J-1} \left[w_y(j) \cdot y(j) \right] \right) \qquad (6.8)$$

The scaled output value is the final ODG:

$$ODG = y_{scaled} = b_{min} + (b_{max} - b_{min}) \cdot y \ . \qquad (6.9)$$

The neural network was trained on a merged data set, containing all accessible data bases except database 3 and CRC’97. For the training of the advanced version of PEAQ, items that are regarded as outside the scope of the measurement method were removed from the training data set. This decision was motivated by the assumption that a neural network tends to produce unreliable mappings when the input data does not contain the information that is responsible for the values of the target data. Left out items were the analogue and artificially created distortions from database 2, the clipped items from database 2, and the items containing single clicks or beeps from the MPEG 1991 and the ITU 1992 database. In the training of the basic version of

PEAQ, which was mainly carried out by OPTICOM, no items have been removed from the training data set.

6.5 Definition of a Distortion Index

The ODG is limited to the quality range covered by the five-grade impairment scale. Due to the saturation of the scale, a distinction among items where the basic audio quality is at one end of the scale cannot be made. For some applications, it may be convenient to have a more linear quality measure covering a somewhat extended quality range. Moreover, the ODG-scale depends on the meaning of the anchor points of the five-grade impairment scale. As the meaning of these anchor points is linked to a subjective definition of quality, it may change over time. For this reason, a technical quality measure should preferably not be expressed as a difference grade, but by a more abstract unit, which maps monotonically to ODGs. In case the anchors of the ODG-scale will change, this measure remains the same, and only the mapping to ODGs has to be adjusted.

A convenient way to derive such a measure is to use the input of the final non-linearity of the output layer of the neural network. At this point, all model output variables are already combined into a single value, but the final clipping to the range of the SDG-scale has not yet taken place. This value is called the *distortion index*, and is defined by

$$DI = w_y[J] + \sum_{j=0}^{J-1}\left(w_y[j] \cdot sig\left(w_x[I,j] + \sum_{i=0}^{I-1} w_x[i,j] \cdot \frac{x[i] - a_{min}[i]}{a_{max}[i] - a_{min}[i]} \right)\right), \quad (6.10)$$

where the inputs and weights are taken from the neural network trained. The relation between distortion index and objective difference grade is given by

$$ODG = b_{min} + (b_{max} - b_{min}) \cdot sig(DI). \quad (6.11)$$

7. Performance of the Measurement Method

The performance of the measurement method was tested by checking the ability of the model to predict the results of subjective listening tests. As the typical application of the model is the quality assessment of perceptual audio codecs, the most relevant test material are audio data used in codec tests together with the subjective gradings derived from these tests. A reasonably good performance in the prediction of the results of psychoacoustical experiments can be expected due to the fact that the model assumptions on which the method is based are derived from such data.

If, however, the method would have a poor performance when simulating psychoacoustical experiments, but still be able to yield reliable predictions on audio codec tests, the latter advantage would supersede this shortcoming. This preference of codec tests instead of psychoacoustical experiments is justified by several considerations:

- Audio codec tests represent the typical application the model is designed for.
- In general, audio codec tests are carried out with a high number of test listeners and an extensive checking of the reliability of the test subjects, whereas commonly used psychoacoustical data is often derived from a very small group of listeners.
- Compared to psychoacoustical experiments, the range of test data is much larger in codec tests, and the signals are in general more complex.

Nevertheless, it has to be noted that certain kinds of degradations might still be missing in codec tests or might be highly correlated with other artefacts, and consequently deficiencies of a measurement method when detecting such artefacts might not be recognised. One example is the band-limitation introduced by some codecs. In older codecs band-limitations were always highly correlated with other artefacts, and needed not to be explicitly measured. In a later listening test, band-limitations occurred independently of other artefacts, and the measurement method had to be changed.

The verification results presented in this section are divided into three parts. The first part shows results of DIX during its development. The second part shows the results of DIX and the competing perceptual measurement methods in a comparative test carried out by the ITU-R task group 10/4. The third part shows results of different versions of PEAQ in the final ITU-R validation test.

7.1 Applied Criteria

7.1.1 Correlations between Model Predictions and Subjective Gradings

The most widely known similarity measure is the cross correlation coefficient. Even though it might not be the best possible criterion, it is easy to handle, as it summarises the correspondence between model predictions and subjective gradings into one single value. The absolute value of the correlation coefficient depends strongly on the distribution of the subjective gradings. Therefore, if one measurement method yields a higher correlation on one data set than on another one, this does not necessarily mean, that it performs better on the first data set than on the second one. In general, a given accuracy of model predictions yields higher correlation coefficients for data sets that cover the full quality range than on data sets that only cover a part of the quality range, and higher correlations on data sets with most items at the upper and lower border of the quality range than on data sets with equally distributed quality. Nevertheless, the correlation coefficient gives a realistic impression of the relative performance of different models on the same data set.

7.1.2 Average Prediction Error

The average prediction error is the mean squared difference between model predictions and subjective gradings after re-scaling the model predictions by a linear regression. It is a direct measure of the relative accuracy of the model and depends only slightly on the quality range of the test items (as long as "end-of-scale effects" are not dominating the test results). For this reason, it is a more convenient measure than the cross correlation coefficient when comparing results that were achieved with different data sets. When looking at the same data set, there is an unambiguous monotonic relation between correlation and average prediction error.

7.1.3 Absolute Error Scores

For the first comparative test within the ITU-R TG 10/4, another performance criterion has been defined that is similar to the average prediction error, but takes the subjective confidence intervals into account. The absolute error score (AES) has been defined as

$$AES = 2 \cdot \sqrt{\frac{1}{N} \cdot \sum_{n=0}^{N-1} \left(\frac{ODG(n) - SDG(n)}{\max[CI(n), 0.25]} \right)^2}\,, \tag{7.1}$$

where N is the number of test items, and $CI(n)$ is the confidence interval of the SDG. The intention of this performance criterion was to assign a lower weight to test items where the subjective gradings are inconsistent. This should take the fact into account that an estimate of a listening test result cannot be expected to be more accurate than the listening test result itself. The AES depends strongly on the minimum value to which the confidence intervals are limited. If this value is too low, test items with a very low confidence interval will mainly determine the value of the AES. The most reasonable value for this minimum confidence interval would be the expected accuracy of the final measurement method. From previous results, the expected accuracy of an ODG is in a range of +/- 0.5. In contrast, the clipping value used was 0.25 which corresponds to a range of +/- 0.125, which is four times smaller than would be sensible when related to the expected accuracy.

7.1.4 Tolerance Scheme

Another possible performance criterion is the application of a predefined tolerance scheme on the model predictions. Within ITU-R TG 10/4 a tolerance scheme was defined as follows :

- If the SDG is above -1.5, the width of the tolerance range is given by the confidence interval of the subjective gradings.
- If the SDG is below -2.5, the width of the tolerance range is twice the confidence interval of the subjective gradings.
- For SDGs between -1.5 and -2.5, the width of the tolerance range relative to the subjective confidence intervals is interpolated between the values given above.
- If the SDG is smaller than -1.9, the tolerance range on the negative side goes down to the end of the scale.

The resulting tolerance scheme is depicted in Figure 7.1.

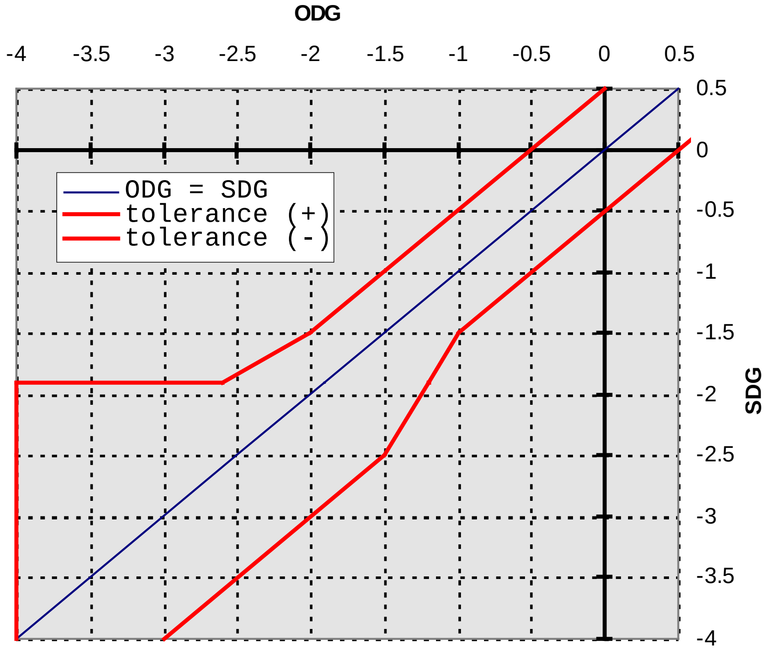


Fig. 7.1: Tolerance scheme proposed for the final ITU validation test (for an assumed confidence interval of +/-0.5 for all test items)

If for any model all predictions are within this tolerance scheme, it is considered as sufficiently accurate in order to be recommended for standardisation.

7.2 Predictions of Comparative Tests among Audio Codecs

During the development of DIX, it has been checked several times against the accessible test data sets. Mostly, the MPEG 1990 test and the ITU 1993 test were used because they provide low subjective confidence intervals. As the quality of the included test items does not change significantly over time, the test items were truncated to a length of six seconds to save computation time.

Figures 7.2 to 7.5 show results of DIX on the MPEG 1990 and the ITU 1993 data set. The subjective difference grades are plotted over the logarithm of the partial loudness of additive distortions ("partial noise loudness"). A mapping between partial noise loudness and subjective difference grades was performed using a sigmoid function.

The linear correlation and standard error between the resulting ODG and the subjective difference grades are also given in the diagrams. The presented results are similar to the results obtained with PAQM [BEE94], and are clearly better than the published results of all other measurement methods at that time. The last diagram shows that the mapping is almost the same among different data sets so that the data sets can be merged.

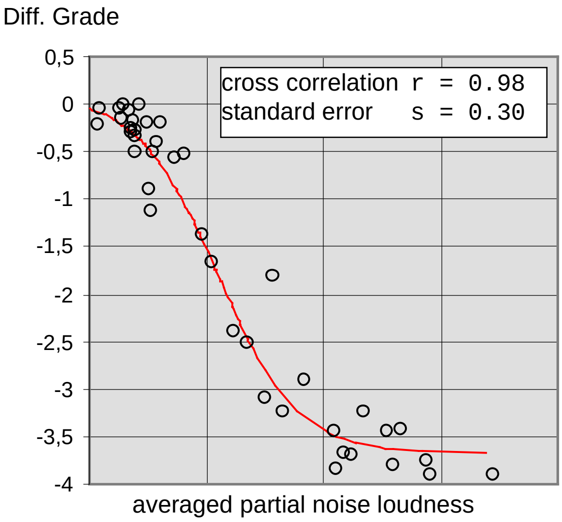


Fig. 7.2: Results for 40 items of the ISO/MPEG 1990 test (headphone presentation) [MPEG90]

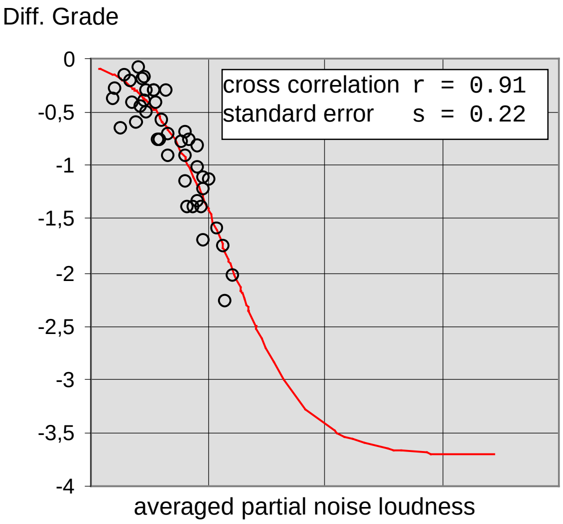


Fig. 7.3: Results for 42 items (cascaded codecs) of the ITU 1993 test [ITU93].

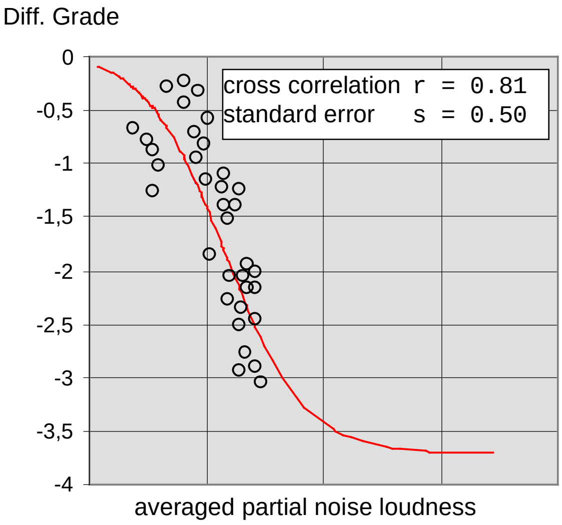


Fig. 7.4: Results for 35 items (commentary stereo) of the ITU 1993 test[ITU93]

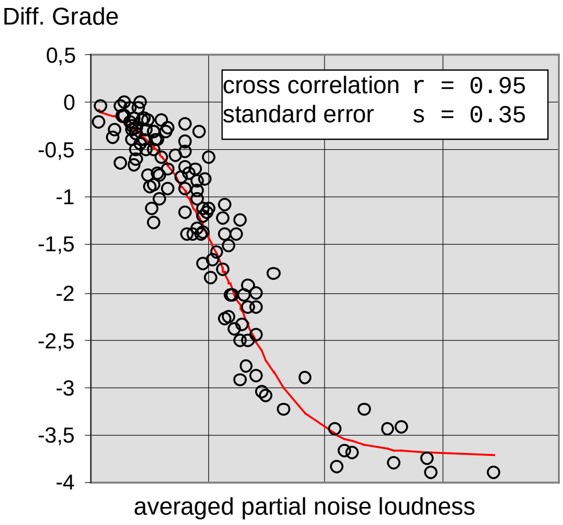


Fig. 7.5: Results for all three tests put together

7.3 ITU-R Comparative Test 1996

In the beginning of 1996, a comparative test was carried out to compare the performance of the six measurement methods that took part in the standardisation efforts of ITU-R TG 10/4. For each of the six measurement methods, up to three versions could take part in the test. A completely new set of 91 test items was assembled for this comparative test (database 2), and listening tests were carried out after the models delivered their predictions of the subjective quality of the test items. In addition, model predictions for a subset of 47 already known test items had to be delivered. As the new data set might contain artefacts that never occurred in the previous data sets, the model proponents were allowed to adjust their models after the listening test results had been published. For this purpose, fifty percent of the data set were made available to the proponents (database 2, first part), and the other half was kept secret to validate the adjusted models after the optimisation phase (database 2, second part). The subset of the 47 old test items was used to check whether the adjustment of the models adversely affected the performance of the models on previous data. The latter case would be an indicator for overfitting, and should be avoided. The whole procedure is divided into three phases. Running the models on the completely unknown data set is the first phase, training the models on half of the data set is the second phase, and running the trained models on the remaining part of the data set is called the third phase.

7.3.1 Results of the First Phase

The performance of all six incorporated models was unexpectedly poor in the first phase of the ITU-R comparative test. The correlation coefficient between model predictions and SDGs never significantly exceeded a value of 0.7.

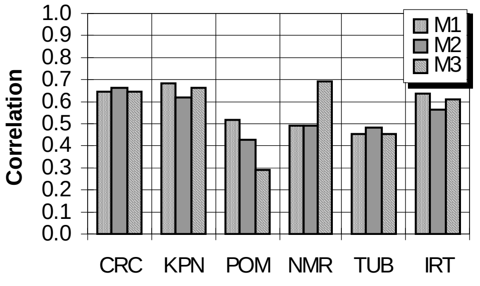


Fig. 7.6: Correlations between model predictions and SDGs for database 2 (91 items).[9]

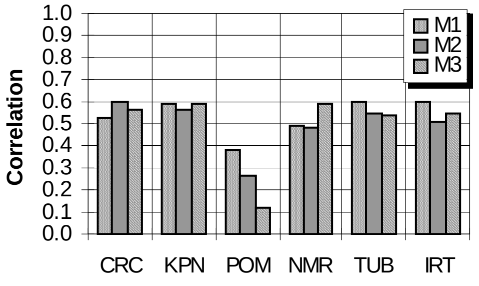


Fig. 7.7: Correlations between model predictions and SDGs for database 2 without the band-limited items (83 items).

As the new database included eight band-limited items, DIX had an additional problem because such linear distortions were separated by the pattern adaptation algorithm. The previously accessible data sets did not include any items where the basic audio quality was determined by linear distortions. Therefore, no mapping was

---

[9] In order to get short, but still meaningful abbreviations, most models are denoted by the name of the organisation instead of the name of the model: CRC ≅ PERCEVAL, KPN ≅ PAQM, TUB ≅ DIX, IRT ≅ "Toolbox". OASE is included as model 3 of NMR.

established between the indicators for linear distortions and SDGs. Consequently, the band-limited items were judged as transparent by DIX (which proved that the separation between linear and non-linear distortions worked perfectly). As the band-limited items were graded very low in the listening test, this shortcoming of DIX caused the correlations between model predictions and subjective grades to become very low (Figure 7.6). Without the band-limited items, the predictions of DIX show a correlation of 0.6 with subjective gradings, which still is very low. However, the other measurement methods performed even worse on the remaining data (Figure 7.7).

The results for the subset of database 1 were clearly better, even though it includes several problematic items. One of the codecs included in these test items introduced a rather large amount of linear distortions, and the audibility of the distortions strongly depended on the chosen playback level. Moreover, one of the included test items (bass guitar, coded with NICAM) has a single click in the middle of the item, but was still judged as transparent in the subjective listening test. When looking at the audio file, it is obvious that this click is caused by an integer overflow. Therefore, it is very likely that the click was introduced when transferring the test item to the 16 bit PCM format used as the input for the measurement methods. As this item apparently is corrupted, it should not be included in the test data set. Other measurement methods had no problems with this items because they were due to their restricted temporal resolution not able to detect this click anyway. Figure 7.8 shows the results of DIX on the subset of database 1, and indicates the problem items.

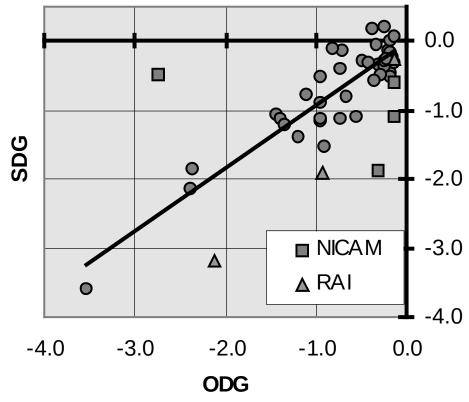


Fig. 7.8: Results of version-3 of DIX for the subset of database 1.

Even when the corrupted item is included, DIX yields the third highest correlation on the subset of database 1 (Figure 7.9). Only PAQM and NMR performed better. Without this item, only PAQM achieves better results than DIX (Figure 7.10).

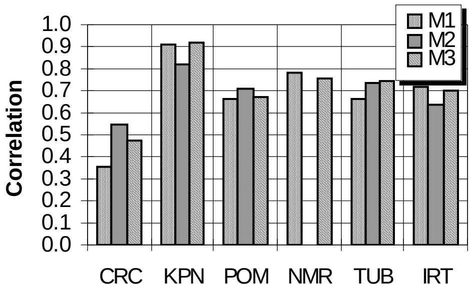


Fig. 7.9: Results for the subset of database 1.

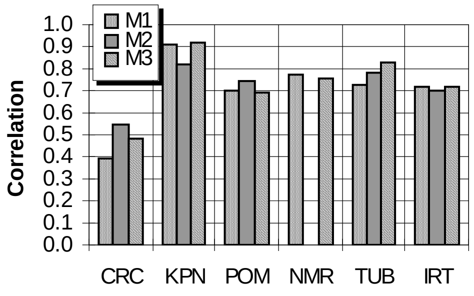


Fig. 7.10: Results for the subset of database 1, without the NICAM bass guitar item, which is probably corrupted.

7.3.2 Second Phase: Fitting the Models to the Database

During the adjustment of the measurement methods to the new database, only few changes have been made to DIX. The problem with the band-limited items could easily be solved by including the partial loudness of linear distortions in the ODG calculation. The problem with the click in the bass guitar item in the subset of database 1 was solved by comparing the linear average of the partial loudness of additive distortions to its square root average. When the ratio between both exceeds a certain threshold, the measured distortions are assumed to result from a single short event, and are suppressed by using the square root average instead of the linear average. This procedure is not really appropriate, but was the only way to handle this corrupt test item without affecting the performance of the model on the remaining data. Furthermore, the time constants for temporal masking in the lower filter bands were reduced, and the averaging between left and right audio channel was slightly modified by replacing the average of both channels by an interpolation between average and maximum. Both changes yielded a slight improvement on the first part of database 2 without affecting the results on the older data sets.

7.3.3 Results of the Third Phase

The training considerably improved the performance of most measurement methods on the first half of database 2 (Figure 7.11). Most of the models show very similar results on the first half of database 3, but only PAQM improved its performance also on the second half (Figure 7.12). When including the results on the database 1 subset in the comparison of the models, DIX performed second best after PAQM (Figures 7.13 and 7.14).

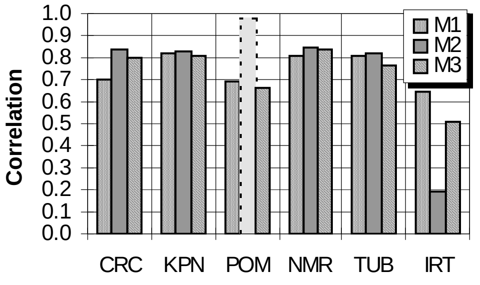


Fig. 7.11: Correlations between model predictions and SDGs for the first half of database 2 (48 items)[10].

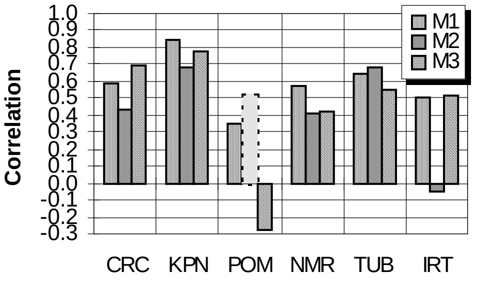


Fig. 7.12: Correlations between model predictions and SDGs for the second half of database 2 (43 items).

7.3.4 Conclusions

Even though some of the models showed an acceptable performance in the third phase of the test, the results of the first phase were indicating that none of the models was yet reliable enough to predict subjective gradings for a completely unknown data set. Therefore, none of the models was recommended for standardisation by the

---

[10] Model 2 of POM appeared to be overfitted and was therefore not considered in the comparison.

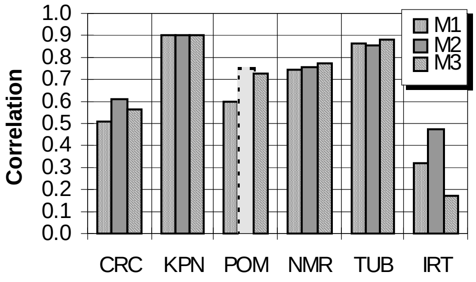


Fig. 7.13: Correlations between model predictions and SDGs for the subset of database 1 (47 items).

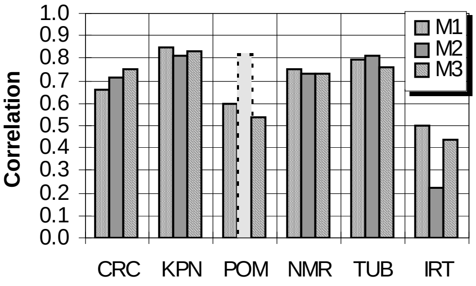


Fig. 7.14: Correlations between model predictions and SDGs for database 2 and the subset of database 1 (138 items).

ITU-R TG 10/4. Instead, the model proponents were urged to establish a joint model that combines features of all original proposals in order to achieve the best possible performance. This lead to the development of PEAQ. As PAQM clearly performed best among the original proposals, it was decided to use it as a reference in the development and validation of PEAQ.

7.4 ITU-R Validation Test 1997

In 1997, the final validation of PEAQ was carried out. As PEAQ includes both filterbank-based and FFT-based features, three different versions of PEAQ were included in the test:

- a purely FFT-based version,
- a purely filterbank-based version,
- a version that combines features of both parts.

As there has not yet been any decision about the use or non-use of neural networks, each version was included twice: once using a neural network, and once not using a neural network. For each version and each mapping strategy, three different mappings were allowed. This resulted in a total of 18 model versions to be tested.

7.4.1 Optimisation and Preselection

During the preparation of the validation test, all model proponents searched for efficient mappings of different model output variables to an estimate for the perceived audio quality. This was not a trivial task because, due to the possible combinations of different parameter settings, different ear models, and different averaging strategies, the total number of available MOVs was increased to more than 200. For this reason, it was not possible to check all MOV combinations for their ability to predict the subjective gradings. With the mapping algorithm described in Section 6.3.3, at least all combinations of two MOVs could be tested. Combinations of more than two MOVs were checked by keeping the 20 best combinations of two, and combining these MOVs with each of the remaining MOVs. When searching for combinations of four, the best 20 combinations of three served as the starting point, and the same was done when searching for combinations of five. Combinations of more than five MOVs were not found by that procedure because the mapping algorithm was not capable to find an optimum mapping when the number of inputs was six or above. This procedure does not guarantee that the best MOV combination is found, but it has a much higher chance of success than a pure random search.

Neural net mappings were searched by either directly using the MOV combinations found by the above procedure, or by combining MOVs that frequently occurred within these MOV combinations. All available data sets were merged together to form the training database. A subset of 20 percent of the test items, showing the same distribution of subjective gradings as the total training set, was taken out of the training database, and was used as validation data during the neural network training.

With the described mapping strategies, 19 potentially successful mappings were found. Most mappings found by other proponents performed clearly worse on the known data sets, and were therefore not considered in the further evaluation. Among the mappings that were further evaluated were six mappings proposed by OPTICOM and one mapping proposed by CRC. All of these mappings were addressed to the purely FFT-based version of PEAQ, and most of them used a higher number of MOVs than the mappings proposed by TUB.

The CRC’97 database, which until that time has not been accessible to the model proponents, was used for an internal validation of the proposed mappings. The mappings that performed best on these data sets were submitted for the “official” validation test. Table 7.1 shows the results of the proposed mappings on this database. The performance on previously known databases was not used as a preselection criterion because, due to the high number of MOVs in some of the proposed mappings, overfitting might have occurred.

ODG	**correlation on the CRC’97 database**
OptOdg0-5H2	0.750
OptOdg1-15H3	0.843
OptOdg2-5H2	0.779
OptOdg3-6H3	0.816
OptOdg4-17H5	0.814
OptOdg5-6H3	0.705
TubFftODG05a	0.842
TubFftODG05b	0.855
TubFftODG03a	0.838
TubNnODG03	0.816
TubNnODG04	0.807
TubNnODG38	0.634
PercOdg	0.764
TubFiltODG06a	0.837

ODG	**correlation on the CRC’97 database**
TubFiltODG05a	0.854
TubFiltODG04a	0.837
FauFiltODG03a	0.864
TubFiltODG02a	0.826
TubNnODG04b	0.858
FauODG07a	0.849
TubODG07b	0.845
FauODG05a	0.849
TubODG05b	0.825
FauODG05c	0.846
TubNnODG05	0.867
TubNnODG07	0.864

Tab. 7.1: Correlations between ODGs and subjective gradings on the CRC’97 database for all mappings evaluated in the preselection phase.

In the ODG-names starting with “Tub” or “Fau”, the number in the end indicates the number of MOVs used as inputs. The letters “Nn” denote neural network mappings. Purely filterbank-based ODGs are denoted by the letters “Filt”, and purely FFT-based ODGs are denoted by the letters “Fft”. If neither of these letters appear in the ODG-name, the mapping uses MOVs of both parts of the ear model. In the ODG-names starting with “Opt”, the number after the hyphen indicates the number of MOVs used as inputs. The ODGs starting with “Opt” or “Perc” all use a neural network and are based on the FFT-based part of the ear model.

Several of the evaluated ODGs show very high correlation on the previously unknown CRC’97 database. The results presented in Table 7.1 lead to the selection of 13 out of the 26 tested ODGs. In addition, another ODG was added by OPTICOM, which was mainly trained on database 2. The purpose of this ODG was to check the assumption that the new database might be more similar to database 2 than to other databases because database 2 was created for the same purpose and is the most recent data set.

The following ODGs were submitted for the official validation test:

- **FFT-based version**: *OptOdg1-15H3*, *OptOdg1-15H3-DB2*, *TubFftODG05a*, *TubFftODG05b*, *TubFftODG03a*, *TubFftNnODG04b*.
- **Filterbank-based version**: *FauFiltODG03a*, *TubFiltODG02a*, *TubFiltNnODG3*, *TubFiltNnODG04a*.
- **Combined version**: *TubODG07b*, *FauODG05a*, *TubODG04a*, *TubNnODG05*.

The submitted ODGs were renamed in a systematic order, only preserving information on the ear model and mapping type, together with a numbering from one to three. The naming convention was: *[ear model][mapping type]ODG[numbering]*. The correspondence between internal and official names is included in Table 7.2.

7.4.2 Results of the First Phase

In the first phase of the ITU-R validation test, the filterbank-based version of PEAQ turned out to be clearly superior to the FFT-based version. Both versions of PEAQ showed a better performance than the reference model (PAQM). However, the absolute performance of the model on database 3 was not very good. When looking at correlations between model predictions and SDGs (Table 7.2), the best FFT-based version performs clearly better than the reference model, but the correlation of 0.69 is still rather poor. The best filterbank-based version performs clearly better, but when compared to its results on the unofficial validation database (CRC’97), the correlation is still lower than expected.

“official” name	internal name	84 items	52 items	42 items
FftODG1	(TubFftODG05b)	0.650	0.671	0.781
FftODG2	(TubFftODG03a)	0.633	0.641	0.759
FftODG3	(TubFftODG05a)	0.671	0.685	0.780
FftNnODG1	(OptOdg1-15H3)	0.693	0.699	0.793
FftNnODG2	(TubFftNnODG04b)	0.629	0.629	0.718

“official” name	**internal name**	**84 items**	**52 items**	**42 items**
FftNnODG3	(OptOdg1-15H3-DB2)	0.619	0.606	0.719
FiltODG1	(FauFiltODG03a)	0.747	0.742	0.861
FiltODG2	(TubFiltODG02a)	0.747	0.739	0.860
FiltNnODG1	(TubFiltNnODG3)	0.765	0.775	0.876
FiltNnODG2	(TubFiltNnODG04a)	0.741	0.735	0.858
CombODG1	(TubODG04a)	0.740	0.744	0.865
CombODG2	(FauODG05a)	0.741	0.754	0.858
CombODG3	(TubODG07b)	0.629	0.640	0.721
CombNnODG1	(TubNnODG05)	0.749	0.756	0.863
B3	*PAQM*	*0.636*	*0.586*	*0.662*

Tab. 7.2: Correlations of the ODGs used in the first phase of the ITU-R validation test with the subjective gradings from database 3.

7.4.3 Second Phase: Fitting the Models to the Database

After the first phase of the validation test, half of the database was released, and the model versions were adjusted to the new database. When selecting the subset of database 3 to be released, care has been taken to ensure that the released part includes examples for all musical excerpts and all introduced artefacts. Moreover, the distribution of subjective gradings was to be similar in both parts of the database. Nevertheless, the deviations between subjective gradings and model predictions were not considered when selecting these data. Unfortunately, all problematic items remained in the hidden part of the data set. The released data set only included items that were already "perfectly" predicted by the measurement methods. As this made an adjustment of the model versions to the new database impossible, ten additional items were released, which included several problematic items. When trying to adjust the model versions to these data, it turned out that simply retraining the mapping functions on the new data did not yield significant improvements. Therefore, a new search for convenient MOV combinations was carried out. The mappings were trained on most of the previously known databases, and the released half of database 3 was used as validation data set during training. The EIA-database and several items from database 2 were not included in the training data set. The EIA-database was excluded because it turned out to be irrelevant when optimising a model for the prediction of database 3: some of the results of the first phase had shown that removing the EIA database from the training set even increased the performance on the validation data set. From database 2, those test items were excluded that were considered as outside the scope of the measurement method. These were the items where the artefacts were caused by error sources other than audio codecs.

Just like in the first phase, the CRC’97 database was excluded from the training procedure, so it could be used for the internal validation and preselection before submitting the new mappings to the official test site. The MOVs in the new mappings where mainly the same as in the first phase, but for some MOVs, the linear temporal average was replaced by the squared average. In the filterbank-based part of the combined version, the time constants used for forward masking were deminuated from eight milliseconds to four milliseconds. Both changes tend to enhance the differences between filterbank-based and FFT-based part of the ear model, which apparently is an advantage when combining both ear models.

The finally submitted ODGs were selected from a total number of 61 MOV combinations. The preselection criterion was again the performance on the CRC'97 database. In addition, the performance on the released part of database 3 was considered. The results of the preselection test are given in Table 7.3. The finally submitted ODGs are printed in bold letters.

data set	**db3**	**db3**	**CRC97**
number of test items	**42**	**52**	**136**
OptOdg0-5H2	0.781	0.712	0.750
OptOdg1-15H3	0.793	0.697	0.843
OptOdg1-15H3-DB2	0.726	0.606	0.823
OptOdg2-5H2	0.829	0.744	0.779
OptOdg3-6H3 (FftNnODG3)	**0.827**	**0.761**	**0.816**
OptOdg5-6H3	0.821	0.749	0.705
OptOdg7-15H4	0.813	0.715	0.000
OptOdg8-14H3	0.782	0.656	0.068
OptOdg9-15H3 (FftNnODG1)	**0.834**	**0.763**	**0.837**
OptOdg10-15H4	0.765	0.709	0.852
TubFftODG1	0.830	0.773	0.686
TubFftODG2	0.853	0.778	0.694
TubFftODG3 (FftODG2)	**0.842**	**0.800**	**0.689**
TubFftODG4	0.828	0.779	0.713
TubFftODG5 (FftODG1)	**0.833**	**0.778**	**0.731**
TubFftODG6 (FftODG3)	**0.832**	**0.748**	**0.729**
TubFftNnODG6g	0.812	0.724	0.787
TubFftNnODG6h	0.811	0.751	0.761
TubFftNnODG6i	0.834	0.761	0.798
TubFftNnODG7g	0.818	0.746	0.774
TubFftNnODG7h	0.813	0.736	0.777
TubFftNnODG7i (FftNnODG2)	**0.830**	**0.765**	**0.779**
TubFiltODG1 (FiltODG2)	**0.901**	**0.830**	**0.862**

data set	**db3**	**db3**	**CRC97**
number of test items	**42**	**52**	**136**
TubFiltODG2 (FiltODG1)	**0.924**	**0.878**	**0.836**
TubFiltODG3	0.917	0.837	0.814
TubFiltODG4	0.902	0.833	0.843
TubFiltODG5	0.904	0.837	0.838
TubFiltODG6	0.908	0.839	0.772
TubFiltODG7 (FiltODG3)	**0.904**	**0.837**	**0.839**
TubFiltNnODG3a (FiltNnODG2)	**0.897**	**0.822**	**0.868**
TubFiltNnODG3b	0.895	0.820	0.871
TubFiltNnODG3c	0.891	0.818	0.877
TubFiltNnODG3d	0.891	0.818	0.879
TubFiltNnODG3e (FiltNnODG1)	**0.910**	**0.832**	**0.830**
TubFiltNnODG3f	0.911	0.826	0.835
TubFiltNnODG4a (FiltNnODG3)	**0.899**	**0.825**	**0.878**
TubFiltNnODG4b	0.808	0.739	0.752
TubFiltNnODG4c	0.805	0.733	0.748
TubFiltNnODG4d	0.836	0.775	0.849
TubODG1 (CombODG3)	**0.904**	**0.863**	**0.777**
TubODG2 (CombODG1)	**0.921**	**0.854**	**0.796**
TubODG3	0.918	0.839	0.827
TubODG4	0.926	0.859	0.805
TubODG5	0.921	0.835	0.822
TubODG6	0.926	0.859	0.805
TubODG7 (CombODG2)	**0.916**	**0.828**	**0.793**

data set	**db3**	**db3**	**CRC97**
number of test items	**42**	**52**	**136**
TubODG8	0.907	0.812	0.820
TubNnODG5a	0.902	0.814	0.856
TubNnODG5b	0.908	0.825	0.888
TubNnODG5c (CombNnODG1)	**0.920**	**0.837**	**0.842**
TubNnODG5d	0.910	0.830	0.885
TubNnODG9a (CombNnODG2)	**0.872**	**0.826**	**0.881**
TubNnODG9b	0.887	0.810	0.844
TubNnODG9c	0.873	0.816	0.871
TubNnODG9d	0.890	0.806	0.807
TubNnODG9e	0.891	0.809	0.820
TubNnODG9f	0.878	0.800	0.872
TubNnODG7a	0.882	0.811	0.863
TubNnODG7b	0.890	0.822	0.816
TubNnODG7c	0.895	0.808	0.824
TubNnODG7d (CombNnODG3)	**0.859**	**0.838**	**0.851**

Tab. 7.3: Correlations of the ODGs used in the optimisation phase of the ITU-R validation test with the subjective gradings from the CRC'97 database and the released part of database 3.

The following ODGs were submitted for the third phase of the official validation test:

- **FFT-based model**: *TubFftODG3*, *TubFftODG5*, *TubFftODG6*, *OptOdg3-6H3*, *OptOdg9-15H3*, *TubFftNnODG7i*.
- **Filterbank-based model**: *TubFiltODG1*, *TubFiltODG2*, *TubFiltODG7*, *TubFiltNnODG3a*, *TubFiltNnODG3e*, *TubFiltNnODG4a*.
- **Combined model**: *TubODG1*, *TubODG2*, *TubODG7*, *TubNnODG5c*, *TubNnODG9a*, *TubNnODG7d*.

Again, the submitted ODGs were renamed in a systematic order, only preserving information on the ear model and mapping type, together with a numbering from one to three. The correspondence between internal and official names is included in Table 7.3 (the official names are given in brackets).

7.4.4 Results of the Third Phase

It turned out that for none of the model versions all model predictions were inside the tolerance range (Figures 7.15 through 7.18, the tolerance range is indicated by the two grey curves). However, the figures indicate that all model versions yielded clearly better predictions than the reference model, and both combined version and filterbank-based version yielded better predictions than the FFT-based version. Applying other performance criteria results in the same ranking. The following performance criteria were used:

- The number of outliers with respect to the confidence intervals (see Table 7.4).
- The number of severe outliers with respect to the absolute distance between model predictions and subjective gradings (see tables 7.5 - 7.6).
- The averaged squared distance between model predictions and the pre-defined tolerance range.
- The average error scores as defined in Section 7.1.3 (see Table 7.7).

- The cross correlations between model predictions and subjective gradings (see Table 7.9 and Figure 7.19).
- The standard error of the model predictions with respect to the subjective gradings (see Table 7.10).

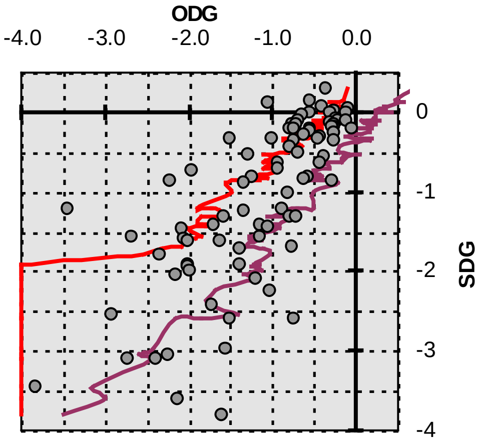


Fig. 7.15: Verification results of the FFT-based version recommended for standardisation (FftNnODG1).

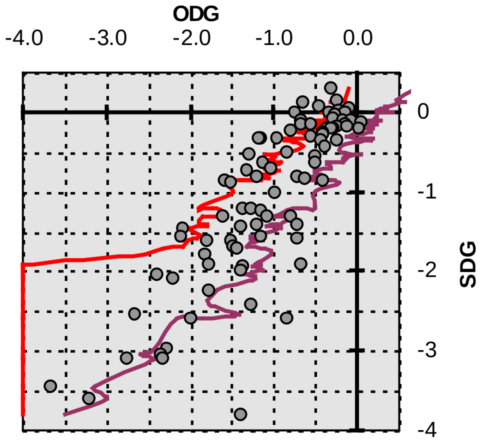


Fig. 7.16: Verification results of the combined version recommended for standardisation(CombNnODG3).

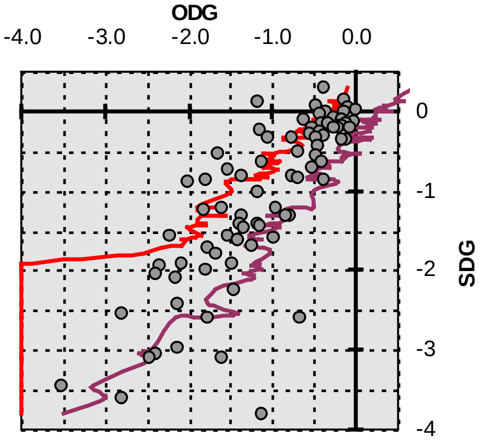


Fig. 7.17: Verification results of the best filterbank-based model (FiltODG3).

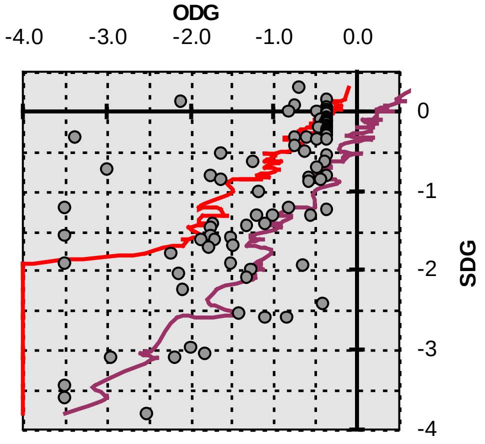


Fig. 7.18: Verification results of the reference model (PAQM).

outlier direction	**FFT-NnODG1**	**FFT-NnODG2**	**Filt-ODG2**	**Filt-ODG3**	**Comb-NnODG3**	**Comb-ODG3**	**PAQM**
too sensitive	10	4	4	4	3	5	
"deaf"	13	13	11	13	12	14	
total	23	17	15	17	15	19	

Tab. 7.4: Number of outliers that deviate from the subjective gradings by more than twice the subjective confidence interval (database 3).

model	**Comb-ODG3**	**Comb-NnODG3**	**FftNn-ODG1**	**FftNn-ODG2**	**Filt-ODG2**	**Filt-ODG3**	**PAQM**
number of outliers	4	4	12	12	9	6	14

Tab. 7.5: Number of outliers that deviate from the subjective gradings by more than one grade on the five-grade impairment scale (database 3).

model	**Comb-ODG3**	**Comb-NnODG3**	**Fft-NnODG1**	**Fft-NnODG2**	**Filt-ODG2**	**Filt-ODG3**	**PAQM**
number of outliers	2	2	3	2	2	2	8

Tab. 7.6: Number of outliers that deviate from the subjective gradings by more than 1.5 grades on the five-grade impairment scale (database 3).

In the first step of the comparison among the different model versions, the best two versions were identified for each ear model. Since all performance criteria pointed in the same direction, this was easily done. The best versions were:

- FFT-based model: *FftNnODG1* and *FftNnODG2*.
- Filter bank- based model: *FiltODG2* and *FiltODG3*.
- Combined model: *CombODG3* and *CombNnODG3*.

When comparing the performance of the different ear models of PEAQ, almost all decision criteria yield the same ordering. Filterbank-based and combined version perform clearly better than the FFT-based versions, and PEAQ generally performs better than the reference model. The only performance criterion that gives a different picture is the absolute error score for the second half of database 3 (Table 7.7). Here, the reference model performs better than PEAQ, and the different versions of PEAQ perform almost equal. The reason for this is that the absolute error scores for database 3 are mainly determined by one single test item, which has a very low subjective confidence interval (compare Section 7.1.3). The different ordering of the absolute error scores as compared to other performance criteria solely reflects that PAQM yields a better prediction for this particular item than PEAQ.

	FFT-NnODG1	FFT-NnODG2	Filt-ODG2	Filt-ODG3	Comb-NnODG3	Comb-ODG3	PAQM
DB3-2nd	2.96	2.79	3.16	3.16	2.91	2.56	2.39
CRC	1.55	1.85	1.61	1.67	1.61	1.90	2.78

Tab. 7.7: Absolute Error Scores for the validation data sets.

The decision between the two best mappings for each model version was much more complicated. Within both FFT-based and combined model, the results on the official validation database (DB3) indicate a different ordering than the results on the internal validation database (CRC'97). The analysis of outliers within database 3 (tables 7.4 - 7.6) corroborates that the combined versions are superior to the FFT-based versions and the reference model. The different mappings within each version perform rather similar, except that *FFTNnODG1* grades several test items too pessimistic (see Table 7.4).

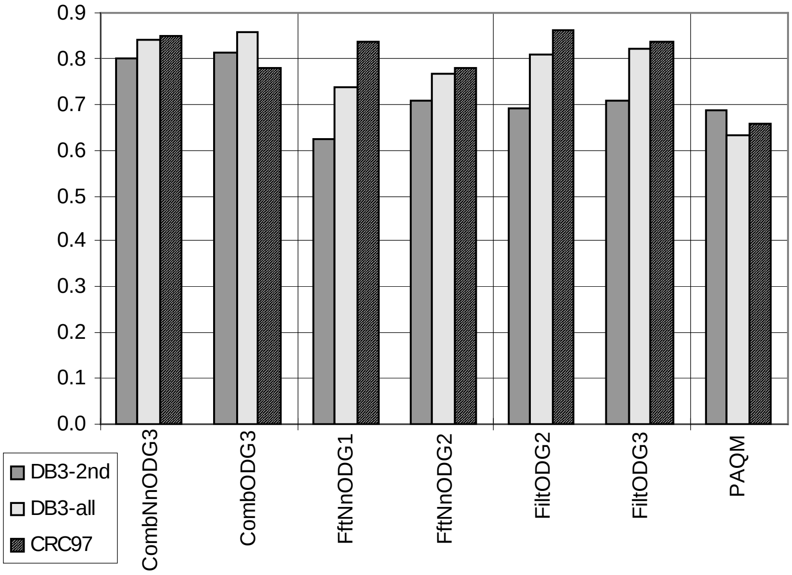


Fig. 7.19: Correlation between SDGs and ODGs for the validation data sets.

With respect to the correlations on database 3 (Table 7.8 and Figure 7.19), *FFTNnODG2* performs better than *FFTNnODG1*, whereas there is no significant difference between the two mappings for the combined model. The correlations on the CRC'97 database *FFTNnODG1* performs best within the FFT-based model, and *CombNnODG3* performs best within the combined model.

	FFT-NnODG1	FFT-NnODG2	Filt-ODG2	Filt-ODG3	Comb-NnODG3	Comb-ODG3	PAQM
DB3-2nd	0.671	0.728	0.738	0.751	0.828	0.826	0.710
CRC	0.837	0.779	0.862	0.839	0.851	0.777	0.656

Tab. 7.8: Correlations between SDGs and ODGs for the validation data sets.

When looking at the performance of the models on the training database (tables 7.9 - 7.10), one has to keep in mind that these results reflect a fitting of the model outputs to the subjective data rather than a true prediction. Therefore, a model that performed bad in the validation is to be considered an unreliable model, no matter how well it performs on the training data sets. On the other hand, the training data set is much larger than the validation data set, and accurate predictions for the training data set are therefore required. This can be taken into account by looking at the most critical data set among training and validation data. As the standard error depends less on the distribution of the subjective gradings than the cross correlation, it should be preferred when looking at the performance of different data sets. In Table 7.10, a ranking is established according to the largest standard error for each mapping. Within the FFT-based version, *FFTNnODG2* performs better than *FFTNnODG1*, and *CombNnODG3* again shows the best performance among all versions. However, the purely filterbank-based mappings perform not significantly worse than the combined model.

In the final decision within the ITU-R TG 10/4, a high weight was assigned to the performance on the CRC'97 database because it provides a larger and more realistic data set than the official validation database. Consequently, *FFTNnODG1* was selected for the basic version of PEAQ, and *CombNnODG3* was selected for the advanced version of PEAQ.

subsets	**number of items**	**Comb-NnODG3**	**Comb-ODG3**	**FftNn-ODG1**	**FftNn-ODG2**	**Filt-ODG2**	**Filt-ODG3**
(DB2-all)	(91)	(0.750)	(0.729)	(0.829)	(0.833)	(0.769)	(0.758)
DB2-clean	79	0.872	0.867	0.839	0.866	0.847	0.842
ITU93	42	0.897	0.891	0.883	0.799	0.917	0.931
MPEG90-all	50	0.968	0.952	0.943	0.943	0.937	0.935
MPEG90-headphone	40	0.971	0.959	0.944	0.945	0.948	0.944
EIA	81	0.783	0.708	0.811	0.796	0.785	0.755
ITU92CO	50	0.686	0.593	0.651	0.525	0.492	0.421
ITU92DI	60	0.887	0.833	0.851	0.849	0.806	0.840
MPEG91	105	0.928	0.931	0.939	0.918	0.871	0.892
MPEG95	132	0.823	0.722	0.829	0.808	0.807	0.831
DB3-all	84	0.840	0.857	0.739	0.764	0.809	0.819
DB3-1st	42	0.859	0.904	0.834	0.830	0.901	0.904
DB3-1st + 10	52	0.838	0.863	0.763	0.765	0.830	0.837
DB3-2nd	32	0.799	0.813	0.622	0.706	0.692	0.709

subsets	number of items	Comb-NnODG3	Comb-ODG3	FftNn-ODG1	FftNn-ODG2	Filt-ODG2	Filt-ODG3
CRC97	136	0.851	0.777	0.837	0.779	0.862	0.839
worst[11]	-	0.783	0.708	0.622	0.706	0.692	0.709
rank	-	1	3	6	4	5	2

Tab. 7.9: Correlations between SDGs and ODGs for all data sets.

subsets	number of items	Comb-NnODG3	Comb-ODG3	Fft-NnODG1	Fft-NnODG2	Filt-ODG2	Filt-ODG3
(DB2-all)	(91)	(0.727)	(0.753)	(0.614)	(0.608)	(0.703)	(0.717)
DB2-clean	79	0.508	0.517	0.565	0.520	0.552	0.560
ITU93	42	0.245	0.252	0.260	0.333	0.221	0.202
MPEG90-all	50	0.385	0.470	0.513	0.514	0.538	0.548
MPEG90-headphone	40	0.368	0.439	0.511	0.507	0.492	0.509
EIA	81	0.494	0.561	0.465	0.481	0.492	0.521
ITU92CO	50	0.455	0.503	0.474	0.532	0.544	0.567
ITU92DI	60	0.474	0.568	0.539	0.543	0.607	0.558
MPEG91	105	0.364	0.354	0.336	0.386	0.478	0.441
MPEG95	132	0.551	0.671	0.543	0.572	0.573	0.540
DB3-all	84	0.603	0.573	0.749	0.717	0.653	0.638
DB3-1st	42	0.576	0.482	0.622	0.628	0.489	0.482
DB3-1st + 10	52	0.645	0.596	0.763	0.760	0.659	0.646
DB3-2nd	32	0.544	0.527	0.707	0.640	0.652	0.637
CRC97	136	0.515	0.618	0.537	0.615	0.498	0.534
worst	-	0.645	0.671	0.763	0.760	0.659	0.646
rank	-	1	4	6	5	3	2

Tab. 7.10: Standard errors between SDGs and ODGs for all data sets.

The results of the finally selected models on all available test items (database 1-3 and CRC'97) are depicted in Figure 7.20. Even though the overall correlation between model predictions and subjective gradings is rather high, for some test items the model prediction deviates severely from the subjective grade.

---

[11] DB2-all and ITU92co are not taken into account because DB2-all includes conditions, that are not in the scope of the measurement schemes and ITU92co covers only half of the possible SDG range (which reduces the correlations, so that a comparison with other subsets would not be fair).

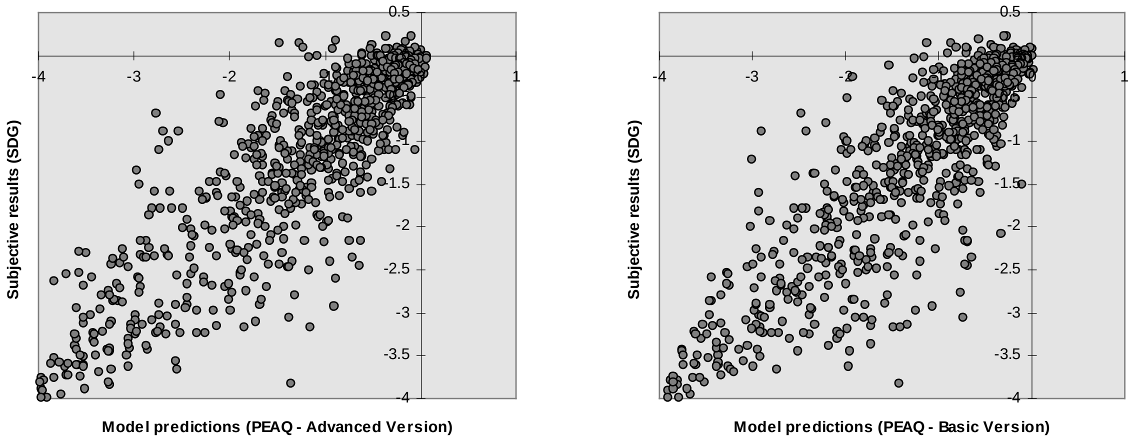


Fig. 7.20: Model predictions versus listening test results for all available data sets (left: advanced version, right: basic version).

The distribution of the prediction errors is close to a Gaussian function. Half of the test items are predicted within an accuracy of 0.3 grades on the five-grade impairment-scale. However, five percent of the test items show prediction errors of one grade and above (Figure 7.21).

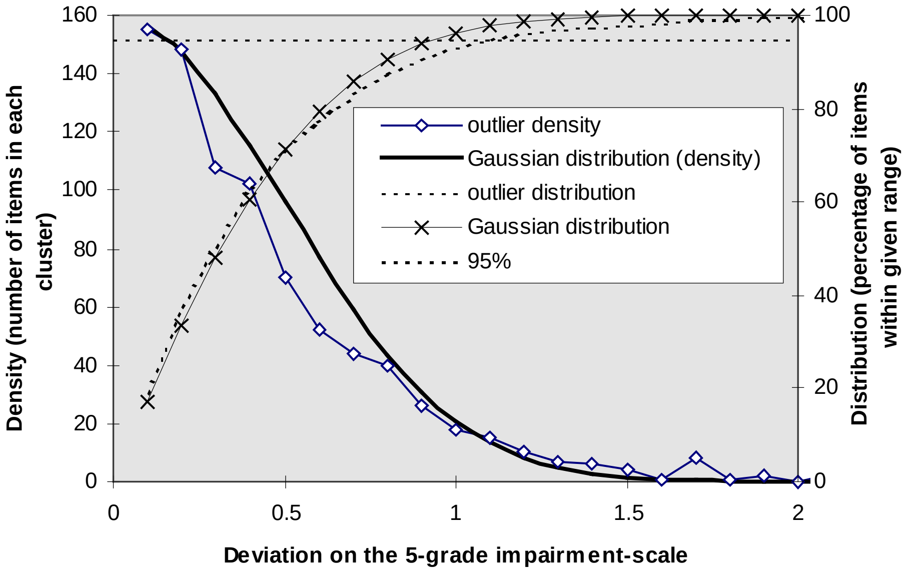


Fig. 7.21: Distribution of the prediction errors for the advanced version of PEAQ. The outlier density is counted in clusters of 0.1 grades on the five-grade impairment-scale.

7.5 Summary

For most kinds of coding artefacts, DIX is an excellent predictor of the perceived audio quality. It has proven to be one of the most reliable measurement methods that currently exist, and a combination of DIX with features of several other measurement methods (PEAQ) was considered as the best available measurement method by the ITU-R TG 10/4. However, there are certain kinds of distortions where DIX, and consequently also PEAQ, performs rather poorly. These problems have been reduced when combining features of different methods into PEAQ, but as DIX forms a major part of PEAQ, and similar problems also occur within the other methods, some conditions remain critical.

The expected accuracy of PEAQ is reduced in the following conditions:

- multiple cascades of identical codecs,
- codecs that introduce modulations due to an overuse of window switching,
- (male) speech items,
- distortions that only occur within few, short sections of the test excerpt.

This list is only one possible interpretation of the observed problems. It is not necessarily complete, and other interpretations of the observed problems may be possible. The last condition in the list is not a severe problem in most applications, since these artefacts are correctly detected by PEAQ, but do not get the appropriate weight in the temporal averaging. The other three conditions are the more serious problem of PEAQ, and their origin is not yet entirely clear. It could be a problem of the way the model output variables are calculated, but might as well reflect a missing detection cue or even a shortcoming of the underlying psychoacoustical models.

8. Simulating Psychoacoustical Experiments with PEAQ

8.1 Psychoacoustical Models and Perceptual Measurement

Psychoacoustical models normally are optimised to predict thresholds or sensational quantities for certain well defined conditions. In general, it is easy to establish such a model as long as the psychoacoustical data to be modelled are consistent. The task is much more difficult when the perception of signals with unlimited complexity is to be modelled. Psychoacoustical models for simple signal constellations mostly fail to predict the perception of more complex signals. This indicates that our understanding of the human auditory system is not deep enough to allow for really generalising models of perception. Therefore, it cannot be expected that a perceptual measurement method, which is optimised on complex signals only, yields accurate predictions for the simple signal constellations used in psychoacoustical standard experiments.

Nevertheless, it is worthwhile to check the behaviour of the measurement method in the context of simple psychoacoustical experiments. Even though optimising the method on such experiments does not increase the performance of the model in real-world applications, the outcome of such experiments may point to shortcomings of the method, and serve as a basis for further improvements.

8.2 Simultaneous Masking

Simultaneous masking curves were measured by adding a distortion (maskee) to the original signal (masker), and adjusting the weight of the distortion until an assumed threshold value for the respective MOV is reached. Both maskers and distortions were either pure tones or critical band noises. The critical band noise was generated by filtering a random signal by a 20th order Butterworth band pass. The critical bands were calculated from the approximation by Schroeder et al. [SCH79], which has also been used to determine the distribution of filter bands in DIX and PEAQ. Whereas for methods using the masked threshold concept, these experiments would simply render the implemented spreading function, for a method based on a comparison of internal representations, spreading function and resulting masking curves are not necessarily identical. Moreover, the MOVs from DIX also incorporate modulations, which makes the result of a simulated psychoacoustical experiment even less predictable.

8.2.1 Masking Properties of Tones and Noises

The masking properties of tones and noises were checked using a 70 dB sine tone or critical band noise at 10 Bark (1277 Hz) as a masker. All four possible constellations, tone-masking-noise, noise-masking-tone, tone-masking-tone, and noise-masking-noise were considered. The settings of PEAQ were identical to those used in the official validation test.

a) General Results

Figures 8.1-8.2 show the results for the MOV *RmsNoiseLoudAsymAvg*, which is one of the main outputs of DIX and PEAQ. The curves for noise-masking-noise and tone-masking-tone correspond closely to the implemented auditory filter shapes, but show a slight dip near above the masker frequency, which is not explicitly modelled. The lower slope is somewhat shallower than expected from the modelled filter shape, and the upper slope is somewhat too steep. The curve for noise-masking-tone is about 10 dB lower at the centre frequency of the masker, but almost identical to the other curves at the slopes. The curve for tone-masking-noise is clearly below the other curves in a range of almost two critical bands around the centre frequency of the masker. The central region of this curve is rather flat, and the slopes at the edges of the masking curve are again similar to the other curves.

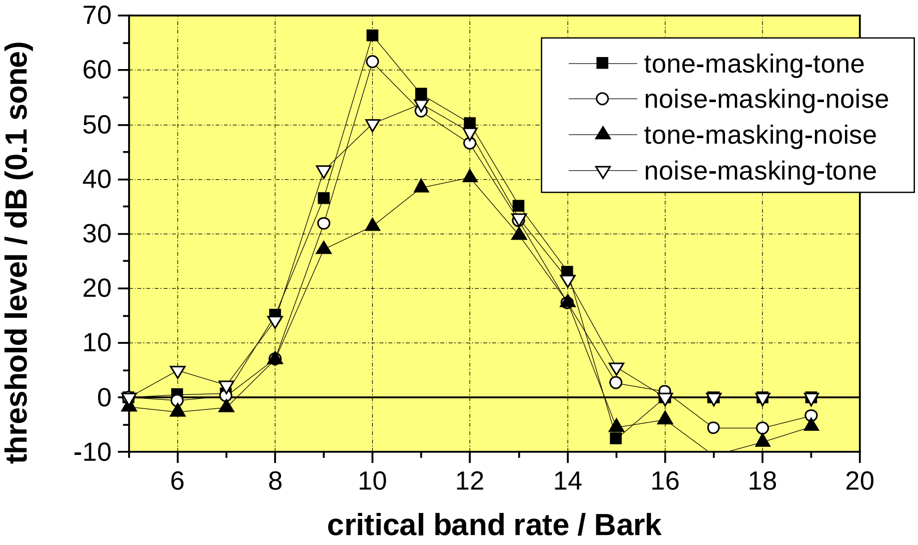


Fig. 8.1: Masking curves for the MOV* RmsNoiseLoudAsymAvg *with an assumed threshold value of 0.1 sone. The masker level is 70 dB.

When trying to identify the appropriate threshold value of the MOV, the decision is somewhat difficult. When predicting results of codec assessment tests, the border where items are transparent is at a noise loudness value between 0.1 and 0.2 sone, but when mapped to ODGs, the difference between these two values is clearly below the accuracy of the perceptual model. On the other hand, Figs. 8.1-8.2 show that in the prediction of masked thresholds, these two values correspond to a difference of more than 5 dB, which surely is significant. As the noise-masking-noise data is least influenced by beats and combination tones, it is taken as the decision criterion. Here, the value of -3 dB compared to the masker level at the centre frequency, which is found when assuming a threshold value of 0.2 sone, is much more realistic than the value of -8 dB when the threshold is assumed at 0.1 sone.

In general, these results correspond nicely to psychoacoustical data, but the curve tone-masking-tone seems to be somewhat too high. The peak at the centre frequency of the masker for tone-masking-tone and noise-masking-noise is caused by the level and pattern adaptation: as masker and maskee are identical at this point, the maskee is considered as a result of an amplification of the masker and is therefore partly suppressed.

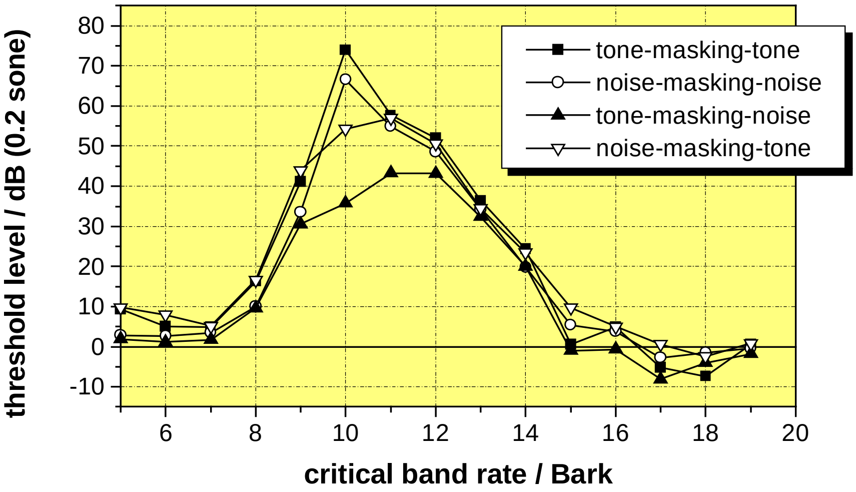


Fig. 8.2: Masking curves for the MOV* RmsNoiseLoudAsymAvg *with an assumed threshold value of 0.2 sone. The masker level is 70 dB.

The peaks of the masking curves for tone-masking-noise and noise-masking-tone are not at the centre frequency of the masker, but between one and two Bark above. This effect results from the concept of comparing internal representations: when assuming linearity and integration over all frequency bands, the masking curves are given by the convolution between two auditory filter shapes rather than the auditory filter shape itself. In the constellations tone-masking-tone and noise-masking-noise, this effect is hidden by the influence of the adaptation, which forces a peak of the masking curve at the centre frequency of the masker. The asymmetry of the filter shapes produces a shift of the maximum of the resulting masking curve. When the lower slope rate is neglected, the theoretical shift of the peak is given by

$$\Delta z = \frac{10}{\ln(10) \cdot \mathit{slope\ rate}} \qquad (8.1)$$

With the observed slope rate of 14 dB/Bark for the 70 dB masker, this would result in a theoretical shift of the peak by 0.3 Bark. Apparently, the observed shift is much higher. However, the theoretical shift was derived under the assumption of linearity, which clearly does not hold for the noise loudness. A better approximation is derived when using the slope of the specific loudness instead of the excitation. As the specific loudness is approximated by a compression of the excitation by raising it to a power of 0.23, the slope rate of the specific loudness function is given by the slope rate of the excitation multiplied by 0.23. Using this slope, the theoretical peak shift is 1.35 Bark, which comes close to the experimental result.

b) Level Dependence

Figure 8.3 shows the level dependence of the modelled masking curves for the constellation noise-masking-noise. This constellation has been used because it showed the most smooth slopes in the above experiments. Obviously, the level dependence of the modelled auditory filter is also rendered by the measured masking curves. The difference in the slope rates is in the correct range of 4 dB/Bark for a level difference of 20 dB, which corresponds to the level dependence found by Terhardt [TER79]. However, in the intermediate level range, the slopes are generally too steep. The measured slope rates are approximately 14 dB/Bark for the 70 dB masker, and 10 dB/Bark for the 90 dB masker. The expected slope rates are 10 dB/Bark for the 70 dB masker, and 6 dB/Bark for the 90 dB masker.

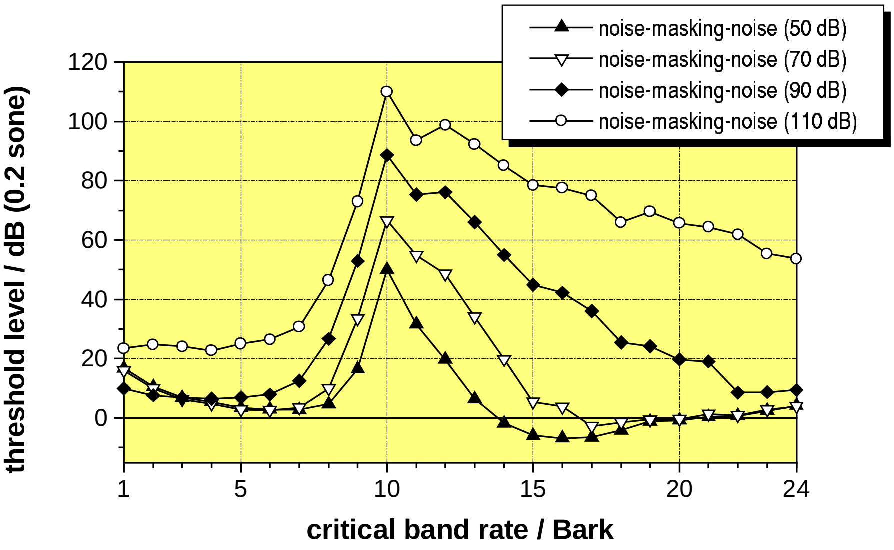


Fig. 8.3: Masking curves for the constellation noise-masking-noise for the MOV* RmsNoiseLoudAsymAvg *with an assumed threshold value of 0.2 sone.

c) Pure-Tone Masking

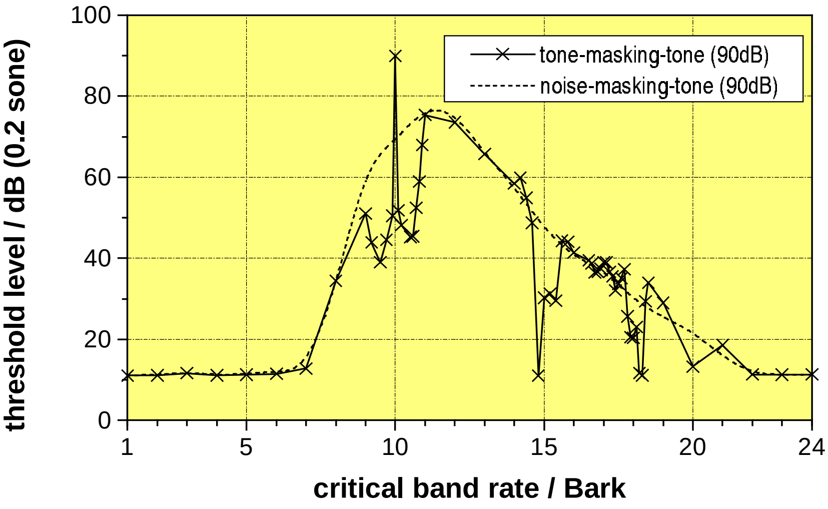


Fig. 8.4: Pure-tone masking curve versus noise-masking-tone for the MOV* RmsNoiseLoudAsymAvg *with an assumed threshold value of 0.2 sone. The masker level is 90 dB.

It is known from psychoacoustical experiments that in the constellation tone-masking-tone (often also: "pure-tone masking" or "two-tone masking") the amount of masking drops considerably in the neighbourhood of the masker frequency and odd multiples of it. This is caused by beats and combination tones. As one advantage of the filter bank is its property to preserve the temporal envelopes in the auditory filter bands, these effects should also occur in DIX and the filterbank-based part of PEAQ. To check this hypothesis, the tone-masking-tone experiments was carried out in more detail around the masker frequency and its third harmonic.

The hypothesis was corroborated by the experiment (Figure 8.4). The masking curves of PEAQ not only show the expected drop in masking around the masker frequency, but even show a significant dip around its third harmonic. This is a rather interesting effect because non-linearities, which normally serve to explain this phenomenon, are not explicitly modelled. Apparently, already the harmonic relation between masker and maskee sufficiently explains this effect.

8.2.2 Additivity of Masking

a) Noise-Masking-Noise

The superposition of two critical band maskers at 10 and 12 Bark (Figure 8.5) does not show the expected growth of masking at the upper slope and in the range between the maskers. The threshold at the lower slope is increased by approximately 10 dB.

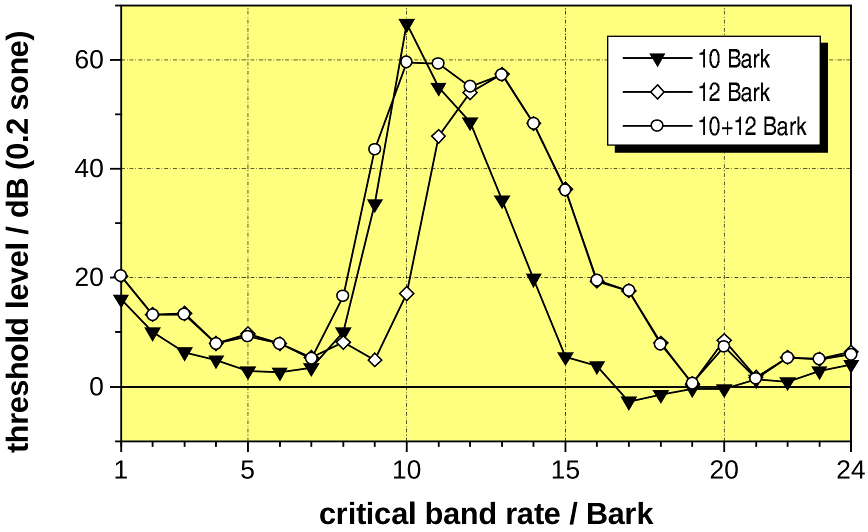


Fig. 8.5: Additivity of masking between a maskers at 10 and 12 Bark for the constellation noise-masking-noise for the MOV* RmsNoiseLoudAsymAvg *with an assumed threshold value of 0.2 sone. The masker level is 70 dB.

When the distance between the critical band maskers is increased to 4 Bark (Figure 8.6), the expected growth of masking is also observed in the range between the maskers, but still not at the upper slope.

Even though the gain of masking at the lower slope of the masking curve is not very apparent due to the high slope rate, its amount is remarkably high. Especially in Figure 8.6, the impact of the 12 Bark masker alone is almost zero at 9 Bark, but it still increases the masking curve of the combined masker by 10 dB compared to the 10 Bark masker alone. Both, the large amount of the growth of masking at the lower slope, and the absence of any growth of masking at the upper slope are not explained by any obvious feature of the model.

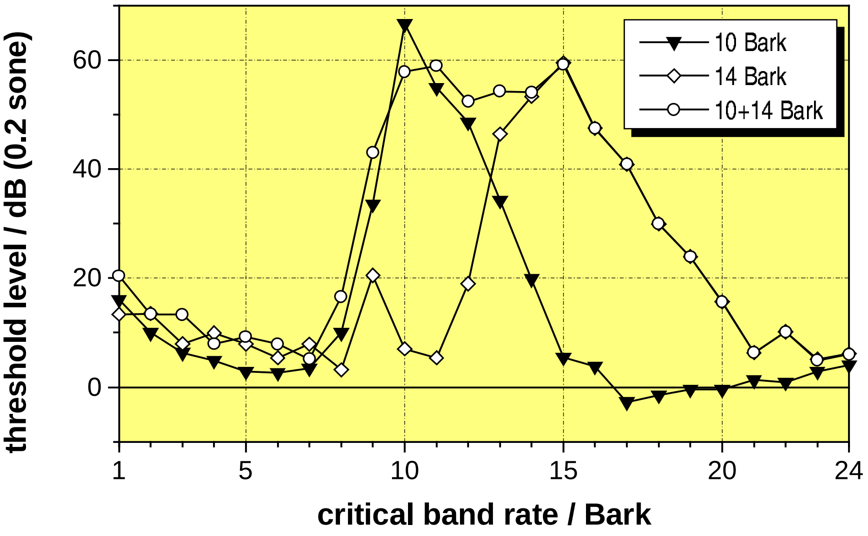


Fig. 8.6: Additivity of masking between a maskers at 10 and 14 Bark for the constellation noise-masking-noise for the MOV* RmsNoiseLoudAsymAvg *with an assumed threshold value of 0.2 sone. The masker level is 70 dB.

The curves in Figures 8.5 through 8.6 also show a decrease of masking at 10 Bark. As the peak of the masking curve at the centre frequency of the masker was caused by the adaptation, the apparent decrease of masking only reflects that with the combined masker the maskee at 10 Bark is treated as a distortion, and therefore detected at a more moderate level.

8.3 Temporal Masking

Temporal masking was investigated using -10 dB full scale tone bursts. The tone bursts were derived by filtering a single click with 8th order Bessel band pass filters. In addition, a short Gaussian pulse with a length of approximately 0.25 milliseconds was generated by filtering a click with a Bessel low pass. Bessel filters have the advantage of short, almost symmetric pulse responses, and the envelope of the pulse response of a higher order Bessel filter approximates a Gaussian shape.

Like pure-tone masking, the masking between tone-bursts is influenced by interferences between masker and maskee. Even with the Gaussian pulse, such interferences are observed in the model. As within each auditory filter, the Gaussian pulse becomes a tone-burst again, interferences still occur within each filter band, and the low number of filter bands in the model might not allow these interference patterns to be compensated in the spectral averaging process (which would happen with a continuous filter distribution). Only with the high frequency tone-burst, no interferences seem to occur. This is explained by the short duration of the high frequency tone-burst, which is below the smallest distance between masker and maskee in the experiment.

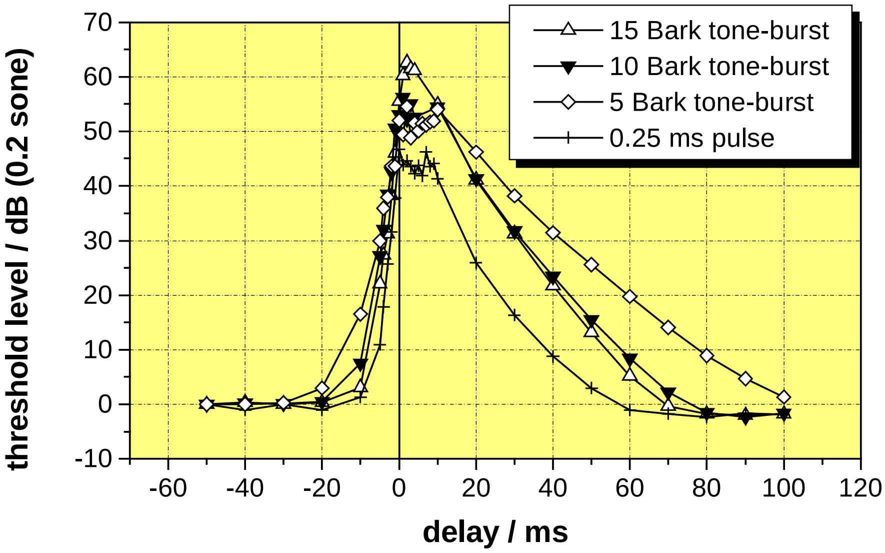


Fig. 8.7: Temporal masking curves for the MOV* RmsNoiseLoudAsymAvg *with an assumed threshold value of 0.2 sone. The masker level is 10 dB below clipping.

Except for the occurrence of interferences within the Gaussian pulse, the results meet the expectations. The descending slopes of the temporal masking curves (forward masking) render the shape of the time-smearing function implemented in the model, including its frequency dependence. Different to the implemented time-smearing function, the masking curves show also a slight frequency dependence of the ascending masking slope (backward masking). However, this might be caused by the increasing length of the critical band tone-bursts at lower frequencies. For the Gaussian pulse, the descend of the masking curve is not perfectly exponential anymore, but tends to flatten out with increasing distance to the masker. This corresponds even better to psychoacoustical masking curves than the implemented exponential functions.

8.4 Summary

In general, the shapes of masking curves derived from the perceptual model of DIX and the filterbank-based part of PEAQ correspond nicely to psychoacoustical data for masking between tones and critical band noises. The model yields the same interferences that are known from pure-tone masking experiments, even at harmonics of the masking tone. The masking asymmetry between pure-tones and noise-like maskers is also correctly predicted. However, DIX predicts the additivity between multiple critical band maskers only at the lower slope and in the range between the individual maskers, but not at the upper slope. This is probably a result of computing the threshold ratio from the envelope modulation after the temporal smearing. For the same reason, the frequency dependence of the threshold ratio for the constellation tone-masking-noise is not correctly predicted.

9. Outlook

As PEAQ, in general, yields rather accurate predictions of the perceived audio quality except for some particular conditions, identifying and removing the problems within these critical test items must be the main goal of further investigations. As pointed out in Section 7.5, problems may occur when one codec is multiply cascaded, a codec makes extensive use of window switching, or the test excerpt includes male speech. The problem with speech items might be influenced by signal recognition processes, and looking into the differences between speech quality measures and the models incorporated in PEAQ could be a starting point for the understanding of this problem. The other problems are probably related to temporal processing. As the temporal resolution has already been a subject of optimisation in DIX, simply increasing the temporal resolution will not solve these problems. However, there are several other aspects of the model related to the handling of time-domain artefacts. Improvements may be possible by investigating the influence of the shape of the temporal masking curves (which is currently rather simple), modelling non-linearities within temporal masking, and finding more sofisticated models of temporal integration.

Another possible improvement might be the inclusion of binaural detection cues. Some binaural modelling has already been implemented in earlier versions of DIX, but the resulting additional MOVs could not be successfully verified when using the available test data sets [TUB97]. This might be explained by the assumption that the present test data did not include any significant artefacts that required binaural processing. Nevertheless, such artefacts might occur in future codecs, and binaural detection cues should therefore be included in future measurement methods. In order to be able of assessing such effects and adjusting the model accordingly, listening tests especially designed to fit distinct parts of the perceptual model would be helpful in the further development of PEAQ.

In all attempts to handle the conditions that are critical for the current version of PEAQ, care has to be taken not to identify the problem items by the wrong detection cues. For example, multiply cascaded codecs can be easily identified by the characteristics of the band-limitation they introduce. However, this is not the artefact that is responsible for the perceived distortions! Tracking to such coincidences could easily look like an improvement of the measurement method, but will yield a wrong perceptual model (which could be easily tricked by a codec developer).

For the assessment of codecs which are somewhere in the gap between high quality audio coding and traditional speech coding (e. g. wide band speech or very low bit-rate audio), neither PEAQ nor established speech quality measures like PSQM may be appropriate. For these applications it may be necessary to develop completely new quality measures where the present methods may serve as a starting point.

10. Summary

The objective quality evaluation of perceptual codecs requires measurement methods which include models of human auditory perception. Although some of the known measurement methods already yielded reasonably accurate predictions for the subjective annoyance of many coding artefacts, the existing methods have numerous shortcomings. Some of the main shortcomings are insufficient temporal resolution and a restriction to steady-state models of auditory perception.

The new introduced perceptual measurement method DIX reduces some of these shortcomings and performes better than most of the formerly known approaches when predicting the perceived quality of coded audio signals. The method is based on a new, non-linear filter bank algorithm that provides level-dependent auditory filter shapes, and a high temporal resolution at a reasonably low computational effort. DIX evaluates the modulation of the temporal envelopes of the auditory filter outputs to predict the different masking characteristics of tonal and noise-like maskers. Moreover, it separates between linear and non-linear distortions, which takes the effect into account that imbalances in the frequency response of an audio device are not as annoying as the same amount of non-linear distortions like, for example, quantisation noise.

A combination of DIX with parts of other perceptual measurement methods (PEAQ) proved to be superior to formerly known perceptual measurement methods, and is the basis for the future ITU-R recommendation "*method for objective measurement of perceived audio quality*". Especially the part of the recommendation that addresses applications requiring maximum possible accuracy („*advanced version*“) is mainly based on DIX.

When simulating psychoacoustical experiments with DIX and PEAQ, the shapes of masking curves for different signal constellations are in most cases very accurately predicted. Especially, some effects occurring in pure-tone-masking, which are not correctly rendered by FFT-based perceptual models, are nicely predicted by the new model. However, for some signal constellations the absolute amount of masking differs from the expected values.

11. References

[BEE92] Beerends, J. G.; Stemerdink, J. A.: A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation. Journal of the Audio Engineering Society, Vol. 40 (12), December 1992, pp. 963-978.

[BEE93] Beerends, J. G.; Stemerdink, J. A.: The Optimal Time-Frequency Smearing and Amplitude Compression in Measuring the Quality of Audio Devices. Contribution to the 94th Convention of the Audio Engineering Society, Berlin, March 1993, Preprint 3604.

[BEE94] Beerends, J. G.; Stemerdink, J. A.: Modelling a Cognitive Aspect in the Measurement of the Quality of Music Codecs. Contribution to the 96th Convention of the Audio Engineering Society, Amsterdam, February 1994, Preprint 3800.

[BEE96] Beerends, J. G.; van den Brink, W. A. C.: The Role of Informational Masking and Perceptual Streaming in the Measurement of Music Codec Quality. Contribution to the 100th Convention of the Audio Engineering Society, Copenhagen, May 1996, Preprint 4176.

[BRA87] Brandenburg, K.: OCF - A New Coding Algorithm for High Quality Sound Signals. International Conference on Audio, Speech, and Signal Processing ’87, Dallas, Texas, USA, April 1987, pp. 141-144.

[BRA89] Brandenburg, K.: Ein Beitrag zu den Verfahren und der Qualitätsbeurteilung für hochwertige Musikcodierung. Dissertation am Lehrstuhl für technische Elektronik der Universität Erlangen-Nürnberg, Erlangen, 1989.

[BRA92] Brandenburg, K.; Sporer, Th.: NMR and Masking Flag: Evaluation of Quality Using Perceptual Criteria. AES 11th International Conference, Portland, Oregon, USA, 1992, pp. 169-179.

[BRO89] Bronstein, I. N.; Semendjajew, K. A.: Taschenbuch der Mathematik. Thun; Frankfurt/Main: Verlag Harry Deutsch, 1989.

[BUU86] Buus, S.; Schorer, E.; Florentine, M.; Zwicker, E.: Decision Rules in Detection of Simple and Complex Tones. Journal of the Acoustical Society of America, Vol. 80 (6), December 1986, pp. 1647-1657.

[CRO83] Crochiere, R. E.;Rabiner, L.: Multirate Filter Signal Processing. Englewood Cliffs, New Jersey: Prentice-Hall, 1983.

[COL93] Colomes, C.; Lever, M.; Rault, J. B.; Dehery, Y. F.: A Perceptual Model applied to Audio Bit-Rate Reduction. Contribution to the 95th Convention of the Audio Engineering Society, New York, October 1993, Preprint 3742.

[COL94] Colomes, C.; Lever, M.; Dehery, Y. F.: A Perceptual Objective Measurement System (POM) for the Quality Assessment of Perceptual Codecs. Contribution to the 96th Convention of the Audio Engineering Society, Amsterdam, February 1994, Preprint 3801.

[CRE85] Cremer, L.; Hubert, M.: Vorlesungen über Technische Akustik. Berlin; Heidelberg; New York; Tokyo: Springer Verlag, 1985.

[DEU92] Deutsch, W. A.; Noll, A.; Eckel, G.: The Perception of Audio Signals Reduced by Overmasking to the Most Prominent Spectral Amplitudes. Contribution to the 92nd Convention of the Audio Engineering Society, Vienna, March 1992, Preprint 3331.

[ENG85] Engeln-Müllges, G; Reutter, F.: Formelsammlung zur Numerischen Mathematik mit BASIC-Programmen. Mannheim; Wien; Zürich: BI Wissenschaftsverlag, 1985.

[FAS76] Fastl, H.: TEMPORAL MASKING EFFECTS: II. CRITICAL BAND NOISE MASKER. Acustica, Vol. 36, 1976, pp. 317-331.

[FIS64] Fischer, R.: Die Lautstärkeempfindung für Dauergeräusche und ihre Nachbildung in einem Meßgerät. Dissertation an der Fakultät für Elektrotechnik der Technischen Universität Berlin, Berlin, 1964.

[GLA90] Glasberg, B. R.; Moore, B. J.: DERIVATION OF AUDITORY FILTER SHAPES FROM NOTCHED NOISE DATA. Hearing Research, Vol. 47, 1990, pp. 103-138.

[GOE68] Goertzel, G.: AN ALGORITHM FOR THE EVALUATION OF FINITE TRIGONOMETRIC SERIES. Am. Math. Monthly, Vol. 65, 1968, pp. 34-35.

[GRE76] Green, D. M.: An Introduction to Hearing. Hillsdale, New Jersey: Lawrence Erlbaum Assoc., 1976.

[HAN96] Hansen, M.; Dau, T.; Kollmeier, B.: OBJEKTIVE SPRACHQUALITÄTS-VORHERSAGE MITTELS EINER GEHÖRORIENTIERTEN VORVERARBEITUNG. Fortschritte der Akustik, DAGA'96, Bonn, February 1996, pp.362-364.

[HÄR97] Härmä, A.; Laine, U. K.; Karjalainen, M.: WLPAC-A PERCEPTUAL AUDIO CODEC IN A NUTSHELL. Contribution to the 102nd Convention of the Audio Engineering Society, Munich, Germany, March 1997, Preprint 4420.

[HEL72] Hellman, R. P.: ASYMMETRY OF MASKING BETWEEN NOISE AND TONE. Perception & Psychophysics, Vol. 11 (3), 1972, pp. 241-246.

[HOL93] Hollier, M.P.; Hawksford, M. O.; Guard, D. R.: CHARACTERIZATION OF COMMUNICATIONS SYSTEMS USING A SPEECHLIKE TEST STIMULUS. Journal of the Audio Engineering Society, Vol. 41 (12), December 1993, pp. 1008-1021.

[HUM89] Humes, L. E.; Jesteadt, W.: MODELS OF THE ADDITIVITY OF MASKING. Journal of the Acoustical Society of America, Vol. 85 (3), March 1989, pp. 1285-1294.

[ISO75] ISO 532: ACOUSTICS - METHOD FOR CALCULATING LOUDNESS LEVELS. 1975

[ITU93] ITU-R: CCIR LISTENING TEST - NETWORK VERIFICATION TESTS WITHOUT COMMENTARY CODECS. Doc. 10-2/43, Canada and Italy, 1993.

[ITU96] ITU-T Recommendation P.861: OBJECTIVE QUALITY MEASUREMENT OF TELEPHONEBAND (300-3400 HZ) SPEECH CODECS. August 1996.

[ITU97] ITU-R Recommendation BS.1116(Rev.1): METHODS FOR THE SUBJECTIVE ASSESSMENT OF SMALL IMPAIRMENTS IN AUDIO SYSTEMS INCLUDING MULTICHANNEL SOUND SYSTEMS. 1997

[ITU98] ITU-R Recommendation BS.1387: METHOD FOR OBJECTIVE MEASUREMENTS OF SUBJECTIVELY PERCEIVED AUDIO QUALITY (UNDER PREPARATION). Geneva, 1998/99.

[KAP89] Kapust, R.: EIN GEHÖRBEZOGENES MEßVERFAHREN ZUR BEURTEILUNG DER QUALITÄT CODIERTER MUSIKSIGNALE. U.R.S.I. - Kleinheuerbacher Berichte, Band 33, Kleinheuerbach, October 1989, pp. 633 - 642.

[KAP93] Kapust, R.: QUALITÄTSBEURTEILUNG CODIERTER AUDIOSIGNALE MITTELS EINER BARK-TRANSFORMATION. Dissertation an der Technischen Fakultät der Universität Erlangen-Nürnberg Erlangen, 1993.

[KAR85] Karjalainen, M.: A NEW AUDITORY MODEL FOR THE EVALUATION OF SOUND QUALITY OF AUDIO SYSTEMS. IEEE International Conference of Acoustics, 1985, pp. 608-611.

[LAI90] Laine, U. K.; Altosaar, T.: AN ORTHOGONAL SET OF FREQUENCY AND AMPLITUDE MODULATED (FAM) FUNCTIONS FOR VARIABLE RESOLUTION FREQUENCY ANALYSIS. International Conference on Audio, Speech, and Signal Processing '90, New Mexico, USA, 1990, pp. 1615-1618.

[LIU93] K.J.Ray Liu: NOVEL PARALLEL ARCHITECTURES FOR SHORT-TIME FOURIER TRANSFORM. IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing, Vol. 40, No. 12, Dec. 1993.

[MOO83] Moore, B. C. J.; Glasberg, B. R.: SUGGESTED FORMULAE FOR CALCULATING AUDITORY-FILTER BANDWIDTHS AND EXCITATION PATTERNS. Journal of the Acoustical Society of America, Vol. 74 (3), September 1983, pp. 750-753.

[MOO89] Moore, B. C. J.: AN INTRODUCTION TO THE PSYCHOLOGY OF HEARING. New York: Academic Press, 1989.

[MOO93] Moore, B. C. J.: CHARACTERIZATION OF SIMULTANEOUS, FORWARD AND BACKWARD MASKING. AES 12th International Conference, Copenhagen, Denmark, June 1993, pp. 22-33.

[MOO97] Moore, B. C. J.; Glasberg, B. R.; Baer, Th.: A MODEL FOR THE PREDICTION OF THRESHOLDS, LOUDNESS, AND PARTIAL LOUDNESS. Journal of the Audio Engineering Society, Vol. 45 (4), April 1997, pp. 224-240.

[MPEG90] ISO/IEC/JTC1/SC2/WG11: MPEG/AUDIO TEST REPORT, Document MPEG90/N0030, October 1990.

[MUM91] Mummert, M.: RÜCKTRANSFORMATION DES KURZZEITSPEKTRUMS DER FOURIER-T TRANSFORMATION UND ANSATZ FÜR EINE GEHÖRGERECHTE TRANSFORMATIONSKODIERUNG. Fortschritte der Akustik, DAGA'91, Bochum, 1991, pp. 753-756.

[OPP75] Oppenheim, A.V.; Schafer, R.W.: Digital Signal Processing. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

[PAI92] Paillard, B.; Mabilleau, P.; Morissette, S.; Soumagne, J.: PERCEVAL: PERCEPTUAL EVALUATION OF THE QUALITY OF AUDIO SIGNALS. Journal of the Audio Engineering Society, Vol. 40 (1/2), January/February 1992, pp. 21-31.

[PAT76] Patterson, R. D.: AUDITORY FILTER SHAPES DERIVED WITH NOISE STIMULI. Journal of the Acoustical Society of America, Vol. 59 (3), March 1976, pp. 640-654.

[PAT86] Patterson, R. D.; Moore, B. C. J.: AUDITORY FILTERS AND EXCITATION PATTERNS AS REPRESENTATIONS OF FREQUENCY RESOLUTION. In: Moore, B. C. J.: (ed.): Frequency Selectivity in Hearing. New York: Academic Press, 1986.

[SCH79] Schroeder, M. R.; Atal, B. S.; Hall, J. L.: OPTIMIZING DIGITAL SPEECH CODERS BY EXPLOITING MASKING PROPERTIES OF THE HUMAN EAR. Journal of the Acoustical Society of America, Vol. 66 (6), December 1979, pp. 1647-1652.

[SCH79a] Schroeder, M. et al.: OBJECTIVE MEASURE OF CERTAIN SPEECH SIGNAL DEGRADATIONS BASED ON MASKING PROPERTIES OF HUMAN AUDITORY PERCEPTION. In: Lindblom; Öhman (eds.): Frontiers of Speech Communication Research. New York: Academic Press, 1979.

[SMI95] Smith III, J. O.; Abel, J. S.: THE BARK BILINEAR TRANSFORM. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, USA, October 1995.

[SPO96] Sporer, Th.: EIN NEUARTIGES VERFAHREN ZUR GEHÖRRICHTIGEN BEURTEILUNG VON AUDIO-SIGNALEN. Proceedings of the 19th International Convention on Sound Design (Tonmeistertagung), Verlag K.G. Saur, München 1997.

[SPO97] Sporer, Th.: OBJECTIVE AUDIO SIGNAL EVALUATION-APPLIED PSYCHOACOUSTICS FOR MODELING THE PERCEIVED QUALITY OF DIGITAL AUDIO. Contribution to the 103rd Convention of the Audio Engineering Society, New York, September 1997, Preprint 4512.

[STE36] Stevens, S. S.: A SCALE FOR THE MEASUREMENT OF A PSYCHOLOGICAL MAGNITUDE: LOUDNESS. Psychological Review, Vol. 43, 1936 pp. 405-416.

[STR80] Strube, H. W.: LINEAR PREDICTION ON A WARPED FREQUENCY SCALE. Journal of the Acoustical Society of America, Vol. 68 (4), 1980, pp. 1071-1076.

[STU92] Stuart, J. R.: IMPLEMENTATION AND MEASUREMENT WITH RESPECT TO HUMAN AUDITORY CAPABILITIES. AES UK DSP Conference, London, September 1992.

[STU93] Stuart, J. R.: NOISE: METHODS FOR ESTIMATING DETECTABILITY AND THRESHOLD. Contribution to the 94th Convention of the Audio Engineering Society, Berlin, March 1993, Preprint 3477.

[TER79] Terhardt, E.: CALCULATING VIRTUAL PITCH. Hearing Research, Vol. 1, 1979, pp. 155-182.

[TER85] Terhardt, E. VERFAHREN ZUR GEHÖRBEZOGENEN FREQUENZANALYSE. In: Fortschritte der Akustik, DAGA'85, Verl.: DPG-GmbH, Bad Honnef, 1985 pp. 811-814.

[TER92] Terhardt, E.: THE SPINC FUNCTION FOR SCALING OF FREQUENCY IN AUDITORY MODELS. Acustica, Vol. 77, 1992, pp. 40-42.

[THI94a] Thiede, Th.: Untersuchungen zur objektiven Qualitätsüberwachung bei bitratenreduzierten Tonsignalen. Diplomarbeit am Institut für Fernmeldetechnik der Technischen Universität Berlin, Berlin, 1994.

[THI94b] Thiede, Th.; Steinke, G.: ARBEITSWEISE UND EIGENSCHAFTEN VON VERFAHREN ZUR GEHÖRRICHTIGEN QUALITÄTSBEWERTUNG VON BITRATENREDUZIERTEN AUDIOSIGNALEN. Rundfunktechnische Mitteilungen, Vol. 38, 1994, pp. 102-114.

[THI94c] Thiede, Th.: GEHÖRRICHTIGE QUALITÄTSBEWERTUNG VON AUDIOSIGNALEN - ÜBERSICHT UND EINSCHÄTZUNG DER GEGENWÄRTIGEN VERFAHREN. Proceedings of the 18th International Convention on Sound Design (Tonmeistertagung), Verlag K.G. Saur, München 1995, pp. 623-642.

[THI96] Thiede, Th.; Kabot, E.: A NEW PERCEPTUAL QUALITY MEASURE FOR BIT RATE REDUCED AUDIO. Contribution to the 100th Convention of the Audio Engineering Society, Copenhagen, May 1996, Preprint 4280.

[THI98] Thiede, Th.; Treurniet, W. C.; Bitto,R.; Schmidmer, Ch.; Sporer, T.; Beerends, J. G.; Keyhl, M.; Colomes, C.; Lever,M.; Stoll,G.; Brandenburg, K.; Feiten, B.: PEAQ - DER KÜNFTIGE ITU-STANDARD ZUR OBJEKTIVEN MESSUNG DER WAHRGENOMMENEN AUDIOQUALITÄT.
Proceedings of the 20th International Convention on Sound Design (Tonmeistertagung), Verlag K.G. Saur, München 1999, pp. 724-766.

[TIE83] Tietze, U.; Schenk, Ch.: Halbleiter-Schaltungstechnik. Berlin, Heidelberg, New York: Springer-Verlag, 1983.

[ZWI67] Zwicker, E.; Feldkeller, R.: Das Ohr als Nachrichtenempfänger. Stuttgart: Hirzel Verlag, 1967.

[ZWI77] Zwicker, E.: PROCEDURE FOR CALCULATING LOUDNESS OF TEMPORALLY VARIABLE SOUNDS. Journal of the Acoustical Society of America, Vol. 62 (3), September 1977, pp. 675-682.

[ZWI80] Zwicker, E.; Terhardt, E.: ANALYTICAL EXPRESSIONS FOR CRITICAL BANDWIDTH AS A FUNCTION OF FREQUENCY. Journal of the Acoustical Society of America, Vol. 68 (5), November 1980, pp. 1523-1525.

[ZWI90] Zwicker, E.; Fastl, H.: Psychoacoustics, Facts and Models. Berlin; Heidelberg: Springer Verlag, 1990.

12. Acknowledgement

The author would like to thank the following people:

Gerhard Steinke and Prof. Klaus Fellbaum for initiating this work.

Prof. Peter Noll for giving me the opportunity to work on this thesis at the Institute of Communications Engineering and Theory of Electricity, and for giving me the freedom to carry out this work without any major restrictions.

Prof. Manfred Krause for many interesting discussions, not only on the subject of this work, but on any questions related to audio and acoustics.

All people involved in the work of ITU-R Task Group 10/4 for the fruitful co-operation, and for hints and comments. In particular Roland Bitto, Thomas Sporer and Dr. Karlheinz Brandenburg of FhG, Dr. John Beerends of KPN, Christian Schmidmer and Michael Keyhl of OPTICOM, William Treurniet of CRC, Catherine Colomes and Michel Lever of CCETT, Susanne Ritscher and Gerhard Stoll of IRT, René Tschannen and Daniel Ledermann of Swiss Telecom, and Thomas Rydén of Teracom.

The people at the audio department of Deutsche Telekom/Berkom for supporting this work. In particular Dr. Bernhard Feiten, Dr. Heinz Schaffner, Wolfgang Hoeg, Ulf Wüstenhagen, and Liane Kernchen.

All of my colleagues at the Institute of Communications Engineering and Theory of Electricity, in particular Martin Drews, Dr. Kai Clüver, Andreas Willig, Guido Heising, Dr. Marcus Purat, Christian Günther, and most of all Ernst Kabot, Florian Lenzner, and Nicole Brandenburg.

Andreas Willig, Dr. Kai Cluever, Manfred Thiede, and Ernst Kabot for thoroughly reviewing this thesis.

And, last but not least, my parents, sister, and friends.

13. Remarks

A large part of this work was supported by *Deutschen Telekom AG / Berkom*. Parts of the described measurement method are subject to copyright protection and to existing or pending patents.

The described measurement method will be a major part of the ITU-R recommendation “*Method for Objective Measurements of Perceived Audio Quality*”.

14. Appendix

14.1 Index

absolute threshold
  absolute threshold of hearing........18
advanced.......................................108
asymmetry of masking.....................17
auditory event..................................90
auditory filters 17
AV...108
backward masking 14, 80
Bark..21
basic 108
bilinear transform 52
binaural masking level difference 22
Butterworth approximation..............52
BV...108
Chebychev approximation 52
cognitive model...............................34
commodulation masking release 22
comparing internal representations...36
critical bands....................................19
  critical band scale 21
DC-rejection-filter 76
degrees of freedom 72, 79
difference limen...............................20
dip listening.....................................22
Disturbance Index............................71
DiX...71
end-of-scale effects........................103
equal loudness contours...................18
equivalent rectangular bandwidth21, 47
ERB
  ERB scale....................................21
excitation...23
excitation patterns............................23
FAMlet..52
FDC-filter.......................................67
FDC-filter bank 74
filter bands
  maximum required number of 78
forward masking........................ 14, 80
Fourier-Time-Transform 55
frequency-sampling structure..... 58, 67
frequency-to-pitch mapping.............46
Functional models 35
gamma-function 56
Gammatone filter bank....................56
half-wave rectification.....................79
Heuristical models...........................35
internal noise....................... 32, 81, 83
just noticeable difference.................20
linear phase 82
masked threshold.............................14
masked threshold concept................35
maskee 14
masker..14
masking..13
masking curves................................16
masking index 17
masking threshold............................14
mel

  mel scale..21

middle ear transfer function.......32, 81

modulation measure..........................93

modulation pattern.....................74, 93

MOV...108

noise disturbance 39

orthogonality 90

  orthogonality relation...................90

partial loudness................................20

partial masking 20, 31

peak detection..................................79

PEAQ....................................107, 134

perceptual streaming........................90

peripheral ear model 34

phon

  phon scale 18

pitch scale..21

post-masking 14

pre-masking.....................................14

resolution

  spectral resolution........................47

  temporal resolution.......................47

ROEX-filter.....................................24

ROunded EXponentials 24

simplified loudness..........................94

simultaneous masking......................14

sone..19

specific loudness.............................31

  specific loudness pattern 23

spectral increment scale...................21

spectral resolution...........................47

SPINC..21

spreading function 37

stapedius...22

temporal masking 14

temporal resolution..........................47

tensor tympani.................................22

threshold factor...............................17

threshold in quiet.............. 18, 81, 112

tuning curves

  neural tuning curves.....................17

  psychophysical tuning curves 17

unsmeared excitation patterns.... 74, 82

worst case masking curve 30, 39

14.2 Table of Figures

Fig. 2.1: Explaining backward masking by filter shapes and threshold. 14

Fig. 2.2: Masking pattern for a narrow band noise pulse (from [FAS76]). 16

Fig. 2.3: Masking curves for simultaneous masking of tones by narrow band noise (from [ZWI67]). 16

Fig. 2.4: Correspondence between masking curves, tuning curves and auditory filter shapes. 17

Fig. 2.5: Equal loudness contours (from [ZWI67]). 19

Fig. 2.6: Comparison of different approximations of the ERB-scale (left: bandwidth, right: ERB-rate). 28

Fig. 2.7: Stages of Auditory Processing. 34

Fig. 2.8: Ordering of different concepts of perceptual models between heuristic and functional models. 35

Fig. 2.9: Masked threshold concept. 36

Fig. 2.10: Comparison of internal representations. 36

Fig. 2.11: Frequency warping when modelling an auditory frequency scale (upper: linear frequency scale, lower: logarithmic frequency scale). 37

Fig. 2.12: Frequency domain smearing by convolving a linear spectrum with a spreading function (example: 4 pure tones + threshold in quiet, linear additivity). 38

Fig. 2.13: Relation between the worst case masking curve(dashed lines) used in NMR and level-dependent masking curves (solid lines). 39

Fig. 3.1: Excitation pattern of a pure tone for an FFT based measurement algorithm that groups the spectral coefficients into rectangular bands. 46

Fig. 3.2: Uncertainty in the estimation of the amount of masking between two narrow band signals. 47

Fig. 3.3: Frequency response of an auditory filter at a centre frequency of 1600 Hz and a 40th order analogue filter consisting of a low pass and a high pass with Butterworth characteristic. 51

Fig. 3.4: Frequency response of an auditory filter at a centre frequency of 1600 Hz and a 12th order analogue band pass with Butterworth characteristic. 52

Fig. 3.5: Approximation for the Bark-scale by the phase response of a first order low pass compared to the commonly used approximation given by Zwicker and Terhardt [ZWI80]. 55

Fig. 3.6: Fourth order Gammatone function compared to an auditory filter shape for intermediate signal levels for a narrow band signal at 1 kHz. 56

Fig. 3.7: Sixth order Gammatone function compared to an auditory filter shape for intermediate signal levels for a narrow band signal at 1 kHz.. 57

Fig. 3.8: Level-dependent tuning curves and their relation to an excitation pattern. 59

Fig. 3.9: Structure of a recursive filter realising a complex linear phase band pass. 62

Fig. 3.10: Amplitude response of one filter as defined by Eq. (3.21) with a centre frequency of 5 kHz and a bandwidth of 1200 Hz. 63

Fig. 3.11: Structure of a recursive filter realising a pair of linear phase band pass filters with 90° phase shift. 63

Fig. 3.12: Filter structure for efficient computation of subsampled band pass values. 65

Fig. 3.13: Amplitude response of a second order FDC-filter with a centre frequency of 5 kHz and a bandwidth of 1200 Hz. 67

Fig. 4.1: Computation of quality indicators in DIX. 71

Fig. 4.2: Structure of the perceptual model of DIX. 73

Fig. 4.3: Processing steps within the FDC filter bank. The middle row shows the pulse response at the different stages, and the lower row shows the corresponding frequency responses. 75

Fig. 4.4: Subsampling in DIX. 81

Fig. 4.5: Side lobes of a second order FDC-filter compared to an auditory filter shape. 83

Fig. 4.6: Excitation pattern at the onset of a 1 kHz sine tone. 84

Fig. 4.7: Impulse response of the FDC filter bank without both spectral smearing and temporal smearing. 84

Fig. 4.8: Impulse response of the FDC filter bank without temporal smearing, but with spectral smearing. 84

Fig. 4.9: Impulse responses of one filter channel (a) prior to and (b, c) after the spreading operation for two pulses of different energy (b: low energy, c: high energy). The shape of the pulse response after the complex spreading operation closely corresponds to the post-masking curves shown in [ZWI67]. 85

Fig. 4.10: Response to a 100 ms noise burst followed by silence. 86

Fig. 4.11: Trailing noise produced by the filter bank as a function of the duration of a 100 dB noise burst. Parameter is the elapsed time after switching off the noise burst. 87

Fig. 4.12: Trailing noise produced by the filter bank as a function of the elapsed time after switching off a 100 dB noise burst. Parameter is the duration of the noise burst. 87

Fig. 4.13: Level adaptation and time alignment within a typical perceptual measurement method. 88

Fig. 5.1: Generic block diagram of the measurement scheme 109

Fig. 5.2: Peripheral ear model and pre-processing of excitation patterns for the FFT-based part of the model. 110

Fig. 6.1: The five-grade impairment scale according to ITU-R BS.1116. 119

Fig. 6.2: The SDG-scale. 120

Fig. 6.3: Mapping from multiple MOVs to one ODG. 124

Fig. 7.1: Tolerance scheme proposed for the final ITU validation test (for an assumed confidence interval of +/-0.5 for all test items) 129

Fig. 7.2: Results for 40 items of the ISO/MPEG 1990 test (headphone presentation) [MPEG90] 130

Fig. 7.3: Results for 42 items (cascaded codecs) of the ITU 1993 test [ITU93]. 130

Fig. 7.4: Results for 35 items (commentary stereo) of the ITU 1993 test[ITU93] 130

Fig. 7.5: Results for all three tests put together 130

Fig. 7.6: Correlations between model predictions and SDGs for database 2 (91 items). 131

Fig. 7.7: Correlations between model predictions and SDGs for database 2 without the band-limited items (83 items). 131

Fig. 7.8: Results of version-3 of DIX for the subset of database 1. 132

Fig. 7.9: Results for the subset of database 1. 132

Fig. 7.10: Results for the subset of database 1, without the NICAM bass guitar item, which is probably corrupted. 132

Fig. 7.11: Correlations between model predictions and SDGs for the first half of database 2 (48 items). 133

Fig. 7.12: Correlations between model predictions and SDGs for the second half of database 2 (43 items). 133

Fig. 7.13: Correlations between model predictions and SDGs for the subset of database 1 (47 items). 134

Fig. 7.14: Correlations between model predictions and SDGs for database 2 and the subset of database 1 (138 items). 134

Fig. 7.15: Verification results of the FFT-based version recommended for standardisation (FftNnODG1). 140

Fig. 7.16: Verification results of the combined version recommended for standardisation(CombNnODG3). 140

Fig. 7.17: Verification results of the best filterbank-based model (FiltODG3). 140

Fig. 7.18: Verification results of the reference model (PAQM). 140

Fig. 7.19: Correlation between SDGs and ODGs for the validation data sets. 142

Fig. 7.20: Model predictions versus listening test results for all available data sets (left: advanced version, right: basic version). 145

Fig. 7.21: Distribution of the prediction errors for the advanced version of PEAQ. The outlier density is counted in clusters of 0.1 grades on the five-grade impairment-scale. 145

Fig. 8.1: Masking curves for the MOV RmsNoiseLoudAsymAvg with an assumed threshold value of 0.1 sone. The masker level is 70 dB. 148

Fig. 8.2: Masking curves for the MOV RmsNoiseLoudAsymAvg with an assumed threshold value of 0.2 sone. The masker level is 70 dB. 149

Fig. 8.3: Masking curves for the constellation noise-masking-noise for the MOV RmsNoiseLoudAsymAvg with an assumed threshold value of 0.2 sone. 150

Fig. 8.4: Pure-tone masking curve versus noise-masking-tone for the MOV RmsNoiseLoudAsymAvg with an assumed threshold value of 0.2 sone. The masker level is 90 dB. 150

Fig. 8.5: Additivity of masking between a maskers at 10 and 12 Bark for the constellation noise-masking-noise for the MOV RmsNoiseLoudAsymAvg with an assumed threshold value of 0.2 sone. The masker level is 70 dB. 151

Fig. 8.6: Additivity of masking between a maskers at 10 and 14 Bark for the constellation noise-masking-noise for the MOV RmsNoiseLoudAsymAvg with an assumed threshold value of 0.2 sone. The masker level is 70 dB. 152

Fig. 8.7: Temporal masking curves for the MOV RmsNoiseLoudAsymAvg with an assumed threshold value of 0.2 sone. The masker level is 10 dB below clipping. 153

14.3 Abbreviations

CCETT	Centre Commun d'Etudes de Télédiffusion et Télécommunications (French telecommunication provider)
CCIR	former name of the *ITU*
CRC	Communications Research Centre (Canada)
DB	database
DIX	Disturbance Index (name of a measurement method)
FhG	Fraunhofer Gesellschaft (in particular the *Fraunhofer Institute for Integrated Circuits*)
IRT	Institut für Rundfunktechnik GMBH
ITU	International Telecommunication Union
JND	just noticeable (level) difference
KPN	Royal PTT Nederland (Dutch telecommunication provider)
MOS	Mean Opinion Score
MOV	Model Output Value
OASE	Objective Audio Signal Evaluation (name of a measurement method)
ODG	Objective Difference Grade
PAQM	Perceptual Audio Quality Measure (name of a measurement method)
PEAQ	Perceptual Evaluation of Audio Quality (name of a measurement method)
PERCEVAL	Perceptual Evaluation (name of a measurement method)
POM	Perceptual Objective Measure (name of a measurement method)
SDG	Subjective Difference Grade

Curriculum Vitae

Person:

Name:	Thilo Volker Thiede
Date of birth:	10.11.1967 in Berlin/Germany

Jobs:

since 01.02.1999	Working in the DSP department of Tøpholm & Westermann ApS (Widex) in Copenhagen.
15.12.1994 - 31.10.1998	Research assistant at the Institute of Communications Engineering and Theory of Electricity of the Technical University of Berlin (Prof. Peter Noll). Working on auditory models and filter banks for the quality assessment of audio codecs. Participation in the ITU-R Task Group 10/4.
1991 - 1993	Student employee in a research project dealing with multichannel sound (“Orthophonie”) at the Institute for Communication Research of the Technical University of Berlin (Prof. Manfred Krause). Working on the design of microphone arrays with controllable directivity.

Education:

1994	Diploma thesis in the field of perceptual audio quality assessment.
1988 - 1994	Studying electrical engineering at the Technical University Berlin with main emphasis on acoustics and psychoacoustics. Other main subjects were telecommunications, electronics, and measurement techniques.
1987 - 1988	Practical at Siemens AG in Berlin.
1987	School leaving exam.
1980 - 1987	High school at “Waldoberschule” in Berlin.
1974 - 1980	Primary school at “Steuben Grundschule” in Berlin.