Psychoacoustics of detection of tonality and asymmetry of masking: implementation of tonality estimation methods in a psychoacoustic model for perceptual audio coding

Psychoakustik der Tonalitätswahrnehmung und der Maskierungsasymmetrie: Implementierung von Tonalitätsschätzverfahren in einem psychoakustischen Modell für die wahrnehmungsbasierte Audiocodierung

Der Technischen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg
zur
Erlangung des Doktorgrades
Doktor-Ingenieur (Dr.-Ing.)

vorgelegt von
Armin Taghipour
aus Rasht, Iran
Als Dissertation genehmigt
von der Technischen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung: 15.04.2016
Vorsitzender des Prüfungsorgans: Prof. Dr. Peter Greil
Gutachter: Prof. Dr.-Ing. Bernd Edler

em. Prof. Brian C. J. Moore, Ph.D, FMedSci, FRS, Dr. h.c.
Acknowledgements

This doctoral thesis has been developed during my employment at the International Audio Laboratories Erlangen (AudioLabs), which is a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and the Fraunhofer Institute for Integrated Circuits IIS. The financial support of these institutions and the Deutsche Forschungsgemeinschaft is gratefully acknowledged.

This work has benefited from a number of people too numerous to name individually. My greatest debt is to my advisor Prof. Dr. Bernd Edler for his continuous support. I wish to thank him for his academic supervision and appreciate his contributions of time, ideas, and inputs.

I would like to express my deepest gratitude to Prof. Brian C. J. Moore, Ph.D, who gave me an opportunity for a research collaboration with the Auditory Perception Group of the University of Cambridge. I am thankful for his time, inputs, and guidance. I also wish to thank all his staff for their support during my stay in Cambridge.

A special thank goes to the AudioLabs’ board members Dr. Bernhard Grill, Prof. Dr. Jürgen Herre, and Prof. Dr. Albert Heuberger, coordinators Dr. Frederik Nagel and Dr. Stefan Turowski, and administrative staff Ms. Tracy Harris, Ms. Day-See Riechmann, and Ms. Elke Weiland. I also wish to thank colleagues and students who collaborated with me in psychoacoustic studies or engineering projects presented in this thesis. I want to thank Ms. Masoumeh Amirpour and Ms. Shujie Guo for their contributions to the psychoacoustic experiments. I would like to thank Mr. Maneesh Jaikumar, Mr. Hao Chen, and Mr. Bharadwaj Desikan for their contributions to the engineering projects. I am thankful to all my colleagues at the AudioLabs and at the Fraunhofer Institute IIS, in particular, Mr. Alexander Adami, Mr. Stefan Bayer, Dr. Christian Borß, Mr. Soumitro Chakrabarty, Ms. Alexandra Craciun, Dr. Sascha Disch, Ms. Maya Eismann, Ms. Esther Fee Feichtner, Mr. Christian Helmrich, Dr. Yesenia Lacouture-Parodi, Dr. Nadja Schinkel-Bielefeld, Mr. Oliver Scheuregger, Mr. Fabian-Robert Stöter, and Mr. Nils Werner for their helpful comments, their technical and moral support, and for their friendship.

I am very thankful to the many subjects who participated in psychoacoustic experiments and listening tests. This thesis would not have been possible without their time and effort. I appreciate their help.

I would not have been able to write this thesis without the support of my family: my parents, Pourshad and Narjes Zari, who have supported me unconditionally in all stages of my education, my sister, Azadeh, who has been a faithful companion throughout my life especially during the last ten years in Germany, and my beloved partner, Eva, who has supported me during writing of this thesis and has always brought joy to my life even in difficult times. This thesis is dedicated to them.
Abstract

Perceptual audio coders exploit the masking properties of the human auditory system to reduce the data rate of the input audio signal. Commonly, the input signal is split into frames and decomposed into a number of subbands, such that it is divided into time-frequency segments. A psychoacoustic model (PM) is used to estimate the masking threshold evoked by the audio signal as a measure for the maximum inaudible quantization noise. Thus, it enables a perceptually suitable bit allocation for the time-frequency segments.

The effectiveness of the masking depends on whether the audio signal is tone-like or noise-like. This property is often referred to as “tonality.” Generally, tone-like sounds with no or slow envelope fluctuations and/or very regular envelope fluctuations are less effective maskers than noise-like sounds with strong, irregular, and more rapid envelope fluctuations. Perceptual audio encoders model this “asymmetry of masking” by means of tonality estimation methods.

Further, the effectiveness of the input signal in masking the quantization noise is likely to depend on the bandwidth and center frequency of the audio signal, but the exact dependency of the masked threshold on the masker’s bandwidth and center frequency had not previously been measured. The effect of bandwidth had been investigated in a few studies, however, only at medium center frequencies. A new psychoacoustic experiment was carried out to investigate masking when the noise masker had a narrower bandwidth than the noise probe, which is the most relevant case in perceptual audio coding. Statistical analysis of the data showed that the masked thresholds increased with increasing masker bandwidth and were lowest for medium center frequencies.

Another effective parameter is the duration of the stimulus. Both perceptually and physically, a certain amount of time is needed to identify the tonality or noisiness of a sound. Hence, to exploit the fact that quantization noise is masked more effectively by a noise-like signal than by a tone-like signal, the analysis frame length used in a perceptual coder must be sufficiently long. However, the required duration may vary depending on the center frequency and on the width of the subbands. A number of psychoacoustic experiments were conducted to investigate the duration required to discriminate a tone from a narrowband noise of the same center frequency. Duration thresholds decreased with increasing bandwidth and with increasing center frequency.
up to medium frequencies. An analysis of the stimuli and the perceptual data indicates that
the duration thresholds depended mainly on the detection of amplitude fluctuations in the noise
bursts.

In addition to the mentioned psychoacoustic studies, the present thesis deals with the
development of tonality estimation methods for psychoacoustic models. There have been a
few approaches to exploit the asymmetry of masking in conventional perceptual audio coding.
These methods were found to give quite accurate estimates of masked thresholds for the extreme
cases of high tonality or strong noisiness. However, a deeper analysis of the data on masking
asymmetries and the results of more recent psychoacoustic studies, including the ones from the
present thesis, provided opportunities for improved perceptual models. Hence, in the course of
this research, several measures were designed and implemented in a PM to estimate the degree
of tonality/noisiness.

The tonality measures were implemented in a filter bank-based PM, which was developed
in two versions. The initial version of the model was designed based on the Bark scale; the
magnitude frequency responses of the filters take into account the spreading of the simultaneous
masking. The second version of the PM was designed and computed based on the ERBN-number
scale and rounded exponential filters. The two versions of the model differ in the number, the
shape, the center frequencies, and the widths of the bandpass filters. The PM, in both its versions,
was developed such that a first approximation of the masking threshold of the input audio signal
in the subbands is derived from the filter outputs. Furthermore, a tonality estimation method
scales the masking thresholds depending on the estimated tonality/noisiness of the masking
audio signal.

Five different tonality estimation approaches are presented here. In an approach using
linear prediction, a high prediction gain was taken as an indication of high tonality. In another
approach, a discrete partial spectral flatness measure (PSFM) was calculated, which decreased
with increasing tonality. An amplitude modulation ratio (AM-R) was developed based on a
measure of the characteristics of amplitude modulations in the envelope of the subband signal;
low level of modulation in the envelope indicated high tonality. Another method was developed
based on the “auditory image model,” whereby high tonality was indicated by high intra-channel
correlations and high correlations between neighboring channels. Finally, in the fifth method,
fluctuations in the envelope of the subband signal were analyzed; slow fluctuations were interpreted
as high tonality. The results of several quality rating tests, which are documented in this thesis,
showed higher audio quality scores for PSFM and AM-R.

The two versions of the PM, including the PSFM tonality estimation method, were imple-
mented in a standard perceptual audio codec and were compared to a state-of-the-art psychoa-
coustic model. No significant difference was found between the two versions of the proposed
model. Also no significant differences were detected between the two models and a version of the
standard codec, in which additional coding tools were activated (e.g., “temporal noise shaping” and “spectral band replication”). However, when all three psychoacoustic models were applied to an identical standard coding scheme, both versions of the proposed model scored significantly higher than the conventional PM.

This shows that replacement of the transform-based psychoacoustic models by filter bank-based models with specially designed spectral analysis and frequency- and bandwidth-dependent tonality analysis has the potential to improve audio codecs.
Zusammenfassung

Wahrnehmungsbasierter Audiocodierverfahren nutzen die Eigenschaften des menschlichen auditorischen Systems aus, um die Datenrate des Audio-Eingangssignals zu reduzieren. Normalerweise wird das Eingangssignal gefenstert und in Teilbänder zerlegt, so dass daraus mehrere Zeit-Frequenz-Abschnitte entstehen, die in der Regel effizienter codiert werden können als das Eingangssignal. Ein psychoakustisches Modell (PM) berechnet eine Schwelle, die angibt, wie viel Quantisierungsrauschen durch das Eingangssignal maskiert werden kann. Das schließlich ermöglicht die gehörangepasste Quantisierung für die Zeit-Frequenz-Abschnitte.


Des Weiteren weisen die Ergebnisse früherer Studien darauf hin, dass die Effektivität des Eingangssignals bei der Maskierung des Quantisierungsrauschens von der Bandbreite und der Mittenfrequenz des Audiosignals abhängt, wobei die exakte Abhängigkeit der Maskierungsschwelle von der Bandbreite und der Mittenfrequenz noch nicht vollständig erforscht wurde. Die Rolle der Bandbreite wurde bisher nur für mittlere Mittenfrequenzen untersucht. Im Rahmen der vorliegenden Arbeit wurde ein psychoakustisches Experiment durchgeführt, bei dem das maskierende Rauschen eine schmalere Bandbreite hatte als das Testrauschen. Das ist der relevanteste Fall in der Audiocodierung. Die statistische Analyse der Daten zeigte, dass die Maskierungsschwelle mit steigender Bandbreite der Maskierer stieg, und dass sie für die mittleren Frequenzen am niedrigsten war.

Ferner zeigten psychoakustische Studien mit Stimuli verschiedener Dauer, dass die Wahrnehmung der Tonalität stark von der Stimulusdauer abhängt. Daher muss die Analysedauer in den psychoakustischen Modellen lang genug sein, um die Tatsache ausnutzen zu können, dass das Quantisierungsrauschen stärker von einem rauschartigen Signal maskiert wird, als von einem tonalen Signal. Allerdings könnte die notwendige Dauer abhängig von der Mittenfrequenz


Die Tonalitätsschätzverfahren wurden in einem Filterbank-basierten PM implementiert, das in zwei Versionen entwickelt wurde. Der ursprüngliche Entwurf des PM basierte auf der Bark-Skala. Die Betragsfrequenzgänge der Filter berücksichtigen die Ausbreitung der simultanen Maskierung, d.h. die Mithörschwelle. Eine weitere Version des Modells wurde auf Basis der ERB_n-Skala entworfen. Die beiden Modelle unterscheiden sich hinsichtlich der Anzahl, der Form, der Mittenfrequenzen und der Bandbreiten der Bandpassfilter. Das PM in beiden Varianten ist so konstruiert, dass eine grobe Schätzung der Maskierungsschwelle bereits auf der Basis der Filterausgänge gewonnen werden kann. Ein Tonalitätsschätzverfahren skalierter dann diese Maskierungsschwelle mit dem Tonalitätsgrad der Teilbandsignale.


Die beiden im Rahmen dieser Arbeit entworfenen PM – inkl. PSFM – wurden in ein Standard-Audiocodierverfahren eingesetzt und in Hörversuchen mit einem Standard-PM verglichen. Die Analyse der Daten zeigte signifikant höhere Qualitätsratings für die beiden neuen Modelle im

Die Ergebnisse der Hörversuche zeigen, dass die im Rahmen dieser Arbeit entwickelten Filterbank-basierten psychoakustischen Modelle, sowie die hier präsentierten Tonalitätsschätzverfahren, ein vielversprechender Ansatz für die Weiterentwicklung der wahrnehmungsbasierten Audiocodierverfahren darstellen.
Contents

Acknowledgements i

Abstract ii

Zusammenfassung v

Abbreviations 3

1 Introduction 5
   1.1 Perceptual audio coding 5
   1.2 Asymmetry of masking 6
   1.3 Challenges 6
   1.4 Overview of chapters 8

2 Psychoacoustic basics 10
   2.1 Human peripheral auditory system 10
   2.2 Thresholds of audibility of tone or noise pulses 14
   2.3 Masking 16
   2.4 Critical bands 20
   2.5 Asymmetry of masking 24

3 Psychoacoustic studies 27
   3.1 Experiment I: dependence of perception of tonality on frequency, bandwidth and duration 27
   3.2 Experiment II: thresholds for perception of rapid changes in tone bursts 34
   3.3 Discussion of Experiment I and II 41
   3.4 Experiment III: masked threshold of noise in presence of narrower bands of noise 43

4 The Psychoacoustic model 51
   4.1 Perceptual audio coding 52
   4.2 Psychoacoustic models for perceptual audio coding 56
   4.3 The model 58
   4.4 An ERB_N design of the PM 61
4.5 The MUSHRA test ................................................................. 64

5 Tonality estimation methods ............................................... 67

5.1 A linear prediction method (Predictor) .......................... 67
5.2 Partial spectral flatness measure (PSFM) ...................... 70
5.3 Amplitude modulation ratio (AM-R) ......................... 78
5.4 Auditory image correlation (AIC) ............................... 83
5.5 Temporal envelope rate (TE-R) ................................. 87
5.6 Conclusions ................................................................. 92

6 Comparison of the PM to a conventional (MPEG) PM ...... 94

6.1 Coding scheme .............................................................. 94
6.2 MUSHRA test .............................................................. 97
6.3 Discussion ................................................................. 99

7 Conclusions ................................................................. 100

References ................................................................. 112
## Abbreviations

<table>
<thead>
<tr>
<th>Abbr.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAC</td>
<td>advanced audio coding</td>
</tr>
<tr>
<td>acf</td>
<td>autocorrelation function</td>
</tr>
<tr>
<td>AI</td>
<td>auditory image</td>
</tr>
<tr>
<td>AIC</td>
<td>auditory image correlation</td>
</tr>
<tr>
<td>AIM</td>
<td>auditory image model</td>
</tr>
<tr>
<td>AM</td>
<td>amplitude modulation</td>
</tr>
<tr>
<td>AM-R</td>
<td>amplitude modulation ratio</td>
</tr>
<tr>
<td>ANOVA</td>
<td>analysis of variance</td>
</tr>
<tr>
<td>ANSI</td>
<td>American National Standards Institute</td>
</tr>
<tr>
<td>BM</td>
<td>basilar membrane</td>
</tr>
<tr>
<td>BMM</td>
<td>basilar membrane motion/movement</td>
</tr>
<tr>
<td>CB</td>
<td>critical bandwidth</td>
</tr>
<tr>
<td>DAB</td>
<td>digital audio broadcasting</td>
</tr>
<tr>
<td>DFT</td>
<td>discrete Fourier transform</td>
</tr>
<tr>
<td>DVD</td>
<td>digital versatile disc</td>
</tr>
<tr>
<td>ERB</td>
<td>equivalent rectangular bandwidth</td>
</tr>
<tr>
<td>ERBN</td>
<td>normal-hearing ERB of the auditory filters</td>
</tr>
<tr>
<td>ETSI</td>
<td>European Telecommunications Standards Institute</td>
</tr>
<tr>
<td>FFT</td>
<td>fast Fourier transform</td>
</tr>
<tr>
<td>FIR</td>
<td>finite impulse response</td>
</tr>
<tr>
<td>GUI</td>
<td>graphical user interface</td>
</tr>
<tr>
<td>hcl</td>
<td>half-wave rectification, compression, and lowpass filtering</td>
</tr>
<tr>
<td>HE</td>
<td>high-efficiency</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>IEC</td>
<td>International Electrotechnical Commission</td>
</tr>
<tr>
<td>IIR</td>
<td>infinite impulse response</td>
</tr>
<tr>
<td>IMDCT</td>
<td>inverse modified discrete cosine transform</td>
</tr>
<tr>
<td>IRN</td>
<td>iterated rippled noise</td>
</tr>
<tr>
<td>ISO</td>
<td>International Organization for Standardization</td>
</tr>
<tr>
<td>LSD</td>
<td>Fisher’s protected least significant difference test</td>
</tr>
<tr>
<td>MDCT</td>
<td>modified discrete cosine transform</td>
</tr>
<tr>
<td>mp3</td>
<td>MPEG-1 Layer III</td>
</tr>
<tr>
<td>MPEG</td>
<td>moving picture experts group</td>
</tr>
<tr>
<td>MUSHRA</td>
<td>multiple stimuli with hidden reference and anchor</td>
</tr>
<tr>
<td>NAP</td>
<td>neural activity pattern</td>
</tr>
<tr>
<td>PCM</td>
<td>pulse code modulation</td>
</tr>
<tr>
<td>PM</td>
<td>psychoacoustic model</td>
</tr>
<tr>
<td>PNS</td>
<td>perceptual noise substitution</td>
</tr>
<tr>
<td>PSFM</td>
<td>partial spectral flatness measure</td>
</tr>
<tr>
<td>roex</td>
<td>rounded-exponential</td>
</tr>
<tr>
<td>rms</td>
<td>root mean square</td>
</tr>
<tr>
<td>SBR</td>
<td>spectral band replication</td>
</tr>
<tr>
<td>SFM</td>
<td>spectral flatness measure</td>
</tr>
<tr>
<td>SL</td>
<td>sensation level</td>
</tr>
<tr>
<td>SNR</td>
<td>signal-to-noise ratio</td>
</tr>
<tr>
<td>SPL</td>
<td>sound pressure level</td>
</tr>
<tr>
<td>STI</td>
<td>strobed temporal integration</td>
</tr>
<tr>
<td>TDAC</td>
<td>time-domain aliasing cancellation</td>
</tr>
<tr>
<td>TECC</td>
<td>temporal envelope cross-correlation</td>
</tr>
<tr>
<td>TE-R</td>
<td>temporal envelope rate</td>
</tr>
<tr>
<td>TETC</td>
<td>temporal envelope threshold crossing</td>
</tr>
<tr>
<td>TNS</td>
<td>temporal noise shaping</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Perceptual audio coding

Perceptual audio coders exploit the masking properties of the human auditory system to reduce the requirements for storage and transmission of digital audio via compression (Bosi et al., 1997; Brandenburg and Bosi, 1997; Brandenburg and Stoll, 1994; Painter and Spanias, 2000). Familiar examples of perceptual coders are mp3\(^1\) (ISO/IEC 11172-3, 1993; Brandenburg and Stoll, 1994; Stoll and Brandenburg, 1992), MPEG-AAC\(^2\) (Bosi et al., 1997; Brandenburg and Bosi, 1997) and digital audio broadcasting (DAB) (ETSI EN 300 401, 2006).

Perceptual audio encoders work in the following way. The input audio signal, which is typically pulse code modulated (PCM) (Oliver et al., 1948), is split into brief segments, called “frames.” The frame length may be fixed or it can vary according to the characteristics of the signal. Within each frame, by means of a filter bank or a transform, the signal is decomposed into a number of frequency bands, called “subbands.” The signal in each subband is represented as a sequence of binary digits (bits). The greater the number of bits, the lower is the quantization noise relative to the audio signal. The most efficient coding, i.e. the minimum number of bits required to represent the audio signal in the subband without audible artifacts, is achieved if the quantization noise lies just below the masked threshold. A psychoacoustic model (PM) is used to estimate the masked threshold of the quantization noise for each subband and frame. This enables a proper bit allocation for each subband and frame.

\(^1\)MPEG-1 Layer III
\(^2\)Advanced audio coding
1.2 Asymmetry of masking

Generally, the quantization noise in a given subband covers the whole spectral range of that subband, but the audio signal often has a narrower bandwidth. The effectiveness of the audio signal in masking the quantization noise depends on its bandwidth and center frequency and especially on whether it is tone-like or noise-like (Hall, 1997; Johnston, 1988a; Painter and Spanias, 2000). In the coding literature, this property is often referred to as “tonality.” Generally, a tone-like masker is less effective than a noise-like masker of equal power/energy (Gockel et al., 2002; Hall, 1997; Hellman, 1972). This phenomenon is called “asymmetry of masking.”

Hellman (1972) showed that a noise masks a tone more effectively than a tone masks a noise. Hall (1997) extended these results to the case where both the signal and masker were noise bands centered at 1000 Hz, but their bandwidths differed. For probe bandwidths exceeding that of the masker, the situation that is predominantly relevant for perceptual coders (where the quantization noise acts as the probe and the signal is the masker), thresholds decreased with increasing signal bandwidth and with decreasing masker bandwidth. Hall suggested that, for this situation, the results might be explained in terms of the temporal structure of the stimuli, specifically the extent to which the amplitude modulation pattern of the signal resembled that of the masker. Verhey (2002) performed a similar study using a center frequency of 2000 Hz and found similar results.

In perceptual audio coders, the audio content may include complex tones as well as noise-like sounds. Gockel et al. (2002) measured the asymmetry of masking between harmonic complex tones and wideband noise. Noise was a stronger masker.

One can summarize these results by saying that sounds with no or slow envelope fluctuations and/or very regular envelope fluctuations are less effective maskers of a noise (with equal or greater bandwidth) than sounds with strong, irregular and more rapid envelope fluctuations.

1.3 Challenges

1.3.1 Lack of relevant psychoacoustic data

The effectiveness of the audio input in masking the quantization noise is likely to depend on its bandwidth and center frequency, but the exact dependence of the masked threshold on the audio signal’s bandwidth and center frequency has not previously been measured. Although data from Hall (1997) and Verhey (2002) give some information about the role of bandwidth, in these studies only medium center frequencies (1000 and 2000 Hz) have been investigated.

Additionally, the stimuli of these studies have bandwidths of up to 256 Hz. However, since in perceptual audio coding the subbands are designed based on fractions of the bandwidth of the
hypothetical auditory filters, data for stimuli with larger bandwidths were needed. Further, it was desirable to investigate the role of the bandwidth with an experimental design in which the widths of the noise stimuli vary systematically relative to the bandwidths of the auditory filters.

In order to investigate the effect of local bandwidth and center frequency, a psychoacoustic experiment (Experiment III) was designed and carried out. Masking was investigated when the noise masker had a narrower bandwidth than the noise probe, which is the most relevant case in perceptual audio coding. The role of bandwidth expressed in Hz, as well as in fractions of the bandwidth of the auditory filters, and the effect of the center frequency was investigated. The experiment and a discussion of the results is presented in this thesis.

Another effective parameter is the duration of the stimulus. Both perceptually and physically, a certain duration is required to identify the tonality of a sound; a very short burst of narrowband noise and a tone burst of the same duration and with the same center frequency sound very similar and are physically very similar (Bos and de Boer 1966; Moore 2012; Viemeister and Plack 1993). The main physical difference is that the envelope of a narrowband noise fluctuates over time, but for short-duration signals this fluctuation can be hard to detect (Stone et al., 2008). Hence, to exploit the fact that quantization noise is masked more effectively by a noise-like audio signal than by a tone-like audio signal, the analysis length in a psychoacoustic model must be sufficiently long. However, the required duration may vary depending on the center frequency and on the width of the auditory filters. As the width of a subband is increased, the amplitude fluctuations in a noise-like audio signal become more rapid (Bos and de Boer 1966) and this would be expected to decrease the duration required to distinguish a tone from a noise. In the course of the present research, three psychoacoustic experiments have been carried out to investigate the duration required to discriminate a tone from a narrowband noise of the same center frequency. Two of the experiments (Experiments I and II) will be presented in detail in this thesis. The main goal of these experiments was to estimate duration thresholds as a function of center frequency and of the bandwidth of the noise.

1.3.2 Tonality/noisiness estimation methods

There have been a few attempts to implement asymmetry of masking in psychoacoustic models of perceptual audio coders (Johnston, 1988b; Hall, 1998; van de Par et al., 2005). Under the notion of “perceptual entropy” (Johnston, 1988a) this phenomenon has been introduced to the audio coding standards. Examples of tonality estimation methods are the “spectral flatness measure” (Johnston, 1988b) and the “chaos measure” (Brandenburg, 2002; ISO/IEC 11172-3, 1993). These state-of-the-art approaches were found to give quite accurate estimates of masked thresholds for the extreme cases of high tonality or strong noisiness. However, Painter and

For the most part of this thesis, not the total bandwidth of a signal is the focus, but the spectral extent of a spectral component in the signal.
Spanias (2000) suggested that a deeper understanding of masking asymmetry may provide opportunities for improved perceptual models. Specifically, in harmonic complex sounds or in speech signals, different spectral regions can progress over time in different ways in terms of tonality/noisiness.

With this background, the second emphasis of this doctoral research was to investigate, introduce, and implement measures for estimation of the degree of tonality/noisiness in psychoacoustic models used in perceptual audio encoders. This manuscript presents a number of these methods (e.g., Chen et al., 2014; Taghipour et al., 2010, 2014b). They model the effects of masker tonality in a filter bank-based PM.

The bandpass filters of the PM were designed to approximate relevant frequency characteristic of the human auditory system. Thus, an initial estimate of masked threshold for a noise masker for each filter was obtained from the short-term signal level at the output of that filter. Different approaches were proposed to adjust (reduce) the estimated thresholds depending on the estimated tonality of the masker. In a “linear prediction” approach (Taghipour et al., 2010), a high prediction gain, indicating a sharp spectral peak, was taken as indicating high tonality. In another approach (Taghipour et al., 2014b), which, in some aspects, was similar to that of Johnston (1988b,a), a discretized partial spectral flatness measure was calculated for each frame and filter; this measure decreased with increasing tonality. A third approach (Chen et al., 2014) was based on a measure of the amount of amplitude modulation in the envelope of the time-domain representation of the signal; low level of modulation in the envelope indicated high tonality. Another method was developed based on the “auditory image model,” whereby high tonality was indicated by high intra-channel correlations and high correlations with the neighboring band (Taghipour et al., 2016a). In a fifth method, fluctuations in the envelope of the filter outputs were analyzed; slow fluctuations were interpreted as high tonality (Taghipour et al., 2016a).

1.4 Overview of chapters

The thesis is structured in seven chapters. In Chapter 2, some of the psychophysiological aspects and characteristics of the human peripheral auditory system are described. A few temporal aspects of hearing are discussed. A brief discussion of temporal and spectral aspects of masking follows.

In Chapter 3 two psychoacoustic studies (Experiment I and II) are presented in which the duration thresholds for distinguishing tone and noise bursts were investigated. A third psychoacoustic study (Experiment III) is presented which investigated the effect of the center frequency and bandwidth on masking when the noise masker had a narrower bandwidth than the noise probe.

Chapter 4 describes basics of psychoacoustic models in perceptual audio coding. A psychoa-
acoustic model is presented which is used and developed further in the course of this thesis. Two versions of the PM are presented, which differ from each other in the characteristics of their filter banks. Moreover, a subjective quality assessment test is described, which is commonly used for evaluation of the quality of coded audio signals.

Chapter 5 deals with various methods for the estimation of tonality/noisiness of the time-frequency segments of the audio signal. These methods were developed in the course of this doctoral thesis. They use different measures for estimating the degree of tonality/noisiness of the signal or the degree of amplitude fluctuations in the envelope of the audio input. Different studies are presented which compare the methods to each other using the quality assessment test described in Chapter 4.

The psychoacoustic model in both its versions was implemented in a standard encoder. Details of the implementation are discussed in Chapter 6. The two versions of the proposed PM were compared to a conventional PM. The results of the study are shown and discussed in detail.

Chapter 7 concludes the research and proposes further areas of research in both fields of experimental psychoacoustics and development of methods for estimation of tonality/noisiness in perceptual audio codecs.

Portions of this work have been published before. The psychoacoustic studies have been presented in [Taghipour et al. (2012, 2013a, 2014a, 2016c,b)]; the psychoacoustic model and the tonality estimation methods have been partially presented in [Taghipour et al. (2016a, 2010, 2013b, 2014b) and Chen et al. (2014)]. These references are cited repeatedly throughout the thesis; however, since an exhaustive citation is cumbersome, it has been avoided whenever possible. For all other references, an attempt has been made to cite carefully and exhaustively.
Chapter 2

Psychoacoustic basics

Sound is an oscillation in pressure which propagates through a medium (typically air) and is perceived by the human auditory system. Psychoacoustics deals with the relationship between the physical characteristics of sounds and how those sounds are perceived (Moore 2012).

As mentioned in Chapter 1, psychoacoustic models used in perceptual audio codecs try to model some psychoacoustical processes of the human auditory system. For example, they particularly estimate thresholds of audibility, e.g., by considering absolute threshold of hearing, and by estimating temporal and spectral masking thresholds for frames of the input audio signal. The goal of a compression algorithm is to reduce the data rate, while retaining as much quality (i.e., acoustic fidelity) as possible. On this account, it is important to take a closer look at the peripheral auditory system and some of the psychoacoustic phenomena.

This chapter gives a brief overview of the physiology and psychophysics of the peripheral auditory system. A number of temporal auditory phenomena are discussed. Further, temporal and spectral masking and the concept of critical bands are explained. Most importantly, asymmetries of masking (differences between noise and tone as masker or probe) are dealt with. For this, several reference studies are discussed.

2.1 Human peripheral auditory system

The human peripheral auditory system works as a preprocessing system (Fastl and Zwicker 2007). It sends the preprocessed signal to the higher neural levels. Most of the modeling for the audio coding purposes consists of either modeling this peripheral structure or some psychoacoustic phenomena, and much less of modeling the higher neural processing. The peripheral auditory system can be divided in three sections: outer, middle, and inner ear, each of which will be described briefly in the following.
2.1.1 Outer ear

As shown in Figure 2.1 (Fastl and Zwicker 2007), the most external part of the human auditory system is the outer ear. It is composed of the pinna and the auditory canal (external auditory meatus) (Moore 2012). The outer ear’s function is to collect sound energy and to transmit this energy through the ear canal to the eardrum, the so called tympanic membrane (Fastl and Zwicker 2007).

The pinna plays an important role in sound localization (Moore 2012; Schandry 2011). The outer ear canal protects the eardrum and the middle ear from damage and enables the inner ear to be positioned very close to the brain, allowing a short travel time for the action potentials (Fastl and Zwicker 2007). It also exerts a strong influence on the frequency response of the hearing organ; it increases the hearing sensitivity for the frequency region near 3 kHz (Fastl and Zwicker 2007).

![Figure 2.1: A schematic drawing of the outer, middle and inner ear; redrawn from Fastl and Zwicker 2007.](image-url)
2.1.2 Middle ear

The region extending from the eardrum (the inner end of the outer ear) to the oval window (a
door to the inner ear) is called the middle ear. As depicted in Figure 2.1 (Fastl and Zwicker,
2007), the middle ear consists of three small bones, or ossicles, named malleus, incus, and stapes.
Malleus is attached to the eardrum, and stapes lies on top of the oval window (Fastl and Zwicker,
2007; Moore, 2012). Air pressure changes in the outer ear are transmitted through the middle ear
by the ossicles and are transferred into motions of the fluid in the inner ear (Fastl and Zwicker,
2007; Moore, 2012). This is the most important function of the middle ear. It basically acts as
an impedance-matching device. This transmission is most efficient at middle frequencies (Moore,
2012).

2.1.3 Inner ear

Not all sections and mechanisms of the inner ear will be discussed in detail, but only the ones
which are most relevant to perceptual audio coding. The cochlea is shaped like the spiral shell of
a snail (see Figure 2.1). It is filled with fluids, and has bony rigid walls (Moore, 2012). Along its
length, it is divided by two membranes, Reissner’s membrane and the basilar membrane (BM)
(Moore, 2012; Schandry, 2011).

As explained in Section 2.1.2, the oscillations in the middle ear are brought to the oval
window, the entrance to the inner ear, by the stapes footplate (Fastl and Zwicker, 2007). A
pressure difference is applied across the BM which causes it to move (Moore, 2012). Basilar
membrane movement (BMM) leads to a motion of cochlea’s fluids.

The basilar membrane is stretched from the basal end of the cochlea (the side of the oval
window) along the spiral shape deep to the apical end of the cochlea. There is a small gap
(opening) between this end of the BM and cochlea’s bony wall. This gap is named helicotrema
(Schandry, 2011). Along the length of the BM, two outer chambers of the cochlea, the scala
vestibuli and the scala tympani, are located. The motions of the fluids in the cochlea happen
in these two chambers. The path can be described as follows: oval window, scala vestibuli,
helicotrema, scala tympani, and finally round window. The cochlea is closed, so the pressure
introduced to the oval window has to be sent out of the cochlea in the form of some pressure at
the round window. For a better understanding, a schematic of an unrolled cochlea is drawn in
Figure 2.2.

For an incoming sinusoidal signal, as the oscillation progresses along the BM, the amplitude
of the wave increases first and then decreases, rather abruptly, with increasing distance from the
basal end (Moore, 2012; Schnupp et al., 2011). For simplicity, if the introduced signal is a pure
tone, the placement of the peak amplitude of the oscillation corresponds to its frequency. The
mechanism which leads to this can be explained as follows. At the basal end, the BM is relatively
narrow and stiff, while at the apical end it is wider and much less stiff (Moore (2012) Schnupp et al. (2011)). So, the mechanism can be explained as a counteraction of two factors: stiffness of the BM vs. inertia of the fluid. While the stiffness of the BM decreases along the path from base to apex, the inertia increases with the amount of the fluid. This is shown in Figure 2.2 where high intensity colors indicate a high level of stiffness or inertia. Thus, based on the frequency of the spectral components of the incoming sound, at some point the movement can be redirected instead of steadily passing the helicotrema. High frequency components cannot move lots of fluids, but can be redirected by the narrow, but stiff, part of the BM near the basal end; whereas low frequency components can move lots of fluids, and because of their slow oscillation they can be first redirected in a much less stiff part of the BM, near the apex. This frequency-placement representation in the inner ear (and in the neural response) is referred to as “tonotopy” (Fletcher, 1940).

Because of the tonotopy, the analysis of an incoming audio signal in the cochlea, as well as in the further neural coding, can be modeled by means of a transform, e.g. Fourier transform (Oppenheim and Schafer 2010 Proakis and Manolakis 2007), or with help of a filter bank (Oppenheim and Schafer 2010 Proakis and Manolakis 2007). The human auditory system analyzes an audio signal not only as a temporal event, but also based on its spectral components. Thus, perceptual audio encoders analyze the signal, in one of the first steps, by means of a transform or a filter bank (Bosi 2003).
2.2 Thresholds of audibility of tone or noise pulses

Thresholds for detection of tone or noise pulses have been of interest in many studies (Exner, 1875; Exner, 1876; Fastl and Zwicker, 2007; Fletcher, 1940; Moore, 2012). The general assumption is that these thresholds are dependent on level, duration, frequency and bandwidth of the burst.

2.2.1 Absolute threshold

In a silent background, in presence of no other signals, there is a minimum sound pressure level at which a tone is detected as an incoming/present signal. This is called the threshold of audibility in quiet, or absolute threshold. This means that, despite the physical existence of a tone (or noise), it might be inaudible for the human auditory system due to its low intensity level. Absolute threshold is depicted in Figure 2.3 (Fastl and Zwicker, 2007). Perceptual audio coding systems model the absolute threshold in different ways (Bosi, 2003; Painter and Spanias, 2000).

2.2.2 Temporal effects

Moore (2012) mentioned that it has been known for many years (Exner, 1876) that the absolute thresholds of sounds depend upon duration. This discussion can be divided into two subjects: temporal resolution and temporal integration.

![Figure 2.3: The absolute threshold of hearing in quiet. SPL stands for sound pressure level.](image-url)
2.2.2.1 Temporal resolution

Several studies have investigated temporal resolution of the human auditory system which refers to the ability of the auditory system to respond to rapid changes in the envelope of a sound over time (Viemeister and Plack 1993). Due to differences in methods and objectives, they measured, to some extent, different aspects of this phenomenon; for instance, the duration of a brief gap detected between two stimuli (Eddins et al. 1992; Exner 1875; Fitzgibbons and Wightman 1982; Penner 1977; Shailer and Moore 1983), time interval between two clicks of time-reversed stimuli (Green 1973; Ronken 1970), or the least amount of detectable fluctuations/modulation in the envelope of a stimulus (Bacon and Viemeister 1985; Dau et al. 1997b; Moore and Glasberg 2001; Viemeister 1979).

Moore (2012) and Viemeister and Plack (1993) gave good summaries of the literature. The following discussion only concentrates on the studies which will help to discuss the results of the psychoacoustic studies presented in Sections 3.1 and 3.2.

Exner (1875, 1876) showed that two noise signals separated from each other by a gap of less than approximately 2 ms cannot be perceived as two separate signals. Other researchers tried to reduce the salience of spectral cues during temporal acuity measurement tasks. Penner (1977) used broadband white noise, since a short gap should not change the long-term magnitude spectrum of such stimuli. Ronken (1970) noted that the long-term magnitude spectrum of a stimulus should not change when it is “time-reversed.” Patterson and Green (1970) and Green (1973) made use of specific stimuli1 which were identical in spectra, however, different in phase. Most of these studies discovered a temporal resolution of about 2-3 ms.

Eddins et al. (1992), Fitzgibbons and Wightman (1982), and Shailer and Moore (1983) investigated gap detection thresholds in bands of noise. Using octave-band noise, Fitzgibbons and Wightman (1982) found an improvement of temporal resolution with increasing center frequency. Also Shailer and Moore (1983) found a similar effect of frequency: gap detection thresholds decreased monotonically with increasing center frequency. Shailer and Moore (1983) suggested that the temporal response of the auditory filters is responsible for the limited temporal resolution at lower frequencies. Further, they argued that the results of experiments using wideband stimuli (Penner 1977) might reflect the ability of subjects to make use of information available in higher frequency regions (Shailer and Moore 1983). Eddins et al. (1992) found no significant effect of center frequency given conditions of equivalent absolute bandwidth (in Hz). With respect to the role of the “bandwidth,” both latter studies showed that the gap thresholds decreased monotonically with increasing bandwidth (Eddins et al. 1992; Shailer and Moore 1983).

1Huffman sequence (Huffman 1962; Patterson and Green 1970).
2.2.2.2 Temporal integration

Another aspect of temporal analysis of sound in the human auditory system is temporal integration. Exner (1876) showed that the absolute thresholds of hearing depended upon duration. Furthermore, he suggested that there should be some kind of integration over time that helps subjects to achieve a perception/sensation of the tonal characteristics of a tone. That is, a minimum number of periods (16 to 17) are necessary in order to perceive a tone as such (Exner, 1876). He suggested that this number must be roughly the same for the tones of different frequencies. This was further investigated by Hughes (1946) and Garner and Miller (1947). Hughes (1946) noted that the threshold of audition rose as the time of presentation of the stimulus was decreased. He suggested that the change in level (in dB) should be linearly proportional to the reciprocal of the duration of the signal. Garner and Miller (1947) investigated the relation between the duration of a tone and its threshold intensity. In the presence of noise as a masker, they measured the intensity of the tone at the threshold of audibility. Garner and Miller (1947) concluded that there should be a linear relation between the logarithm of duration and threshold level (in dB). They suggested temporal integration of up to 200 ms.

Based on Zwislocki (1960) and Penner (1972), Moore (2012) argued that there have been doubts on whether there is an integration of energy over time in the human auditory system, or if rather the neural activities are integrated. Viemeister and Wakefield (1991) suggested “multiple looks” could be responsible for, or at least aid in, decreasing the threshold of audibility.

In the course of this doctoral thesis, the role of duration in the perception of tonality/noisiness is tested experimentally. Two such experiments are presented in Chapter 3, where both the role of “duration” and the effects of the frequency and bandwidth of the stimuli are discussed.

2.3 Masking

We all experience in our everyday life that the presence of a loud sound can make other sounds partially, or totally inaudible. This phenomenon is called auditory masking. Masking has been defined as “the process by which the threshold of audibility for one sound is raised by the presence of another (masking) sound” (ANSI S1.1, 1994; Moore, 2012). Three aspects of masking are of particular interest to perceptual audio coding: (1) temporal and (2) spectral masking, and (3) masking asymmetries (Section 2.5). In the following, it will be distinguished between the “masker” and the (masked) “probe.”

2.3.1 Temporal and spectral masking

As illustrated in Figure 2.4, from a temporal point of view, masking can be divided into two categories: simultaneous or nonsimultaneous. The focus of discussion on simultaneous masking is...
its spectral spread, while backward and forward masking, as forms of nonsimultaneous masking, are weaker versions of the same effect.

### 2.3.1.1 Simultaneous masking

Simultaneous masking occurs when the masker and the probe are present at the same time. Similar to the absolute threshold and duration thresholds of distinction between noise and tone, simultaneous masking is also dependent on duration. Almost all previous arguments and discussions regarding the effects of duration also hold for simultaneous masking. The following discussion is based on the assumption that the masker and the probe are presented for long enough (e.g. over 200 ms) so that the temporal effects of duration are saturated.

Although [Mayer](1876) discussed the idea of masking and published some qualitative observations, the first “quantitative” study on masking seems to be that of [Wegel and Lane](1924). [Wegel and Lane](1924) investigated the amount of masking as a function of level of the masker and frequency of the probe. The frequency of the masker was kept constant. The resulting masking curves showed a spread of masking over neighboring frequencies which decayed with growing spectral distance from the frequency of the masker on both sides.

Due to the use of pure tones as probe and masker, beats occurred ([Mayer](1894) [Wegel and Lane](1924) [Egan and Hake](1950)). Hence, the tip of the masking curves showed dips, since beats...
Figure 2.5: These curves illustrate the difference between the masked audiograms of a pure tone of 400 Hz and of a narrowband of noise (90 Hz wide, centered at 410 Hz) with each presented at an overall SPL of 80 dB. The solid and the dashed lines show the masking curves of the pure tone and the narrowband noise, respectively. A tone probe was used in both cases. Comparison of the two curves shows how much the masking pattern of a pure tone is complicated by beats, harmonics, and difference tones \cite{Egan1950}. The dips are due to beats. These masking audiograms are based upon the average data of 5 normal-hearing subjects; redrawn from \cite{Egan1950}.

make it easier to hear the signal in the presence of the masker, and they produce irregularities in the acoustical pattern. The shapes of the masking curves were later confirmed by \cite{Egan1950}. Masking tends to vanish when the spectral distance from the frequency of the masker is large enough. \cite{Egan1950} showed that using a narrowband noise masker overcomes the problem of beating. Figure 2.5 shows the masked thresholds of a pure tone probe in presence of a pure tone or a band of noise as masker \cite{Egan1950}. The results indicated differences in the levels of the masking curves of tone and noise maskers (which is related to Section 2.5).

\cite{Hawkins1950} investigated masked thresholds of pure tone probes in the presence of white noise maskers. Figure 2.6 illustrates some examples of masked threshold contours for white noise maskers of different levels. A series of good examples for masking...
Figure 2.6: Idealized masked threshold contours. These curves show the monaural thresholds for pure tones when masked by various levels of ideal white noise having uniform power spectral density; redrawn from [Hawkins and Stevens, 1950].


[Fletcher and Munson, 1937] carried out several experiments on the relation between masking and perceived loudness. Masked thresholds and masking contours of different loudness levels were estimated experimentally. One important outcome was that, for a band of noise of a given center frequency, there was a specific “critical bandwidth,” for which the noise would produce about the same masking as a band of greater width, provided the intensity per cycle was the same and that the masked tone was in the same frequency band as the noise ([Fletcher, 1940] [Fletcher and Munson, 1937]). This will be discussed further in Section 2.4. For further reading, [Fastl and Zwicker, 2007], [Fletcher, 1940], and [Moore, 2012] are recommended. They give good reviews of the literature on simultaneous masking and the resulting auditory patterns.

[^2]: A masker which causes uniform masking contours; i.e., parallel horizontal masking curves.
[^3]: A bandpass noise with a large bandwidth.
2.3.1.2 Backward masking

When the probe precedes the masker, it can still be partially (or totally) masked by it. This effect is called backward masking, pre-stimulatory masking, or pre-masking; see Figure 2.4. Fastl and Zwicker (2007) mentioned that the effect is understandable if one realizes that each sensation does not exist instantaneously, but requires a build-up time to be perceived. A quick build-up time for a loud masker, and a slower build-up time for the faint probe can be the reason for this effect. Backward masking seems to be only effective for a very short period of time, i.e., up to 20 ms (Fastl and Zwicker 2007).

However, Moore (2012) argued that the amount of backward masking obtained depends strongly on how much practice the subjects have received, and that practiced subjects often show little or no backward masking (Miyazaki and Sasaki 1984; Oxenham and Moore 1994).

2.3.1.3 Forward masking

Forward masking happens when the masker has already been vanished or been switched off, i.e., when the masker is no longer physically existent; nevertheless, it still produces masking (Fastl and Zwicker 2007). As shown in Figure 2.4, the strength of forward masking decays with time until it becomes zero after 100-200 ms (Fastl and Zwicker 2007; Moore 2012).

Although the basis of forward masking is still not completely understood, the following factors may contribute (Moore 2012): (1) ringing of the BM’s response to the masker, (2) short-term adaptation in the auditory system, (3) persisting neural activity evoked by the masker, (4) possible persisting inhibition evoked by the masker, and (5) activation of the efferent system by the masker (leading to reduced gain of the active mechanism). For a further discussion on the effect and basis of the forward masking, see Moore (2012). Because of its importance, forward masking is frequently modeled in psychoacoustic models of perceptual audio encoders. This holds also for the model presented in Chapter 4.

2.4 Critical bands

As mentioned before, Fletcher (1940) and Fletcher and Munson (1937) investigated masked thresholds of a pure tone in presence of a band of noise masker as a function of the bandwidth of the masker. The results showed that when the bandwidth of a masker is increased (assuming constant noise power density), the masked threshold of a pure tone probe at the center of the noise increases only up to a certain “critical bandwidth.” Beyond that, although the bandwidth increases, as does the total power of the noise, these additional noise portions have no masking effect on the probe. Fletcher (1940) suggested that there should be a number of hypothetical auditory filters in (the BM of) the human peripheral auditory system.
The concept of the critical bandwidth (CB) is explained by Moore (2012) as the following: when trying to detect a probe in a noise background, the listener is assumed to make use of an auditory filter with a center frequency close to that of the probe. This filter passes the probe but removes a great deal of the noise. Only the components in the noise that pass through the filter have any effect in masking the probe (Moore, 2012).

In perceptual audio coding, the frames of an incoming signal are decomposed into spectral subbands. The characteristics of the subbands are based on the idea of the auditory filters. The bandwidth and the spacing of the subbands are related to the critical bandwidths. Two different measures/scales are commonly used in perceptual audio coding and in auditory models, which are discussed in the following.

2.4.1 Bark

Zwicker (1956, 1961) and Zwicker et al. (1957) measured the bandwidth of the auditory filters experimentally in various contexts (“absolute threshold,” “phase,” “loudness summation,” and “masking”). Zwicker and Terhardt (1980) suggested the following formula for the critical bandwidth as a function of the center frequency \( f \):

\[
CB = 25 + 75[1 + 1.4(0.001f)^2]^{0.69}
\]  

(2.1)

where CB and \( f \) are expressed in Hz. Another, simple approximation of the critical bandwidth is (Fastl and Zwicker, 2007):

\[
CB = \begin{cases} 
100 & \text{if } f < 500 \\ 0.2f & \text{else}
\end{cases}
\]

(2.2)

with CB and \( f \) in Hz. These approximations of the critical bandwidth are depicted in Figure 2.7 in black (Equation 2.1) and blue (Equation 2.2). Zwicker and Terhardt (1980). The so called “Bark number scale” with unit Bark is defined by the function (Fastl and Zwicker, 2007):

\[
\text{Bark number} = 13 \arctan(0.00076f) + 3.5 \arctan((\frac{f}{7500})^2)
\]

(2.3)

where \( f \) is in Hz. The auditory-perception-adapted Bark number scale is shown in Figure 2.8 in blue. This scale helps in the design of subbands, e.g., for the spacing of the center frequencies, or it can be used as a spectral scale (instead of frequency) for analysis or plotting.

\[\text{In memory of Heinrich Barkhausen, a German physicist and acoustician.}\]
Figure 2.7: Critical bandwidth, or (normal-hearing) equivalent rectangular bandwidth, as a function of frequency (in kHz). The blue curve is a simple approximation for Bark bandwidths; redrawn from Fastl and Zwicker (2007), Moore (2012), and Zwicker and Terhardt (1980).

2.4.2 ERB$_N$

Patterson (1974, 1976) investigated the shape of the auditory filters with the so called notched-noise method. A wideband noise with a deep notch with sharp edges was used to mask a tone probe. The notch was centered on the tone, and the masked threshold was measured as the width of the notch was increased (Patterson, 1976). By means of similar stimuli, Jurado and Moore (2010); Moore et al., (1990) investigated the bandwidth of the auditory filters with low center frequencies (center frequencies below 500 Hz). In contrast to the data of Zwicker et al. (1957), they found that the bandwidth of the auditory filters further decreased with decreasing center frequency.

Glasberg and Moore (1990) documented detailed experimental results on the shape and bandwidth of the auditory filters as a function of the center frequency of the stimuli. Since the bandwidth was expressed in the form of the “equivalent rectangular bandwidth” of the filters, the...
name “ERB<sub>N</sub>”<sup>5</sup> was chosen (Glasberg and Moore, 1990; Moore, 2012). An equation describing the value of the ERB<sub>N</sub> as a function of center frequency <i>f</i> is (Glasberg and Moore, 1990; Moore, 2012):

\[ \text{ERB}_N = 24.7(0.00437f + 1) \]  

(2.4)

where ERB<sub>N</sub> and <i>f</i> are expressed in Hz. Similar to the Bark scale, an “ERB<sub>N</sub> number scale” was introduced based on ERB<sub>N</sub> estimations, which has the units Cam (Glasberg and Moore, 1990; Moore, 2012):

\[ \text{ERB}_N \text{ number} = 21.4 \log_{10}(0.00437f + 1) \]  

(2.5)

where again <i>f</i> is expressed in Hz. The red curve in Figure 2.7 shows estimates of the auditory filters bandwidth based on ERB<sub>N</sub>. For middle frequencies, Bark and ERB<sub>N</sub> estimates are very close to each other. The major difference is at lower frequencies, where ERB<sub>N</sub> estimates lead to narrower bandwidths. The ERB<sub>N</sub> number scale is shown in Figure 2.8 in black.

The slopes of the magnitude response of the auditory filters change with the level of the stimuli. In general, they have an asymmetric shape. As an example, Glasberg and Moore (2000)\footnote{N stands for young normal-hearing subjects.}

---

Dissertation, Armin Taghipour
investigated the shape of the filters as a function of the level. For a detailed discussion on auditory filters, their shapes and bandwidths, see (Moore 2012).

Both “Bark” and “ERB\_N” systems are used in the course of this doctoral thesis in the following chapters. For the psychoacoustic studies in Chapter 3, the ERB\_N system is used. The initial bandpass filters of the psychoacoustic model and the tonality estimation methods in Chapters 4 and 5 are mostly based on the Bark system. A second version of the PM, presented in Chapter 4, is based on ERB\_N.

2.5 Asymmetry of masking

As explained in Chapter 1, psychoacoustic models used in perceptual coders (see Chapter 4) should take into account an effect called the “asymmetry of masking.” Hellman (1972) showed that noise masks tone more effectively than tone masks noise. The asymmetry was up to 20 dB. In the case when a tone signal was the masker, loudness of masked and unmasked noise bands of equal width were matched using the method of adjustment (Hellman, 1972). The frequency of the tone and the center frequency of the noise bursts were 1 kHz. Among other parameters, the effect of the bandwidth of the noise was investigated. When the probe was a tone, at low intensity level, wideband and narrowband noise masked the tone similarly (Hellman, 1970). At high intensity levels, wideband maskers showed stronger masking than narrowband noise (Hellman, 1970). When the tone acted as a masker (Hellman, 1972), increasing the noise bandwidth decreased the masking strength of the tone.

Hall (1997) extended these results to the case where both the probe and masker were noise bands, but their bandwidths differed. The effect of bandwidth was investigated. Masked thresholds were highest, and were roughly constant, for probe bandwidths smaller than or equal to that of the masker. For probe bandwidths exceeding that of the masker, thresholds decreased with increasing probe bandwidth and with decreasing masker bandwidth. Hall suggested that two different approaches are needed when modeling the asymmetry of masking. For probe bandwidths smaller than or equal to the masker bandwidth, the long-term average spectrum can be used to predict masked thresholds, based on the assumption that the threshold corresponds to a constant probe-to-masker ratio at the output of the auditory filter centered at the probe frequency. When the probe bandwidth is greater than the masker bandwidth, the temporal structure of the probe should also be taken into account (Hall, 1997). Hence, time-domain models should improve the estimation of the level at the masked threshold.

This was confirmed by Verhey (2002), who replicated Hall’s study and came to the same conclusions with respect to bandwidth. Verhey (2002) was able to simulate the results using a model incorporating a modulation filterbank (Dau et al., 1997b).

Krumbholz et al. (2001) investigated the masking of random noise by iterated rippled noise
(IRN) and of IRN by random noise. IRN is generated by adding a delayed copy of a noise to the original and repeating this operation a number of times. IRN has a noise-like waveform but with a degree of temporal regularity. They found that noise masks IRN more effectively than IRN masks noise. They also concluded that temporal processing needs to be taken into account to predict the asymmetry of masking (Krumbholz et al., 2001).

Gockel et al. (2002) measured the asymmetry of masking between harmonic complex tones and wideband noise as a function of masker level. Both tone and noise stimuli were bandpass filtered. The components of the complex tone were added either in cosine phase, giving a waveform with a high peak factor, or in random phase, i.e. incoherently. Fundamental frequency was either 62.5 or 250 Hz. The rms⁶ level of the masker ranged from 40 to 70 dB SPL⁷.

The noise was a more effective masker than the complex tone. The degree of the asymmetry varied between 10 and 28 dB, with the maximum asymmetry occurring for the highest SPL of 70 dB, the lowest fundamental frequency of the complex tone (62.5 Hz), and the cosine-phase tone. Gockel et al. (2002) suggested the active mechanism in the cochlea was responsible for the dependence on level. However, sound level only played a role when the lowest fundamental frequency was used and the signal with components in cosine phase masked the noise. Gockel et al. (2002) concluded that for noise probes, which are the most relevant with regard to perceptual audio coding, since it is the quantization noise at the output of the decoder that needs to be masked by the audio signal, the masked thresholds depend strongly on whether the masker is a noise or a tone complex.

Generally, one can summarize these results by saying that tone-like sounds are less effective maskers of quantization noise than noise-like sounds. Since this is not a binary distinction, the “degree” of tonality or noisiness plays a crucial role. A better way of describing the results would be by saying that sounds with no or slow envelope fluctuations and/or very regular envelope fluctuations are less effective maskers of a noise (with equal or greater bandwidth) than sounds with strong, irregular and more rapid envelope fluctuations.

As mentioned in Chapter 1 in perceptual audio coding, the effectiveness of the audio input signal in masking the quantization noise is likely to depend on its local bandwidth and center frequency, but the exact dependence of the masked threshold on the audio signal’s bandwidth and center frequency has not previously been measured. Although data from Hall (1997) and Verhey (2002) give some information about the role of bandwidth, in these studies only medium center frequencies (1000 and 2000 Hz) were investigated. Besides, both studies investigated the role of bandwidth expressed in Hz. Bandwidths from 0 to 256 Hz were tested. This is helpful in interpreting the effects in terms of the degree of fluctuation in the envelope of the signals, or in terms of amplitude modulation pattern (Viemeister, 1979; Dau et al., 1997a). However, in audio coding the subbands are designed based on fractions of the bandwidth of the hypothetical

---

⁶Root mean square.
⁷Sound pressure level.
auditory filters. The bandwidths of the auditory filters at medium and high frequencies are higher than 256 Hz. Not only is it necessary to collect data for stimuli with bandwidths beyond 256 Hz, but also, in the experimental design, it might be beneficial to relate the width of the stimuli to the bandwidth of the auditory filters.

Experiment III in Chapter 3 investigated the effect of bandwidth and center frequency of the stimuli on masking. Masking was investigated when the noise masker had a narrower bandwidth than the noise probe (the most relevant case for perceptual audio coding). The effect of the center frequency was explored. Further, the role of bandwidth was investigated expressed in Hz, as well as in fractions of the bandwidth of the auditory filters.

Chapter 5 introduces a number of tonality estimation methods for psychoacoustic models of perceptual audio codecs which are developed in the course of this doctoral thesis. They are compared to each other with objective methods, as well as with subjective quality tests.
Chapter 3

Psychoacoustic studies

In the course of this doctoral thesis, several psychoacoustic studies have been carried out, three of which are presented in this chapter. Two phenomena were investigated: “perception of tonality” and “masking.” The focus of the discussion will be on the effects of duration, bandwidth and center frequency of the stimuli.

3.1 Experiment I: dependence of perception of tonality on frequency, bandwidth and duration

In this section and in Section 3.2, two psychoacoustic studies are presented which investigate duration dependent thresholds of perception, or detection, of tonality.

3.1.1 Theory

Both perceptually and physically, a certain amount of time is required to identify the tonality/noisiness of a sound; a very short burst of narrowband noise and a tone burst of the same duration and with the same center frequency sound very similar and are physically very similar (Bos and de Boer 1966; Moore 2012; Taghipour et al. 2016b; Viemeister and Plack 1993). The main physical difference is that the envelope of a narrowband noise fluctuates over time; but for short-duration signals this fluctuation can be hard to detect (Stone et al. 2008).

Hence, to exploit the fact that quantization noise is masked more effectively by a noise-like audio signal than by a tone-like signal, the analysis frame used in auditory models must be sufficiently long. However, the required duration may vary depending on the center frequency and on the width of the auditory filters. In general, as the width of a filter is increased, the amplitude fluctuations in a noise-like signal become more rapid (Bos and de Boer 1966) and this would be expected to decrease the duration required to distinguish a tone from a noise. No
previous study was found that had measured the duration required to discriminate a tone from a narrowband noise of the same center frequency. Hereafter, this duration is referred to as a “duration threshold.” The main goal of the two experiments presented here and in Section 3.2 was to estimate duration thresholds as a function of center frequency and of the bandwidth of the noise. The data were intended to be useful in improving the design of perceptual coders by indicating whether, in the stage of tonality estimation, the analysis frame length should vary with subband center frequency and bandwidth (Taghipour et al., 2013a, 2014a, 2016b).

In Experiment I, duration thresholds were measured for isolated, short noise and tone bursts (Taghipour et al., 2013a). Imagine a narrow band noise burst with center frequency $f_c$, and a pure tone burst of frequency $f_c$, both of the same duration. As long as the stimuli are long enough, they are easily distinguishable for normal-hearing listeners. However, when the stimulus pairs’ duration decreases, at some point (i.e., duration threshold) they become psychoacoustically identical. Thus, in the present experiment, in each condition, the independent variable was the duration. As the duration of the stimuli increased, the acoustical differences became more audible. As the duration decreased, the signals were not easily to tell apart.

Further, we wanted to investigate the role of the center frequency and bandwidth. As mentioned above, in the course of a run[^1] the duration of the stimuli was chosen to be the independent variable. Frequency and bandwidth were changed systematically from condition to condition. It should be mentioned that the sound pressure levels of the stimuli were kept constant during the whole experiment, since level differences were not a focus of the study.

3.1.2 Method

3.1.2.1 Stimuli

In each trial, two stimuli were presented consecutively with a silent gap of 800 ms between them. One of the following pairs was selected randomly for each trial: tone-tone, tone-noise, noise-tone, or noise-noise (Taghipour et al., 2013a, 2016b). As shown in Figure 3.1 the stimuli were gated with raised-cosine ramps. Since the duration of the stimuli was the independent variable in a run, the lengths of the ramps also varied. For overall durations (including ramps) up to 5 ms, the duration of each ramp was 1 ms; for overall durations between 5 and 10 ms, the duration of each ramp was 2 ms; beyond that, 3 ms ramps were used. As an example, the window function

[^1]: That is, all the successively presented trials in a condition which lead to a decision about the threshold.
Figure 3.1: A window function of 15 ms with raised cosine ramps for gating the stimuli.

in Figure 3.1 was defined as follows:

\[
  w(n) = \begin{cases} 
    0.5 - 0.5 \cos \left( \frac{\pi(n-0.5)}{0.003sf_s} \right) & \text{if } 0 \leq n < 0.003sf_s \\
    1 & \text{if } 0.003sf_s \leq n < 0.012sf_s \\
    0.5 + 0.5 \cos \left( \frac{\pi(n+0.5-0.012sf_s)}{0.003sf_s} \right) & \text{if } 0.012sf_s \leq n < 0.015sf_s \\
    0 & \text{else} 
  \end{cases}
\]  

(3.1)

where \( f_s \) is the sampling frequency, here 48000 Hz. \( s \) stands for “second.”

Figure 3.2 illustrates how such a window slightly spread the magnitude spectrum of short pulses. For relatively long pulses (e.g. of 100 ms) this effect was acoustically negligible. However, for very short durations, the spectra of the stimuli broadened, and this might have led to audible effects (Taghipour et al., 2013a). This is discussed in more detail in Section 3.3.

Figure 3.2: Magnitude spectrum of a tone burst resulting from windowing a pure tone of 1 kHz and duration 15 ms.
The estimated level of the stimuli at the eardrum was 75 dB SPL (see 3.1.2.3 for details of calibration). This level was chosen on the basis of pilot experiments, as it led to a comfortable loudness, given the short duration of the stimuli at the threshold for discrimination (duration thresholds were typically below 18 ms).

The noise bandwidth was specified on the ERB$_N$-number scale (Glasberg and Moore, 1990), which has units Cams (Moore, 2012). This scale was used because of its direct link to the bandwidth of the auditory filters. The bandwidth was 0.25, 0.5, or 1 Cams. The center frequencies were 345, 754, 1456, and 2658 Hz, corresponding to approximately 8.5, 13.5, 18.5, and 23.5 Cams. The tone stimuli were generated deterministically, and their waveforms started at a positive-going zero-crossing. Different random noise stimuli were generated for each trial. For this purpose, white Gaussian noise was digitally filtered in the time domain by 4th order bandpass Butterworth filters. The total energy of each noise burst was adjusted to equal that of the tone burst.

**3.1.2.2 Procedure**

Subjects were asked whether the two signals were the same or different. They were told that the signals would be the same on half of the trials and different on the other half. A response was counted as correct when the subject responded “same” and the two signals were the same (either tone-tone or noise-noise) or they responded “different” and the signals were different (either tone-noise or noise-tone). Otherwise, the response was counted as incorrect. Feedback in the form of a green or red light was provided after each trial via a graphical user interface indicating a correct or an incorrect response, respectively.

A 3-down/1-up adaptive method controlled the procedure (Levitt, 1971; Wetherill and Levitt, 1965). Based on the outcome of a prior study (Taghipour et al., 2012) and a pilot experiment, a fixed step size of 1 ms was chosen. Figure 3.3 illustrates the staircase method. Two randomly chosen conditions were presented interleaved, as suggested by Levitt (1971). The starting duration was between 13 and 21 ms. A run continued until 11 reversals had occurred. The average duration over the 6th to the 11th reversal points was taken as the threshold estimate for that run. For the example in Figure 3.3 the estimated threshold would be

\[
\text{Threshold} = \frac{\sum_{i=6}^{11} d_i}{6} = \frac{(11 + 10 + 11 + 10 + 13 + 12)}{6} = 11.2
\]

where \(d_i\) stands for the duration of the stimuli at the \(i\)-th reversal point. There was a single threshold estimate for each condition and subject.
Figure 3.3: An example of the 3-down/1-up staircase method used in the experiment: the horizontal axis shows the reversal points. “+” and “o” indicate correct or incorrect responses, respectively.

3.1.2.3 Apparatus

The listening room had a background noise level of 25 dBA. Calibration of levels at the eardrum was done by means of an artificial head (KEMAR, GRAS, Holte, Denmark). This was done for a tone burst of 250 Hz and 75 dB SPL. Stimuli were computed digitally with a sample frequency of 48 kHz and a resolution of 16 bits. An RME Babyface digital-to-analog converter (Haimhausen, Germany) was used for playback. As for subjective quality tests of monophonic coded audio signals [ITU-R BS.1534-1, 2003], stimuli were presented diotically via a pair of open electrostatic Stax SR-507 headphones with a SRM-600 driver unit (Saitama prefecture, Japan).

3.1.2.4 Subjects

Thirty self-reported normal-hearing subjects were tested. The “modified Thompson Tau test” and “Dixon’s Q test” were used to check for outliers. Both tests revealed outliers for one or more conditions for three subjects, and all data for these three subjects were excluded from further analyses. Thus, the final statistical analysis was based on thresholds for 27 subjects: 16 males and 11 females. They were aged between 20 and 49 years (mean 28 years, median 27 years).

3.1.2.5 Design

Each subject was tested in three sessions of about 20-25 minutes each, separated by at least half a day. Prior to the main experiment, subjects read a page of instructions. They had 2-3 minutes of training by listening to signal pairs for which they were informed as to which pairs were identical and which were different. A training session of two runs with center frequencies 345 and 1456 Hz and bandwidths of 0.5 and 1 Cam, respectively, followed. This preparation/training phase took 15-20 minutes. In the main experiment, the conditions were presented in a random order.
CHAPTER 3. PSYCHOACOUSTIC STUDIES

3.1.3 Results

The variability of the thresholds was approximately proportional to the threshold values, so geometric mean thresholds across subjects were calculated. The mean duration thresholds across the 27 subjects and their 95% confidence intervals (assuming normally distributed data) are illustrated in Figure 3.4. Duration thresholds are plotted on a logarithmic scale as a function of center frequency (also on a logarithmic scale) with bandwidth as parameter.

A two-way repeated-measures analysis of variance (ANOVA) was conducted on the logarithm of the duration thresholds with factors bandwidth and frequency. There were significant effects of noise bandwidth \( F(2, 52) = 106.4, p < 0.001 \) and frequency \( F(3, 78) = 56.2, p < 0.001 \). There was no significant interaction; \( F(6, 156) = 1.0, p > 0.1 \).

Post hoc pairwise comparisons (here and elsewhere based on Fisher’s protected least significant difference, LSD, test) were conducted to investigate the effects of bandwidth and center frequency. Duration threshold decreased with increasing bandwidth (all pairs \( p < 0.001 \)) and with increasing frequency up to 1456 Hz (all pairs \( p < 0.001 \)). No significant difference was found between the two highest center frequencies (\( p = 0.405 \)).

Figure 3.4: Results of Experiment I (Taghipour et al., 2016b): means and 95% confidence intervals of the duration thresholds across the 27 subjects are plotted as a function of frequency with noise bandwidth in Cams as parameter. Duration thresholds are shown in ms.

Dissertation, Armin Taghipour
When bandwidths were expressed in Cams, as above, the results showed significant effects of both bandwidth and center frequency. However, for a given bandwidth in Cams the absolute widths (in Hz) of the noise bands increased with increasing center frequency. The thresholds may have been strongly influenced by the bandwidth in Hz, since this bandwidth determines the average number of amplitude fluctuations per second, and hence determines how many fluctuations occur within the stimulus duration (Taghipour et al., 2016b). Figure 3.5 shows the geometric mean duration thresholds as a function of the absolute bandwidth in Hz. Each center frequency is represented by a different symbol. A one-way repeated-measures ANOVA was conducted with factor bandwidth in Hz. A significant effect was found; $F(11, 286) = 39.8, p < 0.001$. Post hoc comparisons showed that the duration threshold decreased significantly with increasing bandwidth, whenever the two bandwidths that were compared differed by at least 65 Hz ($p < 0.05$).

Figure 3.5 shows that much of the variability in duration thresholds is accounted for by the width of the noise bands in Hz. In a regression analysis, the percentage of the variance in the data accounted for by bandwidth was 93.2%. This is consistent with the idea that the task was performed by detecting amplitude fluctuations in the noise and that the duration had to be sufficient for a detectable amplitude fluctuation to occur. However, bandwidth in Hz does not account for all of the variability in the data; the curves for the different center frequencies do not overlap completely (Taghipour et al., 2016b).

Figure 3.5: Geometric mean duration thresholds across the 27 subjects of Experiment I plotted as a function of bandwidth in Hz with frequency as parameter (Taghipour et al., 2016b). Duration thresholds are shown in ms.
3.2 Experiment II: thresholds for perception of rapid changes in tone bursts

3.2.1 Theory

Experiment I (Section 3.1) provided baseline data for a relatively simple situation; it dealt with very brief, isolated bursts of up to 50 ms duration. In a perceptual coder, however, the decision about tonality has to be made for each frame, and the decision for a given frame can be influenced by the stimuli in preceding and following frames. For example, it is easier to detect a brief irregularity embedded in a regular or steady sound than it is to detect a brief regularity embedded in an irregular sound (Chait et al., 2007; Cusack and Carlyon, 2003; Pollack, 1968; Rupp et al., 2005). In Experiment II, duration thresholds were measured for tone and noise segments embedded within longer tone pulses (Taghipour et al., 2014a). This corresponds to a situation where performance is expected to be relatively good, since the presence of an embedded noise burst is indicated by a transition from regularity to irregularity.

In Experiment I, some of the thresholds were under 10 ms, which gave the bursts rise and fall periods of under 3 ms. Regardless of the shape of the window, such a rapid, high jump in the intensity (or energy) of the signal is perceived as a click. The shorter the signal the more click-like a tone burst sounds. Any rapid, transient jump in temporal shape of the signal introduces high frequency components (Bos and de Boer, 1966; Moore, 2012). The use of embedding in Experiment II reduced spectral broadening effects associated with short-duration stimuli, although spectral “splatter” could potentially provide a cue for detection of the noise burst (Taghipour et al., 2016b); the possible influence of spectral splatter is discussed later.

With this background, in Experiment II rapid changes/modifications in the middle of stationary tone bursts were investigated. This is closer to the phenomena observed commonly in natural audio signals where effects such as amplitude and frequency modulation or frequency shifting can happen for short segments. The two bursts of sound of the following experiments differed only in a short segment in their temporal center (Taghipour et al., 2014a). As mentioned above, these stimuli are relevant to the design of perceptual coders, since the decision about tonality has to be made not for isolated frames but for frames embedded within other frames.

3.2.2 Method

3.2.2.1 Stimuli

As explained above, while in Experiment I the duration of the stimuli was changed (Taghipour et al., 2013a), in the present study the duration was constant, namely 400 ms (Taghipour et al., 2014a). Both stimuli were gated using a window function with raised-cosine ramps of 30 ms.
Figure 3.6 shows such a window, an example of the resulting tone burst of 500 Hz and its magnitude spectrum. As clearly shown in Figure 3.6 (c) the physical/acoustical nature of the tone does not significantly suffer from the windowing process since the burst has a proper length.

One of the stimuli was a pure tone. The other was the same except for a short segment in its temporal center, which was replaced by a narrowband noise of the same center frequency, generated in the same way as for Experiment I.

To avoid discontinuities, cross-fading was used in the transition from tone to noise and back. The cross-fading windows had raised-cosine ramps. For overall noise durations up to 5 ms, the duration of each cross-fading ramp was 40% of the overall noise duration. For longer noise durations...
Figure 3.7: Cross-fading: the solid line shows the window function for eliminating the middle part of the sinusoid; the dashed line shows the window function for introducing narrowband noise into it.

segments, the duration of the cross-fading ramps was kept constant at 2 ms. An example of the cross-fading windows is depicted in Figure 3.7.

The noise segment that was actually picked (from a long narrowband noise burst) was determined based on the extent to which the waveforms of the tone and noise were similar in amplitude and phase within the two cross-fading ranges. The mean-squared difference between the time waveforms was used as a measure of similarity. As a result, the noise was faded in and out almost in-phase with the sinusoid (for more details, see Taghipour et al. (2014a)). This served to minimize spectral splatter. A segment of a noise-included burst is illustrated in Figure 3.8. The figure shows only ca. 60 ms of a 400-ms burst. The transition sections are smooth. In the middle, the signal has a narrow-band noise shape and on both sides it has the character of a 500 Hz pure tone. In the following, the overall duration of the middle noise segment (including half of the cross-fade ramps) is referred to as the “noise duration,” the transition sections as the “cross-fading ranges,” and the resulting noise-included burst as the “noisy burst,” respectively.

Figure 3.8: A segment of a burst of 400 ms in which the middle part is replaced by narrow-band noise of center frequency 500 Hz. On the left and right sides, the burst has the form of a tone of 500 Hz. Noise duration was 36.5 ms.
For short noise durations the noisy burst sounded similar to a pure tone. However, as the noise duration increased the middle part sounded different, e.g. it was perceived as amplitude modulation or it produced a sense of roughness. For very long noise segments, frequency modulations were also heard.

The estimated level of the stimuli at the eardrum was 65 dB SPL. This level was chosen to lead to a comfortable overall loudness, given the relatively long overall duration of the stimuli (in comparison to Experiment I). As for Experiment I, the noise bandwidth was 0.25, 0.5, or 1 Cams and the center frequencies were 345, 754, 1456, and 2658 Hz.

### 3.2.2.2 Procedure

A two-interval two-alternative forced-choice method was used. As shown in Figure 3.9, subjects were asked to indicate which of the two bursts was a “pure tone.” Feedback in the form of a green or red light was provided after each trial via a graphical user interface indicating a correct or an incorrect response, respectively. This can be observed in Figure 3.9. The independent variable in a run was the noise duration. A hybrid staircase procedure was used. A run started with two “easy” trials using a rather long noise segment. Up to the 3rd reversal point a 1-down/1-up method [Dixon, 1965] [Levitt, 1971] was used to achieve a rapid approach to the duration threshold. After that, a 3-down/1-up method [Levitt, 1971] [Wetherill and Levitt, 1965] was used to estimate the 79% point on the psychometric function. The step size was 5 ms up to the second reversal point, then 2 ms up to the fourth reversal point, and then 1 ms until 10 reversals were obtained. The duration threshold for a run was calculated as the arithmetic mean duration at the last 6 reversal points. Two runs were obtained for each condition and subject. An example of such a run (for frequency 2658 Hz and bandwidth 0.25 Cam) is depicted in Figure 3.10. Various parameters of the test can be observed in the figure.

![Figure 3.9](image-url)

Figure 3.9: The graphical user interface (GUI) for the experiment: (a) beginning of a run, and feedback after a (b) correct or (c) incorrect response.

---

Dissertation, Armin Taghipour
3.2.2.3 Apparatus

The experiment was carried out in a room which had a background noise level of 21 dBA. An artificial head (KEMAR, GRAS, Holte, Denmark) was used for calibration of levels. Stimuli were computed digitally with a sample frequency of 48 kHz and a resolution of 16 bits, converted to analog form using a Lawo DALLIS 941/83 digital-to-analog converter (Rastatt, Germany), and presented diotically via Sennheiser HD 650 headphones (Wedemark, Germany).

3.2.2.4 Subjects

Fifteen normal-hearing subjects participated in the experiment. Audiometric thresholds were measured for frequencies up to 14 kHz using a software-based audiometer (SelfscreeningAudioMeter V1.32, HörTech GmbH, Oldenburg, Germany) and Sennheiser HDA 200 headphones. For all frequencies, the absolute thresholds for all subjects were in the range -20 to +20 dB hearing level (ISO 389-7, 2005).

The “modified Thompson Tau test” and “Dixon’s Q test” were used to check for outliers. Both tests revealed outliers for one or more conditions for four subjects, and all data for these four subjects were excluded from further analyses (Taghipour et al., 2016b). Thus, the final statistical analysis was based on thresholds for 11 subjects: eight males and three females. They were aged between 23 and 37 years (mean 28 years, median 26 years).
3.2.2.5 Design

Each subject was tested in three sessions of about 30-35 minutes each, carried out on different days. Prior to the main experiment, subjects read a page of instructions. A training session of two runs with center frequencies 345 and 1456 Hz and bandwidths of 0.5 and 1 Cam, respectively, followed. This preparation/training phase took 20 minutes. During the main experiment, a break was given after every third run (i.e. after approximately 10-12 minutes). The conditions were presented in a random order.

3.2.3 Results

Again (as in Section 3.1.3), the variability in the thresholds was proportional to the mean. Thus, geometric mean thresholds across subjects were calculated. Figure 3.11 shows the mean duration thresholds and their 95% confidence intervals (assuming that the data were normally distributed) on a logarithmic scale as a function of center frequency (logarithmically scaled abscissa). A two-way repeated-measures ANOVA was conducted on the logarithm of the duration thresholds.

![Figure 3.11: Results of Experiment II (Taghipour et al., 2016b): means and 95% confidence intervals of the duration thresholds across the 11 subjects are plotted as a function of frequency with noise bandwidth in Cams as parameter. Duration thresholds of the noise segments are shown in ms.](image)

Figure 3.11: Results of Experiment II (Taghipour et al., 2016b): means and 95% confidence intervals of the duration thresholds across the 11 subjects are plotted as a function of frequency with noise bandwidth in Cams as parameter. Duration thresholds of the noise segments are shown in ms.
with factors bandwidth and center frequency. There were significant effects of bandwidth \([F(2, 20) = 153.6, p < 0.001]\) and center frequency \([F(3, 30) = 55.4, p < 0.001]\). There was no significant interaction; \([F(6, 60) = 0.9, p > 0.1]\).

Post hoc Fisher’s LSD tests showed that duration thresholds decreased with increasing bandwidth (all pairs \(p < 0.001\)) and that they decreased with increasing frequency, but only up to 1456 Hz (all pairs \(p < 0.01\)). There was no significant difference between thresholds for the two highest frequencies (\(p = 0.2\)).

Figure 3.12 shows the geometric mean duration thresholds as a function of the bandwidth in Hz. Each center frequency is represented by a different symbol. A one-way repeated-measures ANOVA with factor bandwidth, in Hz, showed a significant effect; \([F(11, 110) = 43.0, p < 0.001]\). Bandwidth in Hz accounted for 91.1% of the variability in the thresholds. Post hoc LSD tests showed that duration thresholds decreased significantly with increasing bandwidth whenever the two bandwidths differed by at least 206 Hz (\(p < 0.05\)). As for experiment 1, the data showed that the duration thresholds were strongly influenced by bandwidth in Hz, but that there was also an effect of center frequency. Specifically, for a given bandwidth in Hz duration thresholds were higher (worse) for the 2658-Hz center frequency than for the other center frequencies (Taghipour et al., 2016b).

![Figure 3.12: Geometric mean duration thresholds across the 11 subjects of Experiment II plotted as a function of bandwidth in Hz with frequency as parameter (Taghipour et al., 2016b).](image-url)
3.3 Discussion of Experiment I and II

In the context of perceptual audio coding, Experiment I represents a situation where, e.g., a (transient) cymbal or a harpsichord attack occurs and it is important to know how fast the human auditory system is able to perceive the regular structure of the harpsichord compared to the more irregular noisy structure of a cymbal input. In Experiment II, however, transitions from a regular structure to an irregular structure were investigated; e.g., voiced versus unvoiced speech, or in the case of vibration. This is more related to the fluctuations in the envelope of a signal, or to the modulation patterns. Before comparing the results of the two experiments, a discussion of spectral broadening is needed.

3.3.1 Effects of spectral broadening

Decreasing the duration of the stimuli, as in Experiment I, would have led to a broadening of their spectra (Bos and de Boer, 1966; Moore, 2012). However, since the two stimuli that were being compared always had the same duration and were gated with the same window function, this spectral broadening would not be expected to provide a discrimination cue. If anything, the spectral broadening would make it more difficult to use any cues associated with differences in the spectra of the tone burst and the noise burst (Taghipour et al., 2016b).

In Experiment II, the noise bursts were embedded within longer tone bursts. In this situation, spectral broadening associated with the short noise burst embedded within the pure tone might have provided a cue for discrimination of the noisy stimulus from the pure tone. However, the characteristics of the individual noise bursts were chosen to minimize spectral “splatter” effects. Also, while the duration thresholds obtained in Experiment II varied with bandwidth and center frequency in a similar way to the thresholds obtained in Experiment I, most of the corresponding thresholds were higher in Experiment II; compare Figures 3.4 and 3.11. This suggests that embedding the noise bursts within longer tone bursts made the task somewhat harder rather than providing an extra discrimination cue based on the detection of spectral splatter. Overall, it seems likely that performance of the two tasks was based on the temporal structure of the stimuli rather than on spectral cues (Taghipour et al., 2016b).

3.3.2 Effects of frequency and bandwidth

The results for both experiments showed that duration thresholds decreased with increasing bandwidth. This is consistent with what has been found for another measure of temporal resolution, namely the duration required to detect a gap in a band of noise (Eddins et al., 1992; Moore and Glasberg, 1988; Shailer and Moore, 1983). It seems likely that the tasks in Experiments I and II were performed by the detection of amplitude fluctuations in the noise...
stimuli. One plausible hypothesis is that the duration threshold corresponds to a fixed number of envelope fluctuations, for example a fixed number of envelope maxima. The number of envelope maxima per second increases with increasing bandwidth, and this could account for the dependence of the duration thresholds on bandwidth.

To test this hypothesis, the number of envelope fluctuations per second was determined empirically for the noise stimuli used in Experiment I. This was done by calculating the Hilbert envelope of a relatively long sample of the noise for each bandwidth and center frequency, locating the maxima, and calculating their number. Then, the number of envelope maxima per second was multiplied by the mean duration threshold for the same center frequency and bandwidth. If the hypothesis is correct, the resulting number should be approximately constant and independent of the center frequency and bandwidth (Taghipour et al., 2016b). The outcome is plotted in Figure 3.13. In this figure, the mean number of envelope maxima in each stimulus at the duration threshold is plotted as a function of bandwidth. Each center frequency is represented by a different symbol.

Figure 3.13: Estimated mean number of envelope maxima in the noise bursts over the duration corresponding to the measured threshold, plotted as a function of bandwidth in Hz. Each symbol represents a different center frequency, as indicated in the figure. Data of Experiment I are used for this analysis.
Although the results for the different center frequencies cluster around a single function, it is clear that the function is not independent of bandwidth. Also, it is clear that the number of envelope maxima in the stimuli at the duration threshold is less than 1, especially for small bandwidths. In other words, the envelope needs to go through less than one “cycle” of fluctuation for the fluctuation to be detectable (Taghipour et al., 2016b). The results suggest that duration thresholds are determined not only by how much the envelope fluctuates during the stimulus but also by the rapidity of the fluctuation; rapid fluctuations are harder to detect than slow fluctuations, consistent with temporal modulation transfer functions (Dau et al., 1997b; Viemeister, 1979; Viemeister and Plack, 1993).

A comparison of Figure 3.4 and Figure 3.11 shows that some of the thresholds were lower in Experiment I and they generally covered a smaller range. Isolated noise bursts seem to be easier to distinguish from tone bursts of the same duration than noise bursts embedded in a (longer) tone burst. This might be due to temporal uncertainty in the latter case or to forward and backward masking of the noise burst by the surrounding tone.

### 3.3.3 Conclusions

The results suggest that duration thresholds for discriminating a noise burst from a tone burst of the same duration and center frequency depend strongly on the envelope fluctuations in the noise stimulus. The thresholds also depend on the rapidity of the fluctuations. This information can be used in the design of more effective perceptual coders. The results also show that duration thresholds are slightly higher for noise bursts embedded within a longer tone burst than for noise bursts in isolation. This information can also be used in the design of more effective perceptual coders (Taghipour et al., 2016b).

### 3.4 Experiment III: masked threshold of noise in presence of narrower bands of noise

#### 3.4.1 Theory

Psychoacoustic models of perceptual audio encoders estimate masking effects (Section 2.3) of the audio signals. As discussed in Section 2.5, noise-like signals are more effective maskers than tone-like signals. Although asymmetry of masking between different kinds of stimuli has been studied in the literature, the effects of masker bandwidth and center frequency has not yet been investigated systematically (Taghipour et al., 2016c).

Since, in perceptual audio coding, the quantization noise is spread over the whole width of the subbands, but the bandwidth of the audio signal (masker) in the subbands could vary, a
systematic investigation of the role of masker bandwidth was necessary. Hall (1997) and Verhey (2002) investigated the role of the bandwidth in Hz. However, due to the usage of subbands in the audio coding, it was desirable to investigate the bandwidth not only expressed in Hz, but also expressed in fractions of subbands, the widths of which are related to the widths of the hypothetical auditory filters (or critical bands).

The studies of the asymmetry of masking investigated only the mid-range center frequencies of 1 and 2 kHz (Hellman, 1972; Hall, 1997; Verhey, 2002). A study of masking patterns at lower and higher frequencies was needed. Thus, in order to have more insight into the effect of bandwidth and center frequency on masking, in the course of this doctoral thesis, a masking experiment in various conditions was conducted. It should help us to get a more complete picture of the role of tonality/noisiness of the masker on masking in the context of perceptual audio coding. The effects of center frequency and bandwidth were investigated.

In a few studies, the temporal structure of the probe or masker has been directly manipulated (Gockel et al., 2002; Krumbholz et al., 2001; Treurniet and Boucher, 2001). In the present experiment, the probe was always a noise band, meant to be representative of the quantization noise in one subband of a perceptual coder. As in Hall (1997) and Verhey (2002), the temporal structure of the masker was manipulated indirectly, by varying its bandwidth (Taghipour et al., 2016c). With decreasing bandwidth, the amplitude and frequency fluctuations in a noise band become slower and the signal becomes more like a slowly varying sinusoid (Bos and de Boer, 1966; Dau et al., 1999).

3.4.2 Method

3.4.2.1 Stimuli

Both the masker and the probe were bandpass noise, the masker having a narrower bandwidth than the probe. The probe was meant to represent the quantization noise that might occur in a perceptual coder and the masker was meant to represent the audio signal. A trial consisted of two intervals, separated by 500-ms of silence. One interval contained the masker alone and the other contained the masker plus the probe, which were gated synchronously. The order was selected randomly for each trial. The duration of each stimulus was 400 ms, including 40-ms raised-cosine ramps. Since the bursts were long (enough), effects of the spectral shape of the window were negligible (Taghipour et al., 2016c).

For a signal that has been processed via a perceptual coder, the absolute level of the signal in a given frequency region is not an effective cue to the presence or absence of quantization noise (Taghipour et al., 2016c). In contrast, absolute level can be a cue in masking experiments, since the level of the masker plus signal is usually greater than the level of the masker alone (Bos and de Boer, 1966). Hall (1997) and Krumbholz et al. (2001) employed a roving level procedure.
to make overall level an ineffective cue. In the present study, the level of the masker plus probe was adjusted to be equal to the level of the masker alone (65 dB SPL at the eardrum) to make overall level a less salient cue \cite{Taghipour2016a}. The probe-to-masker ratio was adapted to determine the masked threshold.

In the course of a full up-down run, the independent variable was the probe-to-masker ratio, in dB. Whereas the masker level was constant, the probe level was varied using an amplitude factor given by

$$factor = \sqrt{\frac{E_{\text{next}}}{E_{\text{current}}} = 10^{\frac{dB_{\text{next}}-dB_{\text{current}}}{10}}}$$

where $E$ stands for the overall energy of a signal. The instantaneous amplitude of the probe was multiplied by this factor. The values $dB_{\text{current}}$ and $dB_{\text{next}}$ are SPLs of the probe at the current and next time instances, respectively.

The noise and probe bandwidths were specified on the ERB$_N$-number scale \cite{Glasberg1990,Moore2012}. The center frequencies were 257, 697, 1538, 3142, and 6930 Hz, corresponding to 7, 13, 19, 25, and 32 Cams on the ERB$_N$-number scale. The probe was always 1-Cam wide. The masker bandwidth was 0.1, 0.35, or 0.6 Cams \cite{Taghipour2016a}. These bandwidths were chosen to be representative of the situation in perceptual audio coders: the quantization noise, here represented by the probe, is spread over an auditory filter; the audio signal in a given subband, here the masker, is usually spread over a narrower portion of the subband. White Gaussian noise was digitally filtered with a 4th order Butterworth filter in the time domain to generate the bandpass noises. For each trial, a new random noise was generated for the masker. The same noise sample was used in both intervals of a forced-choice trial (see below). The signal was also generated randomly for each trial. As for subjective quality tests of monophonic coded audio signals \cite{ITU-RBS1534-1}, stimuli were presented diotically. As each condition was tested 3 times for each subject, a total of 45 runs was carried out for each subject: 5 frequencies $\times$ 3 bandwidths $\times$ 3 repetitions $= 45$.

### 3.4.2.2 Procedure

The experiment was carried out with a two-interval two-alternative forced choice method. Subjects were required to indicate which interval contained “an additional noisier signal.” This task was chosen since it resembles the types of tasks used to evaluate perceptual coders \cite{ITU-RBS1534-1}. Feedback in the form of a green light (for correct) or a red light (for wrong) was provided after each trial. The graphical user interface is depicted in Figure 3.14.

As in Experiment II, a hybrid staircase procedure controlled the independent variable, i.e. probe-to-masker ratio. All other parameters were kept constant in the course of a run. A run started with a probe-to-masker ratio at which the signal was clearly audible (e.g. 10 dB). Up to the 3rd reversal point a 1-down/1-up method \cite{Dixon1965,Levitt1971} was used to achieve a
rapid approach to the threshold level. After that, a 3-down/1-up method (Levitt, 1971; Wetherill and Levitt, 1965) was used to estimate the 79% point on the psychometric function. The step size in probe-to-masker ratio was 5 dB up to the second reversal point and 3 dB until 2 more reversals occurred. A fixed step size of 2 dB was used until 6 more reversals occurred, and the threshold was taken as the mean probe-to-masker ratio at the last 6 reversal points, i.e. 3 minima and 3 maxima. Three threshold estimates were obtained for each condition. If the standard deviation of these three estimates exceeded 4 dB, a fourth run was obtained (Taghipour et al., 2016c); more on this in Section 3.4.2.5.

### 3.4.2.3 Apparatus

The listening room had a background noise level of 21 dBA. By means of an artificial head (KE-MAR, GRAS, Holte, Denmark), the levels of the stimuli were calibrated at the eardrum. Stimuli were computed digitally with a resolution of 16 bits, played back at a sample frequency of 48 kHz, converted to analog form using a Lawo DALLIS 941/83 digital-to-analog converter (Rastatt, Germany), and presented diotically via Sennheiser HD 650 headphones (Wedemark, Germany).

### 3.4.2.4 Design

A pilot experiment with six normal-hearing subjects using 9 conditions (3 frequencies and 3 bandwidths) preceded the main experiment. Analysis of the results of the pilot study led to the parameters and design of the final experiment, as described earlier.

In the main experiment, each subject was tested in three sessions of about 45-50 minutes each, separated by at least one day. Prior to the main experiment, subjects read a page of instructions. A training session of four runs followed. This preparation/training phase took 20 to 25 minutes. During the main experiment, a break was given after every third run (i.e. after approximately 10-12 minutes). The conditions were presented in a random order (Taghipour et al., 2016c).
CHAPTER 3. PSYCHOACOUSTIC STUDIES

3.4.2.5 Subjects

Eighteen self-reported normal-hearing subjects participated in the experiment. Audiometric thresholds were measured for frequencies up to 14 kHz using a software-based audiometer (SelfscreeningAudioMeter V1.32, HörTech GmbH, Oldenburg, Germany) and Sennheiser HDA 200 headphones. For all tested frequencies, the subjects had absolute thresholds in the range -20 to +20 dB relative to the mean absolute thresholds for normal-hearing subjects (ISO 389-7, 2005).

Inspection of the results revealed some subjects whose results were highly variable or whose results appeared to be outliers. Two criteria were used to identify such subjects and to exclude their results from the statistical analysis:

1. As mentioned in Section 3.4.2.2, for each subject and condition, if the standard deviation of the threshold estimates across the three runs exceeded 4 dB, an additional estimate was obtained. Subsequently, the closest three estimates were chosen. If the standard deviation of these exceeded 4 dB, all results for that subject were discarded. The results for four subjects were excluded based on this criterion.

2. For each condition, the average threshold for an individual subject had to be within the interval $[\text{Mean} - 2 \times SD, \text{Mean} + 2 \times SD]$, with Mean representing the mean masked threshold across all subjects in that condition and SD representing the standard deviation across all subjects (after applying exclusion criterion 1). Based on this criterion, the data for one subject were excluded.

Consequently, the final statistical analysis is based on thresholds for 13 subjects: seven males and six females. They were aged between 20 and 28 years (mean 24 years, median 25 years).

3.4.3 Results and discussion

Figure 3.15 shows the mean masked thresholds (expressed as probe-to-masker ratio) across subjects and their 95% confidence intervals as a function of frequency with bandwidth as parameter.

A repeated-measures ANOVA was conducted on the masked thresholds with factors bandwidth and center frequency (Taghipour et al., 2016c). The effect of masker bandwidth was significant; $F(2, 24) = 534.71, p < 0.001$. The effect of center frequency was significant; $F(4, 48) = 16.84, p < 0.001$. There was a significant interaction between bandwidth and center frequency; $F(8, 96) = 3.47, p < 0.05$. Post hoc pairwise comparisons (based on Fisher’s protected LSD test) were conducted to investigate the effects of bandwidth and center frequency (Taghipour et al., 2016c). Thresholds increased with increasing bandwidth (all pairs $p < 0.001$). Thresholds for the lowest and highest center frequencies were significantly higher than those for the three
medium center frequencies (all pairs $p < 0.001$), but they did not differ significantly from each other ($p = 0.199$). The threshold for the center frequency of 697 Hz was significantly higher than for the center frequency of 1538 Hz ($p = 0.029$). There were no other significant effects of center frequency (all pairs $p > 0.05$).

The effect of bandwidth is consistent with previous research (Hall, 1997; Verhey, 2002). For example, for a center frequency of 2000 Hz and a probe bandwidth of 256 Hz (just over 1 Cam), Verhey reported that the probe-to-masker ratio at threshold increased from about -22 dB to -4 dB when the masker bandwidth was increased from 16 Hz to 256 Hz. As shown in Figure 3.16, for a center frequency of 1538 Hz, we found that the probe-to-masker ratio at threshold increased from about -23 dB to -6 dB when the masker bandwidth was increased from 19 Hz (0.1 Cam) to 114 Hz (0.6 Cam). The 0.1-Cam wide masker was not very effective presumably because the relatively slow amplitude fluctuations in the masker did not effectively mask the faster amplitude fluctuations in the probe (Bacon and Grantham, 1989; Houtgast, 1989). The probe-to-masker
Figure 3.16: Means of the masked thresholds across the 13 subjects plotted as a function of masker bandwidth in Hz with center frequency (or bandwidth of the probe) as parameter. Thresholds are expressed as probe-to-masker ratios in dB.

The effect of center frequency may reflect several factors. Firstly, absolute thresholds are lower for medium frequencies than for very low or very high frequencies (ISO 389-7, 2005). It is possible, therefore, that the results were influenced to some extent by the reduced sensation level (SL) of the probe at low and high center frequencies (Miller, 1947). However, at the lowest center frequency used, 257 Hz, the mean normal absolute threshold is about 12 dB SPL, and the mean measured probe threshold was about 44 dB SPL, so the probe had an SL of at least 30 dB. It seems unlikely that the SL of the stimuli had any substantial influence on the results (Taghipour et al., 2016c).

Another factor that might have influenced the effect of center frequency is the absolute bandwidth of the probe and masker in Hz (Verhey, 2002). As the center frequency increased, these bandwidths also increased. Hence, the rapidity of the envelope fluctuations in the probe and
masker also increased (Rice, 1954). For very low amplitudes of modulation rates, the hypothetical 
modulation filters in the auditory system appear to be very broadly tuned (in terms of their 
relative bandwidth), but the sharpness of tuning increases somewhat as the modulation frequency 
increases above about 10 Hz (Dau et al., 1997a; Sek and Moore, 2002). The improving selectivity 
in the modulation domain may account for the decrease in masked thresholds as the center 
frequency was increased from 257 to 697 Hz. However, the hypothetical modulation filters 
probably span a limited range of center frequencies, with the highest modulation filter being 
tuned to about 120 Hz (Ewert and Dau, 2000; Moore et al., 2009). For the highest center 
frequency, the probe bandwidth was 773 Hz and much of the modulation spectrum of the probe 
would have fallen above the center frequency of the highest modulation filter (Taghipour et al., 
2016c). This modulation would presumably have been detected using a form of “off-frequency 
listening” in the modulation domain, via the filter with the highest modulation center frequency. 
If the high-rate modulation evoked by the probe was detected via a modulation filter tuned to 
lower modulation frequencies, then that high-rate modulation would have been more effectively 
masked by the lower rate modulation evoked by the (narrower bandwidth) masker. This could 
account for the increase in masked thresholds for the highest center frequency used (Taghipour 
et al., 2016c).

3.4.4 Conclusions

For all five center frequencies, masked thresholds increased with increasing masker bandwidth, 
the increase ranging from about 15 to 23 dB. The probe-to-masker ratio at masked threshold 
was also dependent on center frequency of the probe, being lowest for medium center frequencies 
(Taghipour et al. 2016c).

The results can help in improving the psychoacoustic models used to predict masked thresholds 
in perceptual coders. Firstly, the results indicate that the number of bits allocated in a given 
subband can be decreased for very low and high center frequencies. Secondly, since the masked 
thresholds for the lowest masker bandwidth (0.1 Cam) are quite close to those produced by tone 
maskers, it appears that the effect of tonality “saturates” for bandwidths less than 0.1 Cam, and 
this can be exploited in the models. Finally, the additional detailed data on the maskers with 
intermediate bandwidths (0.35 Cam and 0.6 Cam) combined with data in the literature can be 
used to improve the models for cases where the masker has intermediate tonality (Taghipour 
et al. 2016c).
Chapter 4

The Psychoacoustic model

Sound waves can be captured by sensor devices such as a microphone, whose output is a voltage that varies continuously over time. This is referred to as an “analog” representation. Analog representation of sound is used in vinyl discs and audio cassettes. However, most modern systems work with a digital representation of sound. For this purpose, an analog-to-digital converter is used (Oppenheim and Schafer, 2010; Proakis and Manolakis, 2007). The magnitude of the continuous analog waveform is measured, or “sampled,” at regular time intervals and the samples are stored as a sequence of numbers representing the magnitudes (Oppenheim and Schafer, 2010; Proakis and Manolakis, 2007). The sampling frequency has to be high enough such that the waveform is represented accurately and uniquely. Otherwise, when the signal is converted back to the analog domain, e.g. for loudspeaker playback, audible modifications will occur due to “aliasing”1 (Oppenheim and Schafer, 2010).

The values representing the waveform magnitudes for the samples are saved with limited precision. This process is called quantization (Oppenheim and Schafer, 2010). A quantizer represents the continuously scaled values of the samples by a limited number of magnitudes. This can be done based on the limitations of the auditory system: the quantization process exploits the fact that very small changes in the magnitude, or small errors in representing the magnitude of a sample, are not audible to the human auditory system (Bosi, 2003). A signal that is discrete in time and magnitude is called a digital signal (Bosi, 2003; Oliver et al., 1948) and can be represented as a sequence of binary digits, or bits. A digital signal has a bit rate, also known as a data rate, which is the product of the number of bits used to represent each sample (e.g., 16 bits) and the number of samples per second (e.g., 48000) (Bosi, 2003). A “high-fidelity” digital signal might have a bit rate of 768000 bit/s, meaning that transmission or recording of the signal would require 768000 bits to be transmitted or stored each second.

1For two or more analog signals, a low sampling rate can result in their indistinguishable discrete-time representations. This effect is called aliasing. Thus, the so called “aliasing distortion” might occur when a signal is reconstructed from its discrete-time representation into the analog domain. Aliasing is particularly caused when the sampling frequency is not high enough to cover all the spectral properties of an analog signal.
The input audio signal to a perceptual coder is digital, i.e. already discrete-time and quantized. Typically, a pulse code modulation (PCM) is used for this (Bosi, 2003; Oliver et al., 1948; Oppenheim and Schafer, 2010). Through a careful choice of a high sampling rate and fine quantization steps, this original digital signal should have the same subjective quality as the original analog signal. Further, for quality assessment of perceptual audio codecs, PCM items serve as references.

4.1 Perceptual audio coding

One of the main goals of perceptual signal processing in audio codecs (Bosi et al., 1997; Brandenburg and Bosi, 1997; Brandenburg and Stoll, 1994) is to reduce the data rate of the digital representation of the audio or speech input by exploiting the properties of the human auditory system. This is called "data compression." The goal is to reduce the data rate as much as possible without degrading the audio quality, or degrading it only to a very small extent. Since a deep understanding of both the characteristics of the sound source and the properties of the human auditory system is crucial to the effective design of such systems, expertise in engineering and signal processing, as well as in psychoacoustics, is required.

Techniques of perceptual audio coding have been continually developed in the last decades and their area of application has expanded. In every personal computer system, portable music player, mobile phone device, streaming media or DVD and Blu-ray player, technologies such as mp3 (Brandenburg and Stoll, 1994; ISO/IEC 11172-3, 1993), and MPEG-2/4 advanced audio coding (AAC) (Bosi et al., 1997; ISO/IEC 13818-7, 1997) have found widespread adoption in the marketplace.

In order to keep track of temporal alterations of an audio signal, perceptual codecs commonly split the signal into brief time segments, called “frames,” and process them individually. With the help of a frequency analysis using a time-frequency transform or a filter bank, the content within each frame is divided into a number of frequency bands. As a result, the audio input is split into time-frequency segments.

In the encoding process, the average data rate is reduced by means of quantization of the signal in each frame and band. This introduces a so called “quantization error” to the audio signal. A relatively fine quantization leads to an error which can be interpreted as an additive noise, called quantization noise. The quantization noise in a subband is inaudible at the output of a decoder if it is masked by the content of the audio signal or is below the absolute threshold of hearing. Under this condition, the most efficient coding, i.e. the minimum number of bits required to represent the audio input in the subband, is achieved if the quantization noise lies just below its masked threshold. Therefore, a psychoacoustic model is used to estimate the masked threshold of the quantization noise, allowing a proper bit allocation for that subband.
CHAPTER 4. THE PSYCHOACOUSTIC MODEL

Figure 4.1: Basic structure of a transform-based audio encoder. Frames of the input audio signal are decomposed into their spectral components by means of a transform. The perceptual model (psychoacoustic model) estimates masking thresholds for the frames with which it controls quantization steps for the individual subbands. Typically, entropy coding finalizes the encoding process.

As mentioned above, the main goal of audio coding is to reduce the requirements for storage and transmission of the audio signal data via compression. This is partly done by employing transform, predictive and/or entropy coding. These techniques reduce redundancies in the signal based on its properties.

However, redundancy reduction alone does not lead to a sufficient reduction in the average bit rate of the audio signal. Hence, in perceptual audio codecs, psychoacoustic models are used to control the quantizers of the spectral components of frames. This, consequently, reduces irrelevancies in the audio signal. A general block diagram of such a perceptual audio encoder is shown in Figure 4.1. By means of, e.g., the modified discrete cosine transform (MDCT) or other techniques, a frame of the input signal is decomposed into its spectral components. A psychoacoustic model estimates the masking threshold for the frame. The level of the masking threshold for each spectral region (or subband) determines the step size of the corresponding quantizer. Typically a method of entropy coding follows. Optimally, after this compression procedure, the quantization noise should be imperceptible to normal human listeners at the output of the decoder; i.e., it should lie just below the masking threshold. The coding literature refers to such curves as the “masking curves of the input signal” or as the “masked threshold of the quantization noise.”

Regardless of whether a transform or a filter bank is used for the spectral decomposition

---

2Entropy coding is a lossless compression scheme whereby fixed-length inputs are compressed in the form of variable-length outputs, normally based on the stochastic probabilities of the input [Bosi, 2003].

Dissertation, Armin Taghipour
(Brandenburg et al., 1992), the resulting time-frequency segments give comprehensive information about the distribution of the signal energies. This not only helps reducing redundancies, it also helps to identify the irrelevancies for the psychoacoustic analysis by providing a local time-frequency distribution of the masking effects (Painter and Spanias, 2000). Therefore, the efficiency of the perceptual coder relies to a great degree on the design of the filter bank or the transform. On one hand, “critical sampling” (Oppenheim and Schafer, 2010) is required to restrict the data rate to the lowest rate possible. On the other hand, a desire for a “perfect reconstruction” (or at least an almost perfect reconstruction) restricts the design of the analysis and synthesis filters.

4.1.1 The modified discrete cosine transform (MDCT)

The modified discrete cosine transform (MDCT) is a time-frequency decomposition tool which fulfills the above mentioned criteria. It is a uniformly spaced linear orthogonal lapped block transform (Edler, 1989; Malvar, 1990), based on the theory of time-domain aliasing cancellation (TDAC) (Princen and Bradley, 1986; Princen et al., 1987). Perfect reconstruction from a critically sampled representation of overlapping blocks is achieved during overlap-and-add in the synthesis process (Bosi, 2003; Painter and Spanias, 2000). The MDCT not only fulfills the above mentioned design constraints of an optimal filter bank, but it also achieves maximum redundancy removal by optimally choosing time-varying resolutions for both time- and frequency-domain mappings by employing window switching (Edler, 1989).

A detailed discussion on TDAC and window switching is beyond the scope of this thesis. For further details, see (Bosi, 2003; Edler, 1989, 1995; Malvar, 1990; Painter and Spanias, 2000; Princen and Bradley, 1986; Princen et al., 1987). However, since the MDCT is used for most of the coding in this thesis, it will be discussed in the following.

The analysis MDCT requires $2M$ input samples to generate $M$ spectral coefficients. Given an input $x(n)$, the forward MDCT generates $M$ transform coefficients $X(k)$ as follows (Painter and Spanias, 2000):

$$X(k) = \sum_{n=0}^{2M-1} x(n)h_k(n), \quad 0 \leq k \leq M - 1 \quad (4.1)$$

where $n$ and $k$ indicate the indices of time samples and spectral coefficients, respectively. The analysis transform performs a series of inner products between the $M$ analysis basis functions $h_k(n)$ and the input. The cosine modulated function has the form:

$$h_k(n) = w(n)\sqrt{\frac{2}{M}} \cos \left( \frac{(2n + M + 1)(2k + 1)\pi}{4M} \right) \quad (4.2)$$

where $w(n)$ is a window function used for the TDAC. Whereas the window length is $2M$, the frequency resolution is given by $M$ spectral bins. A frequently used window function in an
oddy-stacked MDCT setting is a sine window (Edler, 1989):
\[
w(n) = \sin\left(\frac{(n + 0.5)\pi}{M}\right)
\]
with \(0 \leq n \leq 2M - 1\). Alternatively, the so called Kaiser-Bessel derived window (Bosi, 2003; Kaiser and Schafer, 1980) can be used, which has higher stopband attenuation.

To compute the MDCT, overlapping windows are used. The overlap size is equal to half of the window size (50% overlapping). Figure 4.2 shows an MDCT analysis block diagram. It can be observed that the MDCT uses a 50% overlap between two adjacent frames, and generates \(M\) spectral coefficients from \(2M\) time-domain samples.

The Inverse MDCT has the form (Painter and Spanias, 2000):
\[
x(n) = \sum_{k=0}^{M-1} [X(k)h_k(n) + X^P(k)h_k(n + M)], \quad 0 \leq n \leq M - 1
\]
where \(X^P(k)\) denotes the previous block of transform coefficients. The inverse MDCT (IMDCT) is depicted in Figure 4.3. First, the IMDCT delivers \(2M\) samples at its output; Figure 4.3 also shows the overlap-and-add synthesis, which leads to frames of \(M\) output samples. The synthesis windows are identical to the analysis windows.
4.1.2 The experimental coding setup

For the studies in Chapter 5, an MDCT-based transform audio coding scheme was chosen. For the experimental coding setup, a fixed input frame length of 1024 samples was used, which is equivalent to an MDCT window length of 2048 samples. Different variants of a PM, which will be described in Section 4.3, were applied for quantizer control. Although entropy coding was not applied, entropy rates were estimated. The different versions were controlled to have a desired average data rate, estimated for a large set of standard test audio signals with varying characteristics. This was possible by applying a scaling factor to the calculated masking thresholds. For the quality assessment tests, scaling factors were determined which resulted in equal average entropy rates of 48 kbit/s (medium quality) and/or 32 kbit/s (low quality).

4.2 Psychoacoustic models for perceptual audio coding

It was mentioned that a psychoacoustic model for audio coding analyzes the input signal and estimates masking thresholds for its frames and bands. The level of a masking threshold in a spectral region gives the required accuracy in the quantization process of the corresponding subband. High masking levels allow large quantization step sizes and consequently higher quantization error. For a subband with low masking threshold, not much quantization noise is allowed. This leads to a fine quantization of some time-frequency components and coarse quantization of others. This is the basic mechanism of how a psychoacoustic model controls the quantization error, and

\footnote{Shannon entropy is the amount of information in a signal; i.e. the amount of unexpected data/information that a message contains \cite{Shannon1948}.}
hence, the quantization noise in the decoder. Accordingly, psychoacoustical irrelevancies in the signal can be removed.

As discussed in Chapter 2, psychoacoustic studies show differences between the masking strengths of tonal and noise-like maskers (Hellman, 1972; Gockel et al., 2002; Hall, 1997; Verhey, 2002). The masking effect of narrowband noise is stronger than that of a tone placed at the center frequency of the noise with the same total energy. Furthermore, the discussion in Section 2.2 and the studies presented in Sections 3.1 and 3.2 showed that the human auditory system needs some time to distinguish between narrowband noise and tone (Taghipour et al., 2012, 2013a, 2014a). This duration threshold lies somewhere between 2 and 30 ms (for frequencies up to 2-3 kHz). The duration threshold depends on the center frequency and bandwidth of the narrowband noise (Taghipour et al., 2013a, 2014a).

Johnston (1988a) introduced “perceptual entropy,” expressed in bits per sample, as a theoretical limit on the compressibility of an audio signal (Painter and Spanias, 2000); i.e. the minimum entropy rate at which, in theory, no perceptual difference should be found between the coded signal and the original signal. In order to estimate masking thresholds for the time-frequency segments of the audio signal, psychoacoustic models typically use a measure for estimation of tonality/noisiness, such as the “spectral flatness measure” (Gray and Markel, 1974; Johnston, 1988a,b) or the “chaos measure” (ISO/IEC 11172-3, 1993; Brandenburg, 2002).

Perceptual audio coders such as mp3 (MPEG-1/2 Audio Layer 3 (Brandenburg and Stoll, 1994; ISO/IEC 11172-3, 1993)) or MPEG-AAC (MPEG-2/4 advanced audio coding (Bosi et al., 1997; ISO/IEC 13818-7, 1997)) use such psychoacoustic models which approximate masking effects based on perceptual entropy and tonality estimation methods. The details of the psychoacoustic models can be found in (Bosi, 2003; Brandenburg, 2002; ISO/IEC 11172-3, 1993; Johnston, 1988b; Painter and Spanias, 2000). Additionally, other PMs were suggested for audio coding algorithms, such as by Baumgarte et al. (1995) and van de Par et al. (2005).

As long as a high bit rate is available, the quantizer control works well, i.e., almost no perceptual differences can be detected between the coded and the original signals in subjective quality tests. However, an important fact is that the masking threshold estimated by the PM is violated when coding at medium and low bit rates. There is still room for improvement in estimating psychoacoustical quantizer control data which brings the best possible audio quality when slight violations of the masking thresholds occur.

Hall (1997) and Painter and Spanias (2000) suggested that there was a need for more accurate masking models. A transform-based power spectrum masking model is not always able to accurately detect rapid modulations in the envelope of the audio signals. In the following, a psychoacoustic model is presented which uses a filter bank. Further, Chapter 5 introduces a number of tonality estimation methods which, combined with the filter bank-based model, can improve the estimation of masking thresholds in perceptual audio coding algorithms.
4.3 The model

Taghipour et al. (2010, 2014b) introduced a filter bank-based psychoacoustic model. Figure 4.4 shows a block diagram of the PM. A filter bank decomposes the audio signal into its spectral components (Taghipour et al., 2010, 2013b, 2014b; Chen et al., 2014). In doing so, a uniform frequency analysis is avoided, which would be the case if a transform was used. Further, forward masking is modeled, as described below.

Based on the degree of tonality of the time-frequency segments, their estimated masking thresholds are scaled. A final masking threshold is built from a weighted combination of the outputs of the individual channels and is mapped onto the coding scheme. Further, the level of the absolute threshold is taken into account. In the following, the blocks of the PM will be described in more detail.

4.3.1 The filter bank

The filters were designed based on the theory of the auditory filters in the inner ear (Fletcher, 1940). The design of the initial filter bank was based on the Bark scale (Fastl and Zwicker, 2007), similar to the subbands in most conventional codecs. For an alternative version of the filter bank based on the ERB\(_N\) (Moore, 2012), see Section 4.4. The center frequencies of neighboring filters are spaced \( \frac{1}{4} \) Bark apart from each other. Figure 4.5 illustrates the construction of the desired frequency responses (Taghipour et al., 2010). The bandpass filters, which take into account the spreading in simultaneous masking, were designed from reversed masking curves (Bosi, 2003; Fastl and Zwicker, 2007). The slope of the magnitude frequency response of a filter was chosen to be 20 dB/Bark towards the low frequencies and -27 dB/Bark towards the high frequencies. For a sampling rate of 48 kHz and with a Bark-based design, the filter bank consisted of a total of 104 filters.

Initially, real-valued filters were computed (Taghipour et al., 2010). However, most of the work in this thesis was done with a complex-valued filter bank (Chen et al., 2014).
CHAPTER 4. THE PSYCHOACOUSTIC MODEL

Figure 4.5: Construction of the desired frequency response for a filter with center frequency \( f_{\text{thr}} \) (Taghipour et al., 2010). Left: influence of maskers at different frequencies on the threshold at \( f_{\text{thr}} \). Right: magnitude response of the filter. The abscissa shows the frequencies on a Bark scale.

et al., 2013b, 2014b). This proved advantageous when the filter outputs were used for further processing or analysis. IIR\(^{4}\) filters were used because, generally, they can be implemented with significantly lower computational complexity than FIR\(^{5}\) filters with similar magnitude frequency response.

In general, an IIR filter can be described by its transfer function given by (Oppenheim and Schafer, 2010):

\[
H(z) = \frac{\sum_{k=0}^{K} b_k z^{-k}}{1 + \sum_{m=1}^{M} a_m z^{-m}}
\]

(4.5)

where \( K \) and \( M \) give the total number of zeros and poles, respectively. The implementation of the filters in the PM was based on this representation.

Other than the parametric description for IIR filter banks such as the gammatone filter bank (Patterson et al., 1995a; Slaney, 1993; van de Par et al., 2005), the filter design consisted of a numerical optimization to approximate the “desired” magnitude frequency responses. For this optimization, an alternative representation of the magnitude transfer function was used, which, in the case of a real-valued filter, can be computed with complex-conjugate zeros and poles:

\[
|H(z)| = G \left| \prod_{k=0}^{K} \left( z - |z_k|e^{j\angle z_k} \right) \right| \left| \prod_{m=0}^{M} \left( z - |p_m|e^{j\angle p_m} \right) \right|
\]

(4.6)

\(^{4}\)Infinite impulse response

\(^{5}\)Finite impulse response

Dissertation, Armin Taghipour
where $z_k$ and $p_m$ are the zeros and poles of the transfer function, respectively. The magnitude frequency responses of the real-valued IIR filters (Taghipour et al., 2010) of the PM were computed as above in the design optimization, whereby $K$ and $P$ both were set to 1, resulting in 4 zeroes and 4 poles.

In the case of the complex-valued version of the filter bank (Chen et al., 2014; Taghipour et al., 2013b, 2014b), the magnitude transfer functions of the filters were computed with the following formula.

$$|H(z)| = \frac{G}{\prod_{k=0}^{K} (z - |z_k|e^{j\angle z_k}) \prod_{m=0}^{M} (z - |p_m|e^{j\angle p_m})}$$

(4.7)

Minimum-phase filters were generated for both the real-valued and the complex-valued versions of the filter bank. The real-valued filter design was done with pairs of conjugate poles and zeros. The complex-valued filters were represented by individual poles and zeros.

### 4.3.2 Modeling of forward masking

Due to the forward masking, the masking thresholds of a frame are relevant for the thresholds of the following frames. In the presented PM, this is modeled with a decaying exponential function. The computation of the exponential decay is carried out in the form of a “resistor–capacitor circuit,” which is shown in Figure 4.6.

#### Figure 4.6: Resistor–capacitor circuit used for the design of the exponential decay of forward masking.

4.3.3 Level scaling based on a tonality measure

A first approximation of the masking threshold is obtained from the outputs of the filters (Taghipour et al., 2010). However, this approximation assumes narrowband noise maskers.

---

Dissertation, Armin Taghipour

*A system is minimum-phase if, both the filter and its inverse are causal and stable. All zeros and poles of a minimum-phase filter are located inside the unit circle (Proakis and Manolakis, 2007).*
The level of the masking curves should be adjusted (reduced) based on tonality/noisiness of the time-frequency segments. Chapter 5 presents several tonality estimation measures for this purpose. Either the input signal or the outputs of the filters were used as the input to the tonality estimation block; more on this is presented in Chapter 5.

### 4.3.4 Weighting/mapping

The model can work with short or long blocks. First, masking thresholds of short blocks of 128 samples are calculated; i.e. an MDCT window length of 256. When long blocks are used, e.g., 1024 samples, the estimated masking thresholds of 16 neighboring short blocks are combined to cover the full window length of 2048 samples. Further, the resulting long block is weighted with a long MDCT window.

Since the spreading in simultaneous masking has already been taken into account in the filter design, no superposition (Baumgarte et al., 1995; van de Par et al., 2005) of the filter outputs is necessary. The masking thresholds estimated by the proposed PM are mapped onto the spectral resolution of the MDCT via linear interpolation.

### 4.3.5 Comparison to the absolute threshold

The effects of transmission of sound through the outer and middle ear are commonly modeled with an outer-middle ear filter (Dau et al., 1997a; Bosi, 2003; Moore, 2014; Painter and Spanias, 2000; Patterson et al., 1995a). In the presented model, however, the final masking threshold is compared to the absolute threshold of hearing. Whenever the levels of the spectral components of the masking threshold are below the absolute threshold, they are replaced by the level of the absolute threshold at those spectral points. It is assumed that a full-power signal is played back at a level just below the threshold of pain; for “threshold of pain,” see Fastl and Zwicker (2007).

### 4.4 An ERB\(_N\) design of the PM

Since, especially at low frequencies, the Bark and the ERB\(_N\) systems differ from each other, a second version of the PM was designed and computed based on ERB\(_N\); see Sections 2.4.1 and 2.4.2 for Bark and ERB\(_N\), respectively. The hypothesis was that this might have an effect on the overall quality of the PM.

This second version of the PM has the same block diagram as shown in Figure 4.4. However, the ERB\(_N\)-based complex filter bank consists of 123 fourth order, minimum-phase, IIR filters. As in Section 4.3.1 the filter bank was designed using a numerical optimization of the desired frequency responses, which were specified using rounded-exponential (roex) functions; for roex
filters, their frequency responses, and examples of auditory filter shapes see (Glasberg and Moore, 1990, 2000; Patterson, 1974, 1976; Patterson et al., 1982; Moore and Glasberg, 1983, 1987).

A part of the filter bank is illustrated in Figure 4.7. The roex filters’ frequency responses are depicted as a function of (a) frequency or (b) ERB$_N$-number. The placement of the center frequencies of the filter is a design parameter. It can be observed in Figure 4.7 (b) that, on an ERB$_N$-number scale, the 123 bandpass filters are linearly spaced such that their center frequencies are $\frac{1}{3}$ Cam apart.[7]

Depicted on a linear frequency scale, as in Figure 4.7 (a), the filters are slightly asymmetric. The slopes are shallower on the low-frequency side. The design of the asymmetry was based on a playback level of approximately 75-85 dB SPL, at which the auditory filters do not show sharp tuning (Glasberg and Moore, 2000; Moore, 2012). The rounded-exponential functions of the frequency responses have the form (Glasberg and Moore, 2000):

$$w(g) = \begin{cases} G (1 + p_l g) e^{(-p_l g)} + (1 + t g) e^{(-t g)} & \text{if } f < f_c \\ (1 + p_u g) e^{(-p_u g)} & \text{if } f > f_c \end{cases}$$

where $f$ is frequency, $f_c$ is the center frequency of the filter, and $g$ is $|f - f_c|/f_c$. In general, $G$ is a gain parameter in the design of roex filters. However, in Equation 4.8, $G$ is acting more as a weighting parameter in the numerator, since the term “$G + 1$” in the denominator makes the filter gain unity at its tip. This is convenient for modeling purposes (Glasberg and Moore, 2000).

Whereas $p_l$ and $p_u$ are the main steepness parameters for the lower and upper slopes, $t$ controls the “tail” slope on the low-frequency side (Patterson et al., 1982). For the initial design of the ERB$_N$-based filter bank, $G$ and $t$ were chosen to have the values 50 and 8, respectively. The values of $p_l$ and $p_u$ varied with the center frequencies of the filters.

For a variety of reasons, only one version of the PM was used for the comparison of the tonality estimation methods presented in Chapter 5, namely the Bark-based model. First, although the choice of the filter bank might cause differences in overall performance of the model, it was assumed to be of little importance for the comparison of the tonality estimation methods. Second, keeping the parameter “filter bank” constant made a systematic investigation of the parameter “tonality estimation method” easier. Third, the ERB$_N$-based filter bank was developed recently and was not available at the time the tonality estimation methods were compared to each other.[8]

Nevertheless, for a comparison of the PM to a conventional PM (see Chapter 6), both versions of the psychoacoustic model were applied to a standard audio codec.

---

[7] The Bark-based model had a spacing of $\frac{1}{4}$ Bark, which led to 104 filters. Thus, a spacing of $\frac{1}{3}$ Cam was sufficient here because it led to 123 filters, which is close to the resolution of the Bark-based model.

[8] It should be mentioned that, prior to the design of the presented ERB$_N$-based filter bank, a gammatone filter bank was implemented in the PM. First objective tests, simulations and informal listening tests showed a poor performance of the model; however, no systematic comparison has been carried out yet.
Figure 4.7: Frequency responses of 3 bandpass filters as a function of (a) frequency or (b) ERB<sub>N</sub>-number. The solid line shows a filter with a center frequency of 984 or, equivalently, an ERB<sub>N</sub>-number of 15.50. The filters with the dotted lines have center frequencies of 941 and 1028 and (center) ERB<sub>N</sub>-numbers of 15.16 and 15.83, respectively.
4.5 The MUSHRA test

As a first indicator, objective measures can be used for assessment of the quality of a coding scheme. However, measures such as signal-to-noise ratio (SNR) are neither very convenient nor reliable, because they do not take into account psychoacoustic principles (Brandenburg and Sporer [1992]). Hence, subjective quality tests have been widely used to compare coding schemes to each other. The most broadly used subjective quality test method in the field of perceptual audio coding is the MUSHRA\textsuperscript{9} test (ITU-R BS.1534-1 [2003] ITU-R BS.1534-3 [2015]). It is a reliable listening test for intermediate audio quality. In MUSHRA tests, the output signals to be rated are intermixed with a reference (original) signal and with a signal that is deliberately degraded by lowpass filtering.

4.5.1 Rating scale

In the MUSHRA test, all the conditions are compared to a known (presented), unprocessed, full-bandwidth reference signal. As shown in Figure 4.8, the subjective quality of each condition is rated on an interval scale (continuous quality scale) spanning the range from 0 to 100, i.e., a quality range from “bad” to “excellent.” For example, a range from 0 to 20 should be chosen for “bad” coding quality.

4.5.2 Conditions

A hidden reference signal is needed which is supposed to be rated at 100. Additionally, ITU-R BS.1534-1 [2003] and ITU-R BS.1534-3 [2015] recommend one or two anchors that are lowpass filtered versions of the reference with cutoff frequencies of 3.5 and 7 kHz. The ratings for the hidden reference and the lower anchors serve as an indication that the subject has understood the test and is able to detect the original signal and lower anchors. If a subject fails in this, his/her results are commonly discarded. In addition to the hidden reference and lower anchors, a MUSHRA test includes two or more codec versions which are supposed to be compared to each other.

The total test duration should be below 40 minutes, otherwise, fatigue could bias the test. Therefore, the number of items and conditions must be chosen carefully.

4.5.3 MUSHRA GUI

For most of the MUSHRA tests presented in this thesis, software, including a graphical user interface (GUI), was used which was developed by Fraunhofer-IIS\textsuperscript{10}. Figure 4.8 shows this GUI.

\textsuperscript{9}Multiple stimuli with hidden reference and anchor
\textsuperscript{10}The Fraunhofer Institute for Integrated Circuits IIS

Dissertation, Armin Taghipour
Figure 4.8: The GUI for the MUSHRA test [Chen et al., 2014]. The example shows the item Pitch-pipe. Here, six different conditions were compared to the reference (and to each other): hidden reference, low anchor (3.5 kHz) and four other coded versions. The conditions were randomly ordered from 1 to 6.

Here, the item Pitch-pipe was used. For each item, the order of conditions was randomized. In this example, six conditions were used: hidden reference, lower anchor, and four coded versions.

### 4.5.4 Items

There is a long list of standard test items with different stochastic, acoustic, and psychoacoustic characteristics ([ITU-R BS.1534-1, 2003]). In the following, a number of these items (mono, sampled at 48 kHz) are listed which were used in objective comparison measures or in the MUSHRA tests presented in Chapters 5 and 6: Castanets (si02), German male speech (es02), English male speech (es04), German female speech (es05), Female vocal (es01) [Suzanne Vega, solo], Triangle, Big band, Glockenspiel (sm02), Plucked strings (sm03), Harpsichord (si01/ts_01), Pitch-pipe, Bagpipe (sm01), Symphony orchestra (sc02), News intro (sc03) [pop music], Noise (Noise48), Pure tone [not included in the list], and Diverging tones [two frequency sweeps in opposite spectral directions; not included in the list]. The latter three signals were only used in objective quality tests and were not used in the MUSHRA tests.
The next chapter deals with a number of tonality estimation methods which were implemented in the PM. The quality of the methods was compared by means of MUSHRA tests. The coded audio items did not have a transparent quality; i.e., for most of the items, a difference to the original was audible for expert listeners on average. There were two reasons for this decision. First, the MUSHRA test is designed for intermediate quality [ITU-R BS.1534-1 (2003)]. Second, we wanted to test the codecs for a data rate at which some quality differences were audible. Third, a test which predominantly contains transparent coded items is tiresome for the subjects. Fourth, a statistical analysis of data for transparently coded items, which are most often rated above the value of 90 on the quality scale, is redundant, since the analysis typically leads to no statistically significant differences. Thus, average entropy rates of 48 or 32 kbit/s were used for the MUSHRA tests in Chapters 5, see Section 4.1.2, For the MUSHRA test in Chapter 6, a bit rate of 40 kbit/s was used.

\[11\] The terms “entropy rate” and “bit rate” are carefully chosen for the cases without (see Section 4.1.2) or with entropy coding, respectively.
Chapter 5

Tonality estimation methods

In the course of this doctoral thesis, several methods have been developed for estimation of tonality/noisiness of audio signals. In the coding literature, such methods are commonly referred to as “tonality estimation methods” (Painter and Spanias 2000; Pulkki and Karjalainen 2015). This name is also adapted in this thesis, although one should be cautious about what the terms tonality and noisiness represent; tone-like segments of audio signals are the segments with no or very regular, slow envelope fluctuations. In contrast, noise-like segments have strong, irregular and more rapid envelope fluctuations. In the following, various methods are presented and compared to each other by means of MUSHRA tests (Section 4.5) and/or objective measures.

5.1 A linear prediction method (Predictor)

Prediction is usually applied in source coding for exploiting correlations. The resulting prediction gain is the ratio of input variance and prediction error variance (Kleijn 2008; Taghipour et al., 2010). For noise-like signals with no or only low correlations (corresponding to flat power spectral densities), there is no prediction gain, so the prediction error variance is equal or almost equal to the input variance. On the other hand, for tonal signal components with distinct spectral structures, the prediction gain is high and the prediction error variance is much smaller than the input variance (Taghipour 2009; Taghipour et al., 2010). Therefore, prediction seems to be an appropriate tool to reproduce tonality-dependent masked thresholds.

Prediction cannot be applied directly to the filter output signals, since, due to filtering, they have distinct spectral shapes even for a white noise input, so that almost equally high coding gains are achieved for tonal and noise-like signals (Taghipour et al., 2010).

In order to overcome this problem, downsampling was applied with factors depending on the bandwidths of the individual filters. However, for some bands this was still not sufficient, since the replicated spectra for a noise input did not add up to a nearly flat spectrum if the center
frequencies were at unfavorable positions with respect to the lowered sampling rate (Taghipour, 2009). Further improvement was achieved by introducing (1) a modulation in each band, which shifted its center frequency to zero, and (2) a low pass filter which attenuated the replica that occurred due to the real-valued modulation process. The resulting system is shown in Figure 5.1.

Since the predictor needs to adapt to varying input characteristics, a lattice structure was selected which adapts the predictor coefficients for each input sample (Markel and Gray, 1976; Taghipour et al., 2010). The reciprocal of the prediction gain is calculated and used for the adjustment of the level of the masked threshold:

\[ q_p = \prod_{i=1}^{P} (1 - k_{i,p}^2) \]  

(5.1)

where \( P \) is the prediction order and \( k_{i,p} \) is the \( i \)-th lattice coefficient of the predictor in band \( p \) (Markel and Gray, 1976). The masked threshold \( thr_p \) of band \( p \) is then derived by multiplying the band variance \( s_p \) with a corresponding correction factor

\[ thr_p = \begin{cases} 
    s_p \cdot q_p^G & \text{if } q_p^G > Q_{min} \\
    s_p \cdot Q_{min} & \text{else}
\end{cases} \]  

(5.2)

where an exponent \( G = 1.2 \) and a limit \( Q_{min} = 0.01 \) is used. Since the downsampling factors can be quite high for the narrow bands at low frequencies, temporal resolution may be lost. By applying polyphase processing (Proakis and Manolakis, 2007) as shown in Figure 5.2 instead of pure downsampling, the original temporal resolution is maintained so that the thresholds adapt quickly in the presence of rapid signal changes. The final thresholds that can be passed to the quantization stage were calculated as averages for every 128 samples.

5.1.1 Evaluation

Figure 5.3 shows the tonality/noisiness estimate by the Predictor for (a) Diverging tones, (b) Pitch-pipe, (c) Harpsichord, and (d) white noise. Dark colors, or numbers closer to 0, show high tonality estimates; orange and yellow (i.e., numbers close to 1) indicate estimations of noisiness.

Figure 5.1: One branch of the PM extended by modulation and prediction for approximation of tonality-dependent masked thresholds (Taghipour et al., 2010).
CHAPTER 5. TONALITY ESTIMATION METHODS

Figure 5.2: Prediction based on polyphase components (Taghipour et al., 2010).

Figure 5.3: Predictor’s noisiness estimate for (a) Diverging tones, (b) Pitch-pipe, (c) Harpsichord and (d) white noise.
A first evaluation of the quality of the psychoacoustic model with the “Predictor” was carried out in a MUSHRA test by which the model was compared to a DFT-based psychoacoustic model (Taghipour et al., 2010). For most of the items, the quality/fidelity ratings of the outputs were very close to each other so that only few significant differences were found. For test items with strong tonal components, such as Triangle and Pitch-pipe, the IIR model demanded more bits and achieved higher scores than the DFT model. For the German male speech signal, the DFT model scored higher. The details of the DFT model and the coding schemes are beyond the scope of this thesis; further details can be found in (Taghipour, 2009); see Taghipour et al. (2010) for a detailed discussion of the psychoacoustic models, coding schemes used in the study, and the results of the MUSHRA test.

5.2 Partial spectral flatness measure (PSFM)

5.2.1 The initial method

Gray and Markel (1974) introduced a continuous-frequency spectral flatness measure (SFM) as a measure of “whiteness” of a spectrum. In its discrete form, the SFM has been widely used as a measure of tonality/noisiness in digital audio and speech coding (Bosi, 2003; Brandenburg and Bosi, 1997; Painter and Spanias, 2000). The SFM is defined as the ratio of the geometric and the arithmetic means of the power spectrum, i.e. power spectral density, of a signal. The power spectrum is evaluated at \( N \) discrete frequencies \( f_k \) by means of a spectral decomposition. The discrete SFM is calculated as the geometric mean over the arithmetic mean of the power spectrum \( S_{xx}(k) \) of corresponding spectral coefficients:

\[
\text{SFM} = \sqrt{\frac{N}{N-1} \prod_{k=0}^{N-1} S_{xx}(k)} \over \frac{1}{N} \sum_{k=0}^{N-1} S_{xx}(k) \]

(5.3)

where \( 0 \leq \text{SFM} \leq 1 \), \( N \) is the number of the spectral components and \( S_{xx}(k) \) is the power spectrum value for the \( k \)-th spectral component. A value of “1” indicates the maximum flatness (noisiness) of the signal.

Johnston (1988b,a) applied the SFM to perceptual audio coding. Based on the idea of perceptual entropy (see Section 4.2), the model deployed the SFM for distinguishing between tone-like and noise-like maskers when calculating masking thresholds (Johnston, 1988b,a; Johnston and Kuo, 2003). The model used the short-time power spectrum with fixed analysis frame length for Bark-wide bands. All the SFM-based models used in perceptual audio coding are similar to...
this and commonly have the following form:

\[
\text{SFM}_{CB} = \frac{\prod_{k \in \text{CB}} S_{xx}(k)}{\sum_{k \in \text{CB}} S_{xx}(k)}
\] (5.4)

where \( M \) is the number of \( S_{xx}(k) \) components which fall within a critical band (CB) corresponding to the current subband.

Similarly, [Taghipour et al., 2013b] introduced the “partial spectral flatness measure” (PSFM) of tonality as the ratio of the geometric and the arithmetic means of the short-time squared magnitude spectrum, \(|S_{st}(k)|^2\), of a subband. In that (initial) model, the magnitude spectrum of each IIR-filter output was individually analyzed by a discrete Fourier transform (DFT) [Oppenheim and Schafer, 2010; Proakis and Manolakis, 2007], and the PSFM was calculated for a range of coefficients around its center frequency:

\[
\text{PSFM} = \sqrt[2N]{\prod_{k=N1}^{N2} |S_{st}(k)|^2} \frac{1}{N} \sum_{k=N1}^{N2} |S_{st}(k)|^2
\] (5.5)

where \( 0 \leq \text{PSFM} \leq 1 \), and \( N = N2 - N1 + 1 \). \( N1 \) and \( N2 \) were chosen in such a way that for each filter output the range extended to twice its effective bandwidth (details in [Taghipour et al., 2013b]). Figure 5.4 shows the psychoacoustic model with the initial PSFM.

In order to take into account the outcome of the psychoacoustic studies of Chapter 3 regarding the role of “center frequency” and “bandwidth,” four different block lengths were used for the spectral analysis (DFT): 512, 1024, 2048, and 4096. The higher block lengths were used for the filters with lower center frequencies and narrow widths. A block length of 512 samples was used

Figure 5.4: PM with the initial PSFM. Only one output channel of the filter bank is shown. 104 DFTs are applied to the outputs of the 104 filters. Spectral resolution of PSFM calculation varies with the DFT length; it is higher for low frequencies. Four different DFT lengths are used.
Table 5.1: Estimated entropy rates in kbit/s for different audio samples (Taghipour et al., 2013b).

<table>
<thead>
<tr>
<th></th>
<th>Castanets</th>
<th>Ger. male speech</th>
<th>Female vocal</th>
<th>Orchestra</th>
<th>Harpsi-chord</th>
<th>Pitch-pipe</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor</td>
<td>73.12</td>
<td>57.05</td>
<td>50.30</td>
<td>42.97</td>
<td>45.86</td>
<td>50.97</td>
<td>56.50</td>
</tr>
<tr>
<td>PSFM</td>
<td>73.21</td>
<td>49.88</td>
<td>53.71</td>
<td>47.37</td>
<td>52.71</td>
<td>42.07</td>
<td>71.10</td>
</tr>
</tbody>
</table>

for the highest center frequencies and the largest bandwidths. In doing so, the tendencies shown by the results of experiments of Chapter 3 are taken into account in the design of this measure, although not in the form of an explicit mathematical quantification. The assignment of the block lengths to the center frequencies was done such that the frequency resolution would lead to at least 8 coefficients (N) for the calculation of PSFM in that particular band.\(^1\)

### 5.2.1.1 Evaluation

The outcome of this model was compared to the “Predictor” (Taghipour et al., 2010) by means of a MUSHRA test (Taghipour et al., 2013b). A transform coding scheme was used, such as described in Section 4.1.2. An average entropy rate of 48 kbit/s was used for both models. Table 5.1 shows the individually estimated entropy rates for the six items used in the test and for the white noise signal, which was not included in the MUSHRA test.

24 normal-hearing subjects participated in the MUSHRA test, among which 14 were experts. Since the customer is typically a non-expert listener, we were interested in a comparison between the subjective evaluation by experts and non-experts. The results of one subject were not considered for the evaluation, because of atypical responses for the hidden reference and anchor (ITU-R BS.1534-1, 2003; ITU-R BS.1534-3, 2015). The results of the MUSHRA test with 23 subjects are depicted in Figure 5.5. No overall significant difference was found between the models.

### 5.2.1.2 Experts vs. non-experts

The subjective scores of the experts were compared to those of the non-expert subjects. Figure 5.6 shows the comparison. Average quality scores over all items are shown. The numbers 1 to 4 represent the hidden reference, the lowpass filtered anchor, Predictor, and PSFM, respectively. As expected, quality scores for hidden reference given by the expert listeners was higher (closer to 100) than those given by non-expert listeners. On average, they rated the quality of both coded versions lower than the non-expert subjects did. This is in accord with the literature (Schinkel-Bielefeld et al., 2013).

\(^1\) Only for the filters with the lowest center frequencies, N could fall below 8.
Figure 5.5: Results of the MUSHRA test: mean ratings across 23 subjects (Taghipour et al., 2013b). The abscissa shows different items and their conditions. For each item, the subjective quality ratings of different conditions are shown. The ordinate shows a quality scale between 0 (very bad) and 100 (excellent). Means and their corresponding confidence intervals (95%) are shown.

Figure 5.6: Mean ratings across subjects and items for 14 experts and 9 non-experts (Taghipour et al., 2013b). The numbers 1 to 4 represent the hidden reference, the lowpass filtered anchor, Predictor, and PSFM, respectively. Ordinate shows a quality scale between 0 (very bad) and 100 (excellent). 95% confidence intervals for the means are shown in the figure.
5.2.2 Optimized PSFM

As illustrated in Figure 5.4, in the initial model, “shaped” (filtered) outputs of the preceding filter stage were the inputs of the PSFM. This led to two shortcomings: first, since the PSFM was not calculated from the original input signal, even for a white noise input, the measure never reached the maximum flatness. Second, the calculation had high computational complexity whereby 104 fast Fourier transforms (FFT) were computed (Taghipour et al., 2013b, 2014b).

In order to overcome these problems, Taghipour et al. (2014b) further optimized the PSFM model. The block diagram of the PM including the optimized PSFM is shown in Figure 5.7. For the input signal, short-time spectra of different spectral resolutions are generated by DFTs of different lengths of 4096, 2048, or 1024 for low, middle and high frequencies, respectively. The PSFM is calculated corresponding to the individual band-pass filters, i.e., over twice their effective bandwidth. In doing so, only 3 FFTs had to be calculated for one set of tonality values.

5.2.2.1 Evaluation

Figure 5.8 shows the tonality/noisiness estimates of the PSFM for (a) Diverging tones, (b) Pitch-pipe, (c) Harpsichord, and (d) white noise. In comparison to Figure 5.3, the estimates of tonality/noisiness for the Diverging tones and the white noise signal computed by the PSFM seem to be more accurate.

The PSFM was compared to the Predictor. Both methods were used together with the psychoacoustic model. A transform audio coding scheme with MDCT was used, as described in Section 4.1.2 (Taghipour et al., 2014b). For subjective tests, the codecs under test were designed to have equal average entropy rates of 48 kbit/s (intermediate quality) and 32 kbit/s (low quality). Estimated entropy rates of the individual items are listed in Table 5.2.

By means of a MUSHRA test, the PSFM was compared to the Predictor. Initially, a pilot test was conducted with 3 items at an average entropy rate of 48 kbit/s, in which 5 non-expert listeners participated. The results showed a tendency for higher quality ratings for PSFM for all
Figure 5.8: Noisiness estimate by the PSFM for (a) Diverging tones, (b) Pitch-pipe, (c) Harpsichord, and (d) white noise.

3 items. However, there were no significant differences. Therefore, for the final MUSHRA test, we decided to test the items at two different average entropy rates. The hypothesis was that in a lower quality range the differences would be more distinct (Taghipour et al., 2014b).

Table 5.2: Estimated entropy rates in kbit/s for different items (Taghipour et al., 2014b). Average entropy rates of 48 and 32 kbit/s were chosen.

<table>
<thead>
<tr>
<th></th>
<th>Predict. (48)</th>
<th>PSFM (48)</th>
<th>Predict. (32)</th>
<th>PSFM (32)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female vocal</td>
<td>50.30</td>
<td>51.58</td>
<td>31.58</td>
<td>32.16</td>
</tr>
<tr>
<td>Ger. m. speech</td>
<td>57.05</td>
<td>47.73</td>
<td>36.39</td>
<td>29.76</td>
</tr>
<tr>
<td>Eng. m. speech</td>
<td>51.86</td>
<td>44.31</td>
<td>33.16</td>
<td>28.12</td>
</tr>
<tr>
<td>Ger. f. speech</td>
<td>61.93</td>
<td>53.88</td>
<td>42.66</td>
<td>36.53</td>
</tr>
<tr>
<td>Harpsichord</td>
<td>45.86</td>
<td>53.12</td>
<td>30.86</td>
<td>37.32</td>
</tr>
<tr>
<td>Castanet</td>
<td>73.12</td>
<td>69.08</td>
<td>51.61</td>
<td>48.84</td>
</tr>
<tr>
<td>S. orchestra</td>
<td>42.97</td>
<td>42.55</td>
<td>28.17</td>
<td>27.76</td>
</tr>
<tr>
<td>News intro</td>
<td>47.68</td>
<td>51.05</td>
<td>31.58</td>
<td>33.36</td>
</tr>
<tr>
<td>Pitch-pipe</td>
<td>50.97</td>
<td>48.69</td>
<td>38.00</td>
<td>35.51</td>
</tr>
<tr>
<td>Noise</td>
<td>56.50</td>
<td>66.10</td>
<td>34.65</td>
<td>45.39</td>
</tr>
</tbody>
</table>
For the final MUSHRA test, 9 items (monophonic with 48 kHz sampling frequency) with diverse characteristics were used in two groups: (1) speech and vocal and (2) music. The two groups of sounds were presented to the subjects in two separate sessions. Otherwise the test would have taken more than 50 minutes, including the training phase. Subjects took between 15 and 25 minutes for each session. Prior to the actual test, subjects were given an instruction text and were asked to take a similar MUSHRA training test with 2 audio items and 4 conditions. 15 expert subjects (average age 32) participated in the test. Audiometric thresholds were measured for frequencies up to 14 kHz using a software-based audiometer (SelfscreeningAudioMeter V1.32, HörTech GmbH, Oldenburg, Germany) and Sennheiser HDA 200 headphones. For all tested frequencies, the subjects had absolute thresholds in the range -20 to +20 dB relative to the mean absolute thresholds for normal-hearing subjects [ISO 389-7 (2005)]. The MUSHRA test was carried out diotically with a pair of open electrostatic Stax SR-507 headphones with a SRM-600 driver unit (Saitama prefecture, Japan) in an environment with a background noise of approximately 20 dBA.

Mean ratings across the 15 expert subjects are depicted in Figure 5.9a. The 9 audio items are listed on the abscissa including their 6 conditions. The graph shows average subjective quality ratings for the different conditions of the audio items on the ordinate. For each item, the mean quality scores across subjects and the corresponding 95% confidence interval are shown. Post hoc comparison tests (two-tailed paired t-tests) were carried out between the two “48 kbit/s”-coded versions and between the two “32 kbit/s”-coded versions. Significant differences ($p < 0.05$) are shown with “*” in Figure 5.9a. For most of the items coded at medium entropy rate, no significant difference was found between the methods. However, for the low entropy rate, for 7 items, PSFM was rated significantly higher than Predictor. This is in agreement with our hypothesis that greater differences should be observed at lower entropy rates.

5.2.2.2 Experts vs. non-experts

Additionally, we conducted the same MUSHRA test with 19 non-expert listeners. Based on the responses to the hidden reference and anchor, the results for 4 non-expert subjects were excluded from the final statistical analysis. They either did not understand the task sufficiently or had severe difficulty in distinguishing the hidden reference and anchor from the other conditions. The average age of the 15 non-expert subjects, whose data were used for the statistical analysis, was 23 [Taghipour et al. (2014b)].

Figure 5.9b shows the results of the MUSHRA test with non-expert listeners. As expected [Schinkel-Bielefeld et al. (2013)], the results for the different coded versions were closer to each other and to the hidden reference [Taghipour et al. (2014b)]. Significant differences are marked with “*”, as before. Here as well, the results for the same 7 items show significantly higher ratings for PSFM for the low entropy rate.
Quality ratings were higher for PSFM for non-speech signals, especially at the lower entropy rate \citep{Taghipour2014b}. For speech signals, despite higher compression (lower entropy rate, Table 5.2), PSFM achieved comparable quality to the other model.
CHAPTER 5. TONALITY ESTIMATION METHODS

5.3 Amplitude modulation ratio (AM-R)

5.3.1 The initial method

Amplitude modulation (AM) is mostly used in electronic communication (Proakis and Salehi, 2007). AM produces a signal from an information signal and a carrier wave. The envelope of the resulting signal is in accordance with the information contained in the original signal. Commonly, this envelope can be obtained by means of the “Hilbert Transform” (Chen et al., 2014; Proakis and Salehi, 2007). Considering this principle, Chen et al. (2014) introduced a measure for tonality estimation which will be referred to as amplitude modulation ratio (AM-R) (a similar model has been introduced by Tchorz and Kollmeier (2003) in the context of noise suppression).

The basic idea of using AM-R as a tonality measure is that the envelopes of the outputs of the filters used in the PM can indicate tonality: e.g., assuming that $|H(jw)|$ is the magnitude response of a filter, if $|H(jw_0)| \gg |H(-jw_0)|$, a sinusoidal input with frequency $w_0$ would lead to an output with almost constant envelope, i.e. no AM. This is the basis for distinguishing a tone from a superposition of multiple components which, in contrast, would lead to more envelope fluctuations or, in other words, to a strong AM.

Since psychoacoustic studies confirm that modulation has a strong influence on masking (Dau et al., 1997a; Moore, 2012; Taghipour et al., 2016c; Verhey, 2002; Verhey et al., 2003; Viemeister and Plack, 1993), this link between the degree of modulation and tonality could be exploited to locally adjust the level of the masking curve.

A basic block diagram for calculation of the masking threshold in the psychoacoustic model with the AM-R tonality estimation method is depicted in Figure 5.10. The figure shows three major steps for calculating the AM-R which are discussed in the following.

- Obtaining the envelopes of the filtered signals: using IIR complex-valued filters led to a simple computation of envelopes as the magnitudes of the filter outputs. For AM-R

![Figure 5.10](image)

Figure 5.10: Block diagram of the AM-R tonality estimation method. Only one output channel of the filter bank is shown. AM-R is calculated from the short-time magnitude spectral coefficients of the squared envelope of the filter output. Different DFT lengths are used for the analysis of the filter outputs at low, middle and high frequencies (4096, 2048, 1024, 512 and 256).
calculation, squared envelopes were used, since they resulted in better overall performance (Chen et al., 2014).

- Calculating DFTs for the envelopes: amplitude modulation spectra, which are represented by spectral components of the envelopes, were obtained by performing DFTs (fast Fourier transform in practice) on the filter output envelopes. Different spectral resolutions were generated by using different DFT lengths in different bands \((l = 4096, 2048, 1024, 512, 256)\), which led to higher spectral resolution for the calculation of AM-R in low frequency regions (Chen et al., 2014).

- Calculating the AM-R:

\[
\text{AM-R} = \frac{\sum_{k=1}^{\frac{l}{2}} |A[k]|}{|A[0]|} \tag{5.6}
\]

where \(A[k]\) is the \(k\)-th element of the current amplitude modulation spectrum. \(A[0]\) represents the average envelope, for \(1 \leq k \leq \frac{l}{2}\), and \(A[k]\) represents the variations around this average. Only a restricted spectral region is used for this formula, since modulation frequencies much higher than the bandwidth of a filter can only be caused by components located in other bands (Chen et al., 2014).

In this model, output values of Equation (5.6) were limited to a range between 0 and 1. Values close to 1 correspond to the noise-like characteristic of the signal at that specific spectral segment of the frame. Values close to 0 indicate tonal behavior of the signal in the corresponding spectral part of the frame (Chen et al., 2014).

### 5.3.1.1 Evaluation

Figure 5.11 shows the tonality/noisiness estimate of the AM-R for (a) Diverging tones, (b) Pitch-pipe (short), (c) Harpsichord, and (d) white noise. The AM-R was compared to the PSFM by means of a MUSHRA test. The two tonality estimation methods were implemented in the psychoacoustic model (see Chapter 4) which was applied to transform audio coding, as in Section 4.1.2. Estimated average entropy rates of 48 and 32 kbit/s were used (Chen et al., 2014).

In order to compare the two models in a comprehensive manner, test items were chosen which stress different aspects of the coder (i.e., the test items had various characteristics). All the items were monophonic and were sampled at 48 kHz. Table 5.3 shows the individually estimated entropies in both cases (48 kbit/s and 32 kbit/s) for all six items used in the test.

12 experienced subjects (11 males, 1 female) participated in the MUSHRA test. They were aged between 24 and 45 (average 30, median 28). Audiometric thresholds were measured for frequencies up to 14 kHz using a software-based audiometer (SelfscreeningAudioMeter V1.32, Dissertation, Armin Taghipour.
HörTech GmbH, Oldenburg, Germany) and Sennheiser HDA 200 headphones. For all tested frequencies, the subjects had absolute thresholds in the range -20 to +20 dB relative to the mean absolute thresholds for normal-hearing subjects (ISO 389-7, 2005). For the MUSHRA test, the stimuli were presented diotically via a pair of open electrostatic Stax SR-507 headphones with a SRM-600 driver unit (Saitama prefecture, Japan) in an environment with a background noise of approximately 21 dBA.

Table 5.3: Estimated entropy rates in kbit/s for different items [Chen et al., 2014].

<table>
<thead>
<tr>
<th>Items</th>
<th>PSFM (48)</th>
<th>AM-R (48)</th>
<th>PSFM (32)</th>
<th>AM-R (32)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female vocal</td>
<td>51.58</td>
<td>57.94</td>
<td>32.16</td>
<td>37.68</td>
</tr>
<tr>
<td>Ger. m. speech</td>
<td>47.73</td>
<td>57.27</td>
<td>29.76</td>
<td>37.65</td>
</tr>
<tr>
<td>Harpsichord</td>
<td>53.12</td>
<td>45.85</td>
<td>37.32</td>
<td>30.66</td>
</tr>
<tr>
<td>Castanet</td>
<td>69.08</td>
<td>79.72</td>
<td>48.84</td>
<td>57.96</td>
</tr>
<tr>
<td>Orchestra</td>
<td>42.55</td>
<td>44.85</td>
<td>27.76</td>
<td>28.30</td>
</tr>
<tr>
<td>Pitch-pipe</td>
<td>48.69</td>
<td>40.75</td>
<td>35.51</td>
<td>29.10</td>
</tr>
</tbody>
</table>
Figure 5.12: Mean ratings across 12 expert subjects (Chen et al., 2014). The abscissa shows different items (and their conditions). Subjective quality ratings for different conditions of the items are depicted. The ordinate shows a scale between 0 and 100 (spanning a quality range from bad to excellent). Mean values and confidence intervals (95%) are shown. Significant differences between PSFM and AM-R pairs are indicated by “*”.

The experiment involved comparing the subjective quality of different audio coding methods using the MUSHRA test. The subjects rated the quality of six test items under six different conditions, each coded with either PSFM or AM-R at two entropy rates: 48 kbit/s and 32 kbit/s. The ratings were compared using two-tailed paired t-tests, which showed significant differences at the 0.05 level for some conditions.

At both data rates, the ratings were higher for PSFM for Harpsichord and Pitch-pipe. However, for Castanets and Orchestra, AM-R was rated significantly higher than PSFM. For the item “German male speech”, at the low entropy rate (32 kbit/s), AM-R was rated significantly higher, while for the intermediate entropy rate (48 kbit/s) the two coded versions had comparable quality. Ratings were higher for AM-R for the non-stationary and complex signals while PSFM was rated higher for the stationary signals.
5.3.2 Optimized AM-R

The bandwidth of a noise is related to the rapidity of fluctuations in its envelope (Bos and de Boer, 1966; Taghipour et al., 2016c). Hence, envelope analysis using amplitude modulation allows an estimation of the effective bandwidth of the masking signal (based on the beating components of the AM of a complex tone). Exploiting the estimated bandwidth led to an optimization of the AM-R measure.

The psychoacoustic motivation comes from the dependence of the asymmetry of masking on bandwidth (Hall, 1997; Taghipour et al., 2016c). Hall (1997) suggested that, when the probe bandwidth exceeds the masker bandwidth, the temporal structure of the stimuli should be analyzed appropriately for the asymmetry of masking. The threshold decreases with decreasing masker bandwidth. In order to improve perceptual quality, the latter part was exploited here by weighting the beating (baseband) components of the AM spectrogram with a linearly increasing weighting function. With the filter center frequency being the carrier, the amplitude modulation has two beating components. Thus, it is meaningful to set an upper bound for the calculation, which should be twice the bandwidth of the current sub-band. This was considered in the design of a weighting function $w(k)$ which was added to the formula for the initial AM-R.

The optimized AM-R computes tonality after weighting the modulation frequency components of the modulation spectrogram with a perceptual weighting function $w(k)$. Hence, the optimized formula for the AM-R is as follows.

$$\text{AM-R} = \frac{\sum_{k=1}^{L} |w(k) \times A(k)|}{|A(0)|} \quad (5.7)$$

where $A(0)$ is the dc-component of the envelope and $A(k)$ refers to other beating components. The optimized AM-R is depicted in Figure 5.13.

![Figure 5.13: A block diagram of the AM-R in its optimized version. Only one output channel of the filter bank is shown. Short-time magnitude spectral coefficients of the squared envelope of the filter output are weighted for the calculation. Five different DFT length were used.](image)
5.3.3 Evaluation

Figure 5.14 shows the tonality/noisiness estimate of the optimized AM-R for (a) Diverging tones, (b) Pitch-pipe (short), (c) Harpsichord, and (d) white noise. The AM-R was compared to other tonality estimation methods by a MUSHRA test, which will be discussed in Section 5.5.3. In comparison to Figure 5.11 here the estimates are more towards tonality than noisiness. This can be helpful in cases where the masking threshold should be low to have a more conservative estimate if a signal is predominantly tonal (e.g., Pitch-pipe and Harpsichord).

5.4 Auditory image correlation (AIC)

The auditory image correlation (AIC) is a tonality measure (Taghipour et al., 2016a) based on the so called “auditory image model” (AIM) (Bleeck et al., 2004; Patterson, 2000; Patterson et al., 1992, 1995a; Patterson and Holdsworth, 1996). AIM is a model which simulates the auditory processing of sounds. A functional cochlea simulation transforms a sound into a multi-channel
activity pattern like that observed in the auditory nerve, and then this is converted into an “auditory image” (AI) that is intended to represent our initial impression of the sound (Patterson et al., 1992; Patterson and Holdsworth, 1996).

The AIC is a time-domain based tonality measure, by which the inter- and intra-channel correlations of the multi-channel activity pattern are exploited for the analysis of the envelope fluctuations. Since the PM already provides an auditory filter bank, the filter banks used in AIM were not employed. AIM and the AIC are discussed in the following.

5.4.1 Auditory image model (AIM)

As mentioned above, AIM is a model of auditory processing which was developed to analyze the auditory representation of sounds like music and speech and the sound environments in which they occur. The AI is constructed using three stages: a filter bank simulates the basilar membrane motion in the inner ear. Then, a multi-channel neural activity pattern is computed like that observed in the auditory nerve. Then a bank of temporal integration units, one per channel, converts this neural activity pattern into a dynamic AI (Patterson et al., 1992; Patterson and Holdsworth, 1996).

5.4.1.1 Basilar membrane motion (BMM)

The cochlea simulation is commonly carried out by means of a gammatone (Patterson et al., 1995b; Slaney, 1993) or a gammachirp (Irino and Patterson, 1997) auditory filter bank which performs a spectral analysis and converts the acoustic wave into a multi-channel representation of basilar membrane motion (Patterson and Holdsworth, 1996).

5.4.1.2 Neural activity pattern (NAP)

The outputs of the filters are converted into a neural activity pattern (NAP) which simulates the signal in the cochlear nerve (Patterson and Holdsworth, 1996). In other words, the NAP is a simulation of the signal flowing from the cochlea up the auditory nerve in response to a stimulus (Patterson, 2000). The transduction process includes compression\(^2\) and adaptation\(^3\), as well as suppression of lower activity areas. Additionally, the frequency dependence of phase locking\(^4\) is simulated using a lowpass filter.

---

\(^2\)A large range of input sound levels is nonlinearly compressed into a smaller range of responses on the BM (Moore, 2012).

\(^3\)The rapid decline of the initial high neural spike rate in response to a sound is referred to as adaptation (Moore, 2012).

\(^4\)Phase locking is the tendency for neural firings to occur at a particular phase of the stimulating waveform on the BM (Moore, 2012). Phase locking occurs mainly at low and medium frequencies (Moore, 2012; Schnupp et al., 2011).
5.4.1.3 Strobed temporal integration (STI)

One problem in modeling temporal integration is to determine how the auditory system can integrate over periods of the sound to form a stable auditory image without losing the fine-grain temporal information within the period (Patterson et al., 1992). This is done by a periodicity-sensitive temporal integration mechanism, so-called “strobed temporal integration” (STI), which stabilizes the neural activity pattern and removes global phase differences (Bleeck et al., 2004; Patterson, 2000). In each channel, the activity in the NAP is monitored, looking for large peaks. The adaptive triggering mechanism of the temporal integration tends to output one trigger pulse per period at the time of the largest peak in the period. The multi-channel result of the temporal integration is the AI (Patterson et al., 1992, 1995a).

5.4.2 Computation of AIC

As depicted in Figure 5.15, AIC uses the complex IIR filters of the PM instead of gammatone filters of the AIM (Taghipour et al., 2016a); the real parts of the filter outputs are taken as the BMM. The conversion of the BMM to the NAP was done using half-wave rectification, compression, and lowpass filtering (hcl). Here, half-wave rectification simulates the unipolar response of the hair cells. Even though complex 2D adaptation and suppression based models have been proposed (Bleeck et al., 2004; Patterson, 2000; Patterson and Holdsworth, 1996), here simple log compression was implemented to make the NAP level-independent (Krumbholz et al., 2001). The lowpass filtering, which simulates the progressive loss of phase locking with increasing frequency was implemented with a second order leaky integrator with a cutoff frequency of 1200 Hz (Bleeck et al., 2004; Patterson et al., 1995a).

Figure 5.15: Block diagram of the AIC measure. Only one output channel of the filter bank is shown. The real part of the filter output is used as BMM. hcl (half-wave rectification, compression, and lowpass filtering), acf (autocorrelation function), and cross-correlation with the neighboring channel follow. An AIC value of 0 indicates high tonality.
As the AIC is not designed to produce an AI, but only a tonality estimate, an autocorrelation function (acf) was employed instead of the models suggested for STI (Bleeck et al., 2004; Patterson, 2000). This can be explained in the following way. Based on an early computational version of “the duplex theory of pitch perception” (Licklider, 1951; Patterson, 2000) and Patterson et al. (1995a) suggested using a simple autocorrelation approach for the analysis of tonality or pitch extraction. As the acf is a similarity measure, and thus, helps in detecting repeating patterns in a signal, it is suitable for our purpose; i.e., strong tonality (or strong periodicity) leads to dominant maxima in the output of the acf at positions corresponding to its period.

Since the acfs are built from the filter outputs, the model indirectly takes into account the dependencies of tonality perception on bandwidth (Taghipour et al., 2012, 2013a, 2014a, 2016b).

As mentioned above, in general, tone-like signals can be well discriminated from noise-like signals by means of the maxima of the acf. In the case of the presented model, however, the outputs of the IIR filter bank to a white noise input are “colored” (filtered). Consequently, while white noise has a dirac-shaped acf, the shaped noise at the output of a PM filter has a smeared acf. Hence, errors occur in the estimation of noisiness. This suggests that the individual intra-channel acfs are not sufficient for tonality estimation in the context of masking asymmetry.

The model was revised to combine the inter- and intra-channel correlations. For this, a proposal by Hu and Wang (2004) and Hu and Wang (2011) was used which evaluates the similarities in acfs of the outputs of neighboring PM filters, reflecting their cross-correlation coefficient. This is shown in Figure 5.15. Further, the measure was normalized using the variances of individual channels. This approach normalizes the measure with respect to the white noise, i.e., it achieves maximum noisiness for the case of white noise.

A high intra-channel correlation and a high correlation with the neighboring channel indicate strong tonality. A mathematical modification was carried out to have a measure similar to the PSFM and the AM-R. These measures indicate strong tonality or strong noisiness by values close to 0 or 1, respectively. Thus, the AIC is defined by the following equation:

$$\text{AIC} = 1 - \frac{\sum_{k=1}^{N} \text{acf}_{c,k} \times \text{acf}_{c-1,k}}{\sqrt{\sigma_{c}^{2} \times \sigma_{c-1}^{2}}}$$

(5.8)

where \(\text{acf}_{c,k}\) is the autocorrelation output for channel “\(c\)” and sample index “\(k\)” and \(\sigma_{c}^{2}\) is the variance of the autocorrelation output of channel “\(c\)” (Taghipour et al., 2016a).

Additionally, frequency-dependent lengths of the acf sequences were implemented for different channels in order to take into account the effect of center frequency; however, no enhancement in the audio quality could be observed (in an informal listening test). Hence, it was excluded from the implementation in the current version.

Dissertation, Armin Taghipour
5.4.3 Evaluation

Figure 5.16 shows the tonality/noisiness estimates of the AIC for (a) Diverging tones, (b) Pitch-pipe, (c) Harpsichord, and (d) white noise. The spectro-temporal structure of the Diverging tones is not captured properly by the measure. The AIC was compared to other tonality estimation methods by a MUSHRA test, which will be discussed in Section 5.5.3.

5.5 Temporal envelope rate (TE-R)

Temporal envelope rate (TE-R) is introduced as another approach for analyzing envelope fluctuations in the time domain (Taghipour et al., 2016a). The prime motivation for the design of this approach was to have lower computational complexity. In contrast to the AIC, TE-R avoids the computation of entire autocorrelation functions with very long buffer lengths, which reduces complexity. The TE-R measure is based on two stages, with the first and second stages performing inter-channel and intra-channel correlation analysis, respectively. A combination of these individual measures is used to determine the final TE-R. This was based on objective analyses and informal listening tests, which are beyond the scope of this thesis.
5.5.1 Temporal envelope threshold crossing (TETC)

The temporal envelope threshold crossing (TETC) is an intra-channel temporal envelope measure. It was designed based on the idea of using threshold crossings to determine the noisiness of the temporal envelope: for every channel, the short-time average of the magnitude filter output (temporal envelope) is taken as a threshold value (similar to the value “zero” in zero crossing approaches). Fluctuations around this threshold are counted. The number of crossings is then normalized to a reference crossing value, which is obtained by averaging crossing numbers for a white noise input. A block diagram of the TETC is depicted in Figure 5.17. The fluctuation rate in the temporal envelope increases with increasing bandwidth (see Experiment III in Section 3.4); e.g., for a multi-tone stimulus (Verhey et al., 2003).

5.5.2 Temporal envelope cross-correlation (TECC)

The temporal envelope cross-correlation (TECC) is an inter-channel temporal envelope measure for the TE-R. The TECC analyzes the correlations between different channels of the magnitude temporal envelope outputs of the IIR filter bank. The number of channels for the estimation of the TECC is generally a design parameter. The immediate lower neighboring channel was considered for the current implementation. Additionally, as in Section 5.4.2, this inter-channel measure is utilized to enhance the capability of the intra-channel measure, i.e. TETC, such that

Figure 5.17: Block diagram of the TE-R. Only one output channel of the filter bank is shown. Short-time segments of the envelope are used for the computation of the TETC and the TECC. The TE-R is calculated from the harmonic mean of the TETC and the TECC.
the resulting combined measure, i.e., TE-R, correctly estimates the maximum noisiness value for a white noise PCM input. A block diagram of the TECC is shown in Figure 5.17. Finally, the TETC and the TECC are combined to give the TE-R. For this purpose, the harmonic mean of the TETC and TECC was used. This choice was based on initial objective evaluations \cite{Taghipour_2016a}. However, a weighted combination of the two measures has not been yet investigated systematically.

The TE-R gives a tonality/noisiness value which lies between 0 and 1, indicating strong tonality or strong noisiness, respectively (similar to the PSFM, the AM-R, and the AIC).

### 5.5.3 Evaluation

Figure 5.18 shows the tonality/noisiness estimate of the TE-R for (a) Diverging tones, (b) Pitch-pipe (short), (c) Harpsichord, and (d) white noise.

A MUSHRA test was carried out to compare AIC, TE-R, PSFM, and AM-R. The methods were implemented in the PM, as described in Section 4.1.2. A mean estimated entropy of

![Figure 5.18: Noisiness estimate by the TE-R for (a) Diverging tones, (b) Pitch-pipe, (c) Harpsichord, and (d) white noise.](image-url)
Table 5.4: Estimated entropy rates in kbit/s for different items.

<table>
<thead>
<tr>
<th>Item</th>
<th>AIC</th>
<th>TE-R</th>
<th>AM-R</th>
<th>PSFM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female vocal (es01)</td>
<td>58.40</td>
<td>59.39</td>
<td>61.68</td>
<td>57.06</td>
</tr>
<tr>
<td>Eng. m. speech (es04)</td>
<td>61.44</td>
<td>56.25</td>
<td>54.36</td>
<td>49.97</td>
</tr>
<tr>
<td>Harpsichord (si01)</td>
<td>29.87</td>
<td>30.40</td>
<td>34.31</td>
<td>38.13</td>
</tr>
<tr>
<td>Castanet (si02)</td>
<td>72.17</td>
<td>90.73</td>
<td>79.86</td>
<td>73.32</td>
</tr>
<tr>
<td>S. orchestra (sc02)</td>
<td>39.29</td>
<td>47.82</td>
<td>49.28</td>
<td>45.39</td>
</tr>
<tr>
<td>Pitch-pipe</td>
<td>26.87</td>
<td>24.33</td>
<td>29.24</td>
<td>31.04</td>
</tr>
<tr>
<td>Noise</td>
<td>78.90</td>
<td>96.04</td>
<td>91.29</td>
<td>70.85</td>
</tr>
<tr>
<td>Average (6 test items)</td>
<td>48.01</td>
<td>51.49</td>
<td>51.46</td>
<td>49.15</td>
</tr>
</tbody>
</table>

approximately 48 kbit/s was used. Table 5.4 shows a list of the items used in the MUSHRA test (plus noise) with the corresponding estimated entropy of each condition.

16 well-trained subjects with normal hearing participated in the test. However, an inspection of the results showed that the results of two subjects should be discarded. For several items, they had missed the hidden reference and/or the low anchor. Additionally, their ratings for the coded conditions showed extremely low variability. Hence, the final statistical analysis was based on the results of 14 well-trained subjects: 13 male and 1 female, aged between 21 and 39 (mean 30.1, median 30).

The results of the MUSHRA test are depicted in Figure 5.19. For each item and condition, the mean rating across subjects and the corresponding 95% confidence interval is shown. The scores for the hidden reference and the low anchor are represented by orange and magenta, respectively. Red, green, blue, and black represent AIC, TE-R, the AM-R and the PSFM, respectively.

To analyze the results, a two-way repeated-measures ANOVA was carried out with factors “item” and “condition.” A significant effect of item was found \[ F(5, 65) = 10.0, p < 0.001 \]. The effect of condition was also significant \[ F(5, 65) = 148.6, p < 0.001 \]. There was a significant interaction between item and condition \[ F(25, 325) = 8.4, p < 0.001 \].

It can be observed in Figure 5.19 that the hidden reference was rated significantly higher than all the other conditions; similarly, the low anchor scored significantly lowest (all pairs \( p < 0.001 \)). Since the focus was to compare the four coded versions to each other, an additional two-way repeated-measures ANOVA was carried out with the same factors, but only on data for the four coded conditions. A significant effect of item was found \[ F(5, 65) = 10.0, p < 0.001 \]. A significant effect of condition was found \[ F(3, 39) = 31.1, p < 0.001 \]. There was a significant interaction between item and condition \[ F(15, 195) = 8.6, p < 0.001 \].

Post hoc pairwise comparison tests (Fisher’s LSD) between the items showed that Female

---

\( ^6 \)

The entropy rates for PSFM in Table 5.4 slightly differ from those in Table 5.3. PSFM was modified to take into account spectral spreading. This was a minor modification.
vocal scored significantly higher than all other items (all pairs \( p < 0.05 \)), but Symphony orchestra (\( p = 0.221 \)). Pitch-pipe was rated significantly lower than all other items (all pairs \( p < 0.05 \)), but Castanets (\( p = 0.608 \)). Whereas no significant difference was found between Castanets and Harpsichord (\( p = 0.130 \)), castanets scored significantly lower than English male speech and Symphony orchestra (all pairs \( p < 0.05 \)). No other significant differences were found. Figure 5.19 indicates that the low rating for Castanets is an overall effect (i.e., for all conditions). On the other hand, the overall low rating for pitch pipe seems to be due to low scores for AIC and TE-R.

Post hoc pairwise comparisons (Fisher’s LSD) between the four coded conditions showed that AM-R and the PSFM were rated significantly higher than AIC and TE-R (all pairs \( p < 0.001 \)). There was no significant difference between AIC and TE-R (\( p = 0.820 \)). Similarly, AM-R and PSFM were not rated significantly different (\( p = 0.666 \)).

Due to the interaction between item and condition, it seemed reasonable to analyze possible differences between the conditions, item by item. A separate one-way repeated-measures ANOVA was conducted for each item, with 6 conditions. The effect of condition was sig-
nificant [all $F(5, 65)s > 69.3, p < 0.001$] for all items. Post hoc comparison tests (Fisher’s LSD) showed that the hidden reference was rated significantly higher than the other conditions (all pairs $p < 0.05$). The lower anchor was rated significantly lower (all pairs $p < 0.001$) than the other conditions.

For every item, the differences between the four coded conditions were further analyzed by one-way repeated-measures ANOVA with parameter “condition” (4 conditions) and post hoc comparison tests. For Female vocal, there was a significant effect of condition [$F(3, 39) = 4.0, p < 0.05$]. Post hoc comparisons showed that PSFM was rated significantly higher than AIC ($p < 0.05$) and TE-R ($p < 0.01$), even though this method showed the lowest estimated entropy for this item among the 4 conditions; see Table 5.4. No further significant differences were found.

For English male speech [$F(3, 39) = 0.5, p > 0.5$] and Symphony orchestra [$F(3, 39) = 2.0, p > 0.1$], no significant effect of condition was found.

For Harpsichord, there was a significant effect of condition [$F(3, 39) = 28.9, p < 0.001$]. Post hoc comparisons showed that both the AM-R and the PSFM were rated significantly higher than both AIC and TE-R (all pairs $p < 0.01$). TE-R was rated significantly higher than AIC ($p < 0.05$). No significant difference was found between the two best conditions.

For Castanets, there was a significant effect of condition [$F(3, 39) = 3.1, p < 0.05$] when sphericity was assumed. However, use of the Greenhouse-Geisser correction led to no significance ($p = 0.072$).

For Pitch-pipe, there was a significant effect of condition [$F(3, 39) = 25.6, p < 0.001$]. LSD tests showed that the PSFM was rated significantly higher than AIC, TE-R, and AM-R (all pairs $p < 0.01$). It can be observed in Table 5.4 that PSFM allocates more (estimated) bits to this item than other methods. AM-R was rated higher than AIC and TE-R (all pairs $p < 0.01$). No significant difference was found between the AIC and TE-R.

### 5.6 Conclusions

The results of Section 5.5.3 show that, overall, AM-R and PSFM scored significantly higher than AIC and TE-R, but were comparable to each other. When analyzed item by item, PSFM was rated significantly higher than AM-R in the case of Pitch-pipe. For Female vocal, AIC and TE-R were rated significantly lower than PSFM, but were not different from AM-R. No other significant difference was found between PSFM and AM-R. The analysis of the results of the MUSHRA tests, which were presented in different sections of this chapter, indicate that both PSFM and the AM-R are effective methods for tonality estimation.

Further modifications might optimize the Predictor, the AIC, and the TE-R, such that they score higher in subjective quality tests when used in psychoacoustic models of perceptual audio
encoders. However, their current state is not fully competitive with the AM-R and the PSFM. The varying analysis lengths for the different frequency channels in the PSFM and the AM-R should also be incorporated for the cases of the Predictor, the AIC, and the TE-R.

Whereas the AM-R is an envelope analysis measure, the PSFM is a power spectrum analysis approach. However, the PSFM still possesses two links to the temporal structure of the input signal. First, the ranges over which SFMs are calculated are related to the bandwidths of the subbands, which in turn are related to the modulation frequencies in the envelope of the signal. Second, the DFT lengths vary for different spectral regions: longer and shorter blocks are used for low and high frequencies, respectively. The results of the MUSHRA tests with the PSFM (see Sections 5.2.2, 5.3.1, and 5.5.3), indicate that this flexibility is an appropriate correction for the fact that the “original SFM” (ISO/IEC 11172-3, 1993; Johnston, 1988b,a) is a purely frequency-domain, power spectrum measure and hence, is not able to model all the effects of the envelope fluctuations on masking (Hall, 1997; Painter and Spanias, 2000; Taghipour et al., 2016c).
Chapter 6

Comparison of the PM to a conventional (MPEG) PM

In order to compare the PM to a conventional PM (ISO/IEC 11172-3, 1993; ISO/IEC 13818-7, 1997), the model was applied to a standard codec. Both versions of the PM were tested: the initial Bark-based model and the ERB\textsubscript{N} version (based on the ERB\textsubscript{N}-number scale and roex filters); see Chapter 4 for details. In this chapter, the standard coding scheme is discussed briefly. Further, the results of a MUSHRA test are presented and analyzed.

6.1 Coding scheme

Desikan and Taghipour (2016) integrated the proposed PM in both its versions (Bark-based and ERB\textsubscript{N}-based) in the MPEG-HE\textsuperscript{1}-AAC v2 codec (Herre and Dietz, 2008; ISO/IEC 14496-3, 2009) which is a transform-based perceptual audio codec like the one in Figure 4.1. However, in comparison to the simple block diagram in Figure 4.1, the MPEG-HE-AAC v2 codec incorporates a number of additional coding tools. For the detailed architecture of the MPEG-HE-AAC system, see ISO/IEC 14496-3 (2009).

The main part of an MPEG-HE-AAC encoder is an MPEG-AAC block (Bosi et al., 1997; Brandenburg, 2002; ISO/IEC 13818-7, 1997), which uses the MDCT transform and the window switching technique (Edler, 1989). The analysis filter bank is followed by the so called “temporal noise shaping” (TNS) module. The TNS implements a pre-echo control strategy. It performs a filtering of the spectral coefficients along the frequency axis to implement prediction in the frequency domain (Herre and Johnston, 1996). This permits the coder to exercise control over the temporal fine structure of the quantization noise within each filter bank window (Bosi et al.,

\textsuperscript{1}High efficiency
The TNS improves the coding quality of transient signals by shaping the noise in the time domain. Some pre-echo artifacts can be suppressed this way.

Time-domain prediction follows, which exploits the correlation of a given spectral coefficient in the current frame with the spectral coefficients of the preceding frames. The prediction improves the redundancy reduction of stationary components.

In parallel, perceptual noise substitution (PNS) is incorporated which permits a frequency selective parametric coding of noise-like signal components. The PNS technique is based on the fact that the temporal fine structure of noise-like audio signals is of minor perceptual importance to the human auditory system. Rather than a large set of quantized spectral coefficients, the so called “PNS flag” is transmitted plus the energy information for the selected scale factor band.

Apart from these integrated blocks, MPEG-HE-AAC supports spectral band replication (SBR) as a part of the pre-processing stage of the encoder and post-processing in the decoder. The SBR is a bandwidth extension technology based on the observation that the high frequencies in the spectrum of an audio signal provide only a marginal contribution to the “perceptual information” contained in the signal. SBR exploits this observation to improve the compression; instead of transmitting the upper part of the spectrum, the SBR produces high-frequency components in the decoder from the lower part. The SBR does not explicitly take into account the masking properties of the human auditory system. Therefore, using the SBR would make the assessment of the PM difficult. When the SBR is activated, no information regarding the high frequency components of the input signal is available to the PM; more on this later.

The spectral coefficients are quantized and coded with a masked threshold predicted by a conventional, state-of-the-art psychoacoustic model, which is built upon the notion of perceptual entropy and a “measure of chaos”. The quantization of the spectral coefficients is handled by two-fold nested iteration loops using the analysis-by-synthesis method. In these two-fold loops, the inner loop optimizes the quantizer steps for the available number of bits, while the outer loop optimizes the scale factors by satisfying the noise-to-mask ratio demands of the psychoacoustic model. In order to limit the side information for quantizer control, the MDCT coefficients are grouped into scale factors for each band. For each scale factor band, a non-uniform quantization step size is calculated and the corresponding scale factor is transmitted. Additionally, a mechanism called “bit reservoir” is used to limit the side information for quantizer control.
allows for a locally varying bit rate to cater for the demands of individual frames, while achieving a globally constant bit rate (Bosi et al., 1997; ISO/IEC 11172-3, 1993).

In order to accomplish a versatile integration of the proposed psychoacoustic model in the coding scheme and an unbiased comparison of the quality of the two PMs, the following customizations were incorporated in the existing MPEG-HE-AAC standard codec. As mentioned above, the SBR leads to a version of an audio signal which is bandlimited to low frequencies in the “core” coder. However, it is of strong interest to assess the behavior of the PM also at high frequencies. Hence, as a first step towards the customization, SBR was disabled from the coding chain. Hence, the AAC core operated on the full bandwidth.

A parametric coding scheme takes into account other information for quantizer control. The current PM is not optimized for parametric coding technology. Using the PNS module for parametric coding of noise-like components hinders assessment of the efficiency of the PM in estimating the masking thresholds, which is sensitive to the tonality/noisiness of the input signal; a coding technique which is case-sensitive to the “noise-like” components is not appropriate in this context. Hence, the PNS module was disabled in the codec for both versions of the PM.

As explained in Chapter 4, the PM is designed to work with both short and long coding windows. This makes it highly compatible with the “window switching” technique used in the codec (Edler, 1989; ISO/IEC 13818-7, 1997), which is a pre-echo control strategy (see Section 4.1.1). Therefore, no additional pre-echo control strategy was incorporated, such as the TNS; the TNS module was turned off for both versions of the PM.

All in all, four versions of the codec were prepared:

- The original MPEG-HE-AAC v2, which includes all the coding tools. This codec will be referred to as “MPEG-codec.”
- A customized version of the MPEG-HE-AAC v2 codec, for which the TNS, the PNS, and the SBR were disabled. This codec will be referred to as “MPEG-custom”.
- A version with the Bark based PM including the PSFM tonality measure, for which the TNS, the PNS, and the SBR were disabled as well. This codec will be referred to as “Bark-PSFM.”
- A version with the ERB\(_N\) based PM including the PSFM tonality measure, for which the TNS, the PNS, and the SBR were disabled as well. This codec will be referred to as “ERB\(_N\)-PSFM.”

The three latter versions of the codec have exactly the same coding scheme, so that a direct comparison of their quality by means of a MUSHRA test is plausible (see below). Further, including the condition “MPEG-codec” in the MUSHRA test allows for comparison of the PM to a standard reference codec which uses additional coding tools.

Dissertation, Armin Taghipour
6.2 MUSHRA test

Seven MUSHRA test items were used: Female vocal (es01), German male speech (es02), Harpsichord (si01), Castanets (si02), Glockenspiel (sm02), News intro (sc03), and Pitch-pipe. All the test items were monophonic and sampled at 48 kHz. A bit rate of 40 kbit/s was used for all codecs. For each item, the subjects had to rate six conditions: hidden reference, lower anchor, MPEG-codec, MPEG-custom, Bark-PSFM, and ERB_N-PSFM. The room had a background noise of 23 dBA.

19 normal-hearing expert subjects participated in the MUSHRA test. Since they frequently participate in MUSHRA tests, no training was given. Nevertheless, they read a page of instruction about the specifics of the test. Inspection of the results indicated that all data were valid according to the recommendations of ITU-R BS.1534-1 (2003). The subjects were aged between 21 and 47 years (mean 31.7, median 31): 3 females, 26 males.

6.2.1 Results

Figure 6.1 shows the mean results across the 19 subjects for each item and condition and the corresponding 95% confidence intervals, using parametric statistics. The scores for the hidden reference and the lower anchor are represented by orange and magenta, respectively. Red, green, blue, and black represent MPEG-codec, MPEG-custom, Bark-PSFM, and ERB_N-PSFM, respectively.

A two-way repeated-measures ANOVA was carried out with parameters “item” and “condition.” A significant effect of item was found \( F(6, 108) = 8.5, p < 0.001 \). A significant effect of condition was also found \( F(5, 90) = 304.4, p < 0.001 \). There was a significant interaction between item and condition \( F(30, 540) = 4.5, p < 0.001 \).

It can be observed in Figure 6.1 that the hidden reference and the lower anchor were rated significantly higher and lower than all other conditions, respectively (all pairs \( p < 0.001 \)). Since the focus was to compare the four coded version to each other, an additional two-way repeated-measures ANOVA was carried out with the same parameters, but only on data for the four coded conditions. A significant effect of item was found \( F(6, 108) = 8.9, p < 0.001 \). A significant effect of condition was also found \( F(3, 54) = 11.0, p < 0.001 \). There was a significant interaction between item and condition \( F(18, 324) = 2.2, p < 0.05 \), when sphericity was assumed. A Greenhouse-Geisser correction, however, showed no significant interaction between the two factors \( p = 0.057 \).

Further analyses were carried out in the form of item-wise ANOVAs and post hoc pairwise comparison tests (Fisher’s LSD). Post hoc pairwise comparison tests between the items showed that Castanets (si02) was rated significantly lower than all other items (all pairs \( p < 0.05 \)). Both Female vocal (es01) and Pitch-pipe were rated significantly higher than both German male
speech (es02) and News intro (sc03) (all pairs $p < 0.05$). No other significant differences were found between the items.

Post hoc pairwise comparisons between the four coded conditions showed that MPEG-custom was rated significantly lower than all the other conditions (all pairs $p < 0.01$). No significant difference was found between MPEG-codec, Bark-PSFM, and ERB$_N$-PSFM (all pairs $p > 0.224$).

Possible differences between the four coded conditions were further analyzed, item by item. For Female vocal (es01), a one-way repeated-measures ANOVA showed no significant effects of the parameter “condition” [$F(3, 54) = 1.1, p > 0.3$].

For German male speech (es02), there was a significant effect of condition [$F(3, 54) = 4.0, p < 0.05$]. Post hoc pairwise comparison tests showed that MPEG-custom was rated significantly lower than all the three other coded conditions (all pairs $p < 0.05$). No significant differences were found between MPEG-codec, Bark-PSFM and ERB$_N$-PSFM.
For Harpsichord (si01) \( [F(3, 54) = 0.5, p > 0.6] \), News intro (sc03) \( [F(3, 54) = 0.3, p > 0.7] \), and Pitch-pipe \( [F(3, 54) = 1.6, p > 0.2] \) no significant effect of condition was found.

For Castanets (si02), there was a significant effect of condition \( [F(3, 54) = 6.1, p < 0.001] \). Post hoc comparisons showed that Bark-PSFM scored significantly higher than MPEG-codec \( (p = 0.045) \). MPEG-custom was rated significantly lower than Bark-PSFM \( (p < 0.001) \) and ERB\( N \)-PSFM \( (p < 0.01) \). No other significant differences were found.

For Glockenspiel (sm02), there was a significant effect of condition \( [F(3, 54) = 3.5, p < 0.05] \). Post hoc LSD tests showed that MPEG-custom was rated significantly lower than MPEG-codec \( (p = 0.038) \) and Bark-PSFM \( (p = 0.034) \). No other significant differences were found.

### 6.3 Discussion

The results showed that, overall, MPEG-custom was rated significantly lower than all the other three coded conditions, but that no significant differences were found between MPEG-codec, Bark-PSFM, and ERB\( N \)-PSFM.

Since the coding algorithms for MPEG-custom, BARK-PSFM, and ERB\( N \)-PSFM codecs were identical, the analysis of the results indicates that the psychoacoustic models with PSFM estimated the masking threshold more efficiently and more accurately than the conventional psychoacoustic model. The quality difference between the two MPEG codecs can be explained due to a number of coding tools (i.e., TNS, PNS, and SBR) which were switched off for MPEG-custom. The PM proposed here, in either version (i.e., Bark or ERB\( N \)), was comparable to the MPEG-codec, which used additional coding tools, while these coding tools were switched off in the case of the proposed model. This means that, without the additional coding tools, our models scored better than the MPEG psychoacoustic model; when those additional coding tools were used in the MPEG coding scheme (i.e. MPEG-codec), our models were rated as high as the MPEG codec.

The compatibility of the presented psychoacoustic model with the additional coding tools (i.e., TNS, PNS, and SBR) should be further analyzed and improved. The ERB\( N \)-based psychoacoustic model did not score significantly differently from the Bark-based model.

At low center frequencies, the bandpass filters of the ERB\( N \)-based PM possess narrower widths than those in the Bark model (see Section 2.4). This leads to differences in the estimated levels of the masking thresholds. However, for the purpose of quantizer control, these thresholds are mapped onto a Bark scale, since the scale factor bands in the standard codec are designed in such a way. This might be disadvantageous for the ERB\( N \)-based PM. Further investigation of this hypothesis should be carried out. The model could be developed further to optimize the mapping. Additionally, the application of an ERB\( N \)-based coding scheme should be investigated.
Chapter 7

Conclusions

The experiments on the perception of tonality (see Chapter 3) showed that the discrimination of tonality and noisiness depends on the stimulus duration: the longer the stimuli are presented, the easier it is to distinguish the noise from the tone. An analysis of the role of bandwidth and center frequency indicated that there is a significant effect of both parameters. Duration thresholds decreased with increasing center frequency up to 1456 Hz. Further, duration thresholds decreased with increasing bandwidth when it was expressed in “fraction of bandwidth of the auditory filter.” However, further analysis of the results showed that much of the variability in duration thresholds is accounted for by the absolute width of the noise bands in Hz. This means that the task was performed mainly by detecting amplitude fluctuations in the noise and that the duration had to be sufficient for a detectable amplitude fluctuation to occur.

A comparison of the results of Experiments I and II (see Chapter 3) revealed that isolated noise bursts seem to be easier to distinguish from tone bursts of the same duration than noise bursts embedded in (longer) tone bursts from (longer) tone bursts. This might be due to temporal uncertainty in the latter case or due to forward and backward masking of the noise burst by the surrounding tone.

Whereas the effects of duration, bandwidth, and center frequency were experimentally tested in Experiments I and II, the effect of the parameter “intensity” (or, alternatively, parameter “loudness”) was not investigated. This was due to the fact that, in perceptual audio coding, no control of the playback intensity is available when implementing the system [Painter and Spanias, 2000].

In the development of the tonality estimation methods “PSFM” and “AM-R,” frequency-dependent analysis durations were introduced to take account of the fact that duration thresholds decrease with increasing center frequency and width of the subbands. Shorter buffers were used for filters with higher center frequencies and higher bandwidths.

Experiment III, which was described in Section 3.4 dealt with the most relevant case of
masking for perceptual audio coding, namely when the masker had a narrower bandwidth than the probe. The results showed a probe-to-masker ratio of -23 to 4 dB depending on the condition. Further analysis of the results showed significant effects of bandwidth and center frequency. Masked thresholds were lowest for medium frequencies and increased with increasing masker bandwidth. In other words, it was easier to detect the probe at medium frequencies and when the masker had a narrower bandwidth or, equivalently, lesser amount of fluctuations in its envelope.

In the design of the psychoacoustic model (Chapter 4) and the tonality estimation methods (Chapter 5), this effect of center frequency was not considered explicitly. It was assumed that modeling the absolute threshold of hearing already accounted for the higher sensitivity of the system for medium center frequencies. In future works, this should be tested systematically by means of simulations.

It was discussed that a combination of the results of the studies on asymmetry of masking (Gockel et al., 2002; Hall, 1997; Hellman, 1972; Krumbholz et al., 2001; Verhey, 2002) gives an almost complete picture of this phenomenon. However, the role of the center frequency has not been tested systematically in the context of the asymmetry of masking. As mentioned above, in Experiment III, the effect of center frequency was investigated when the masker had a narrower bandwidth than the probe (Taghipour et al., 2016c). However, comprehensive data are not available for the case when the probe has a narrower bandwidth than the masker. Although this is not very relevant to perceptual audio coding, it might be worth investigating from a psychoacoustic perspective.

Another aspect of masking was discussed partially by Hall (1997), Krumbholz et al. (2001), and Verhey (2002), namely the “type” of fluctuation of the envelope of the audio signal in each subband. As discussed before, the envelope fluctuations reflect the way that the signal varies in amplitude over time. The extent to which one sound masks another is partly determined by the similarity of their envelope fluctuations (Dau et al., 1997a; Sek and Moore, 2002). In psychoacoustic studies, one way of changing the fluctuation rate of the envelope is to change the width of a noise band. Such fluctuations are typically irregular. The narrower the bandwidth, the slower are the envelope fluctuations. This approach was used in this thesis, such as done by Hall (1997) and Verhey (2002). However, an important and relatively unexplored aspect of the envelope is its regularity in time. As a counterpart to the first approach, regular envelope fluctuations can be achieved by imposed amplitude modulation (AM) on a signal (Dau et al., 1997a). The author suggests a comparison of the effectiveness of masking produced by regular envelope fluctuations and by irregular fluctuations of the same mean rate.

In the course of this research, five different tonality estimation methods were proposed, designed and implemented in a psychoacoustic model. Various versions and modifications of the methods were discussed in Chapter 5. They aimed to improve the accuracy of the estimation of masking thresholds and the controlling of bit allocation for the time-frequency segments in
CHAPTER 7. CONCLUSIONS

Two of the methods proved to be more successful than the others when evaluated in subjective quality MUSHRA tests, namely the partial spectral flatness measure (PSFM) and the amplitude modulation ratio (AM-R). Whereas the former method is based on analysis of the short-term power spectrum, the latter is an envelope analysis method. Both methods are subband sensitive; i.e., they take into account the role of the center frequency and bandwidth of the subband in masking. In the PSFM, the lengths of DFTs reduce from low to high frequencies, which leads to higher spectral resolution for low frequencies and higher temporal resolution for high frequencies. Further, the area over which the SFM is calculated is bandwidth-dependent. In the AM-R, the amount of fluctuations in the envelope of the input signal is analyzed directly. Thus, the amount of amplitude modulation is incorporated in the estimation of tonality and, consequently, masking threshold.

The three other methods, namely “Predictor,” “AIC,” and “TE-R” (see Chapter 5) need further improvements. In its current version, the Predictor is not compatible with the complex filter bank. Downsampling, modulation and the polyphase structure of the Predictor led to a high computational complexity. Further, the combination of the intra- and inter-channel measures in AIC and TE-R is an initial proposal. Whereas the author is confident of the need for considering both intra- and inter-channel characteristics, the exact portions have to be investigated in the future.

While the initial psychoacoustic model was designed with a “Bark”-based filter bank, another version of the model was designed and implemented based on “ERB_N.” The MUSHRA test in Chapter 6 showed no significant difference between subjective quality ratings for the two models for intermediate bit rates. The author suggests a further investigation in this area: in their current versions, the models should be compared to each other in the cases of near-transparent coding (e.g., 60 kbit/s) and low quality coding (e.g., 32 kbit/s).

The scale factor bands in the MPEG-AAC encoder and decoder are based on the Bark scale. This might have an influence on the quality of the coding when using masking thresholds estimated by an ERB_N-based model. Since this thesis does not deal with the design of the codec itself, no modification and no investigation was carried out in the coding scheme. The fact that the ERB_N-based model estimates the masking threshold with narrower filters at low frequencies might not be appropriately exploited in the quantizer control. This can be investigated in future work. A possible design for a study to investigate the effects of variations in the PM and the codec design systematically (with respect to the Bark and ERB_N systems) could be as follows. The Bark-based psychoacoustic model, as well as the ERB_N-based model can be implemented in both a Bark-based codec and an ERB_N-based codec.

It should be noted that this work introduced a “first” version of the ERB_N-based psychoacoustic model, which could be improved. The filter design, which is based on a rounded-exponential
function, can be optimized. Although the suggested formulae presented in the literature are somewhat similar to each other (Moore and Glasberg, 1987; Glasberg and Moore, 1990; Patterson, 1976), the parameter choices vary. Most of the publications deal with perceptual data at only a few center frequencies (Moore and Glasberg, 1987; Glasberg and Moore, 1990, 2000; Patterson, 1974, 1976). However, the design of a filter bank requires estimation for all center frequencies. Especially, the frequency-dependent “shapes” of the filters (Glasberg and Moore, 2000), not the placement of the 3-dB cut-offs, were challenging. Further work is needed here. It should be mentioned that using a gammatone filter bank was tested and was less successful in objective tests, simulations and informal listening tests. No MUSHRA test was done including a gammatone filter bank design. The author suggests further optimization of the ERB\textsubscript{N}-based filter bank for the psychoacoustic model.

In their current form, the Bark-PSFM and ERB\textsubscript{N}-PSFM were comparable to the “MPEG-codec” (see Chapter 6), even though less coding tools were activated in their coding path (e.g., TNS). Compared to an MPEG-custom version with the exact same coding path, both Bark-PSFM and ERB\textsubscript{N}-PSFM were rated higher.

The results of the studies in this thesis suggest that filter bank-based models with specially designed spectral analysis and frequency- and bandwidth-dependent tonality estimation provide a promising approach to improve perceptual coding systems. The author recommends such psychoacoustic models instead of the state-of-the-art transform-based psychoacoustic models for conventional audio codecs. As for the tonality estimation, the author recommends the use of either the PSFM or the AM-R in perceptual audio encoders.
References


Dissertation, Armin Taghipour


REFERENCES


Dissertation, Armin Taghipour


REFERENCES


Dissertation, Armin Taghipour


Taghipour, A., Knölke, N., Edler, B., and Ostermann, J. (2010). Combination of different perceptual models with different audio transform coding schemes: implementation and evaluation. 129th AES Convention, San Francisco, USA. paper number 8283.


