Perceptual Modelling for Low-Rate Audio Coding

Christopher R. Cave

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

June 2002

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Engineering.

2002/06/18

Your file Votre référence
ISBN: 0-612-85882-0
Our file Notre référence
ISBN: 0-612-85882-0

In loving memory of my grandfather,
George W. Ledoux

Abstract

Sophisticated audio coding paradigms incorporate human perceptual effects in order to reduce data rates, while maintaining high fidelity of the reconstructed signal. Auditory masking is the phenomenon that is the key to exploiting perceptual redundancy in audio signals. Most auditory models conservatively estimate masking, as they were developed for medium to high rate coders where distortion can be made inaudible. At very low coding rates, more accurate auditory models will be beneficial since some audible distortion is inevitable.

This thesis focuses on the application of human perception to low-rate audio coding. A novel auditory model that estimates masking levels is proposed. The new model is based on a study of existing perceptual literature. Among other features, it represents transient masking effects by tracking the temporal evolution of masking components. Moreover, an innovative bit allocation algorithm is developed that considers the excitation of quantization noise in the allocation process. The new adaptive allocation scheme is applicable with any auditory model that is based on the excitation pattern model of masking.

Sommaire

Des algorithmes sophistiqués de codage audio tirent profit des effets de perception humaine afin de réduire le débit de transmission. L'application du masquage auditif constitue la base pour exploiter la redondance perceptuelle des signaux audio. Généralement, les modèles auditifs estiment l'effet de masque de façon conservatrice parce que ceux-ci ont été conçus pour des codeurs à débit moyen et élevé, lesquels auront suffisamment de bits pour rendre la distorsion inaudible. À bas débit, la précision des modèles auditifs devra être améliorée pour contrecarrer une certaine distorsion audible inévitable.

Cette thèse étudie l'intégration des effets de perception humaine aux codeurs à bas débit binaire. Un nouveau modèle auditif qui estime le niveau de masquage produit par un signal audio est présenté. Entre autres caractéristiques, le modèle représente les effets transitoires du masquage en suivant l'évolution temporelle des composantes de masquage. De plus, un nouvel algorithme d'allocation binaire qui considère l'excitation auditive produite par le bruit de quantification est présenté. Cet algorithme d'allocation adaptif peut être combiné avec tout modèle de masquage basé sur la répartition de l'excitation auditive.

Acknowledgments

Firstly, I would like to acknowledge my supervisor, Prof. Peter Kabal, for his indispensable guidance and valuable support throughout my graduate studies at McGill University. I am grateful to Prof. Kabal and les Fonds de Recherche sur la Nature et les Technologies du Quebec (previously known as FCAR) for providing financial assistance to complete this research. I would also like to recognize Dr. Hossein Najafzadeh-Azghandi for his technical expertise in the field of perceptual audio coding and the numerous discussions that have influenced the outcome of this work.

I would like to thank my fellow Telecommunications and Signal Processing Laboratory graduate students for their close friendship and support; they are Mr. Tarun Agarwal, Mr. Mark Klein, Mr. Wesley Perreira, Mr. Paxton Smith, Mr. Aziz Shallwani and others. They have all contributed to this work by providing technical advice, volunteering for subjective testing or simply creating a pleasant work environment.

I would like to express my deepest gratitude to Miss Nadine Ishak for her wonderful love and support. Her patience and understanding during my graduate studies were greatly appreciated. Nadine has been my closest friend throughout my years at McGill University.

Finally, I am sincerely indebted to my family for their everlasting encouragement, their confidence and the opportunities that they have given to me. They are the reason for who I am and what I have accomplished thus far. To my sister, Miss Michèle Cave, my mother, Mrs. Suzanne Ledoux and my father, Mr. Ronald Cave, thank you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Audio coding is used in a variety of applications such as personal communication systems, internet multimedia and digital broadcast. The digital representation of information bearing signals allows for reliable and efficient transmission or storage. Audio compression algorithms are concerned with the digital representation of sound using a reduced number of information bits. The prevalence of audio coders has increased along with the advent of next generation communication systems, for which rising traffic volumes substantiate the need for bandwidth efficient representation.

Transparent coding is achieved when the original and coded signals are perceptually equivalent to a human listener. The Compact Disc (CD) representation, operating in stereo at 1.41 Mb/s, is a benchmark for transparent quality. Although essentially perceptually flawless, CD representation results in an excessively high data rate for transmission or storage of sound. Audio coding algorithms aim at reducing information bit requirements, while maintaining an acceptable quality for the reconstructed signal.

Traditionally, reductions in data rate have been achieved by exploiting the redundancy that is inherent in audio signals. More recently, advanced audio coding algorithms have been proposed that consider human perception in order to further reduce data rates while maintaining high fidelity. The quality of reconstruction from a coded signal is ultimately established from the perception of human listeners. Perception involves the recognition and interpretation of sensory stimuli. Audio signals stimulate the human auditory system, leading to their perception. Distortion may be introduced when representing sound at a reduced data rate. Perceptual audio coders shape the distortion in frequency such that it

is imperceptible by the human ear. Alternatively, perceptual coding can be described as the representation of information that contributes to the perception of a sound; only signal components that affect perception are represented. The exploitation of perceptual effects in the design of audio coders has led to high compression ratios while maintaining audible distortion at a minimum level. *Auditory masking*, which is introduced in the following section, is the primary perceptual effect that is considered in audio coding.

1.1 Auditory Masking

Auditory masking is the process by which a stronger audio signal inhibits the perception of a weaker signal. The intensity of the weaker audio signal must be raised so as to be heard by a human listener. The *masking threshold* (or masked threshold) corresponds to the increased threshold of audibility, resulting from the presence of the stronger masker signal.

A variety of masking instances occur in everyday life. For example, the music of a car radio can mask the sound of the engine of the car, provided that the music is considerably louder [1]. Similarly, a speaker must raise his/her voice when background noise increases in order to be heard. Masking is a phenomenon in sensory perception that has received significant attention from researchers in the field of psychoacoustics. Psychoacoustic experiments have been performed by several researchers in order to discover and model auditory masking effects. The amount of masking is influenced by various factors including signal level, frequency and duration. The phenomenon of auditory masking is explained in more detail in Chapter 2.

1.2 Perceptual Audio Coding

A number of paradigms have been proposed for the digital compression of audio signals. Accordingly, audio coders are commonly categorized as either *parametric coders* or *waveform coders*. The concept of perceptual audio coding is relevant in the latter case, where auditory perception characteristics are applicable.

Parametric coders represent the source of the signal rather than the waveform of the signal. Such coders are suitable for speech signals since accurate speech production models are available. More specifically, the vocal tract is modelled as a time-varying filter that is

excited by a train of periodic impulses (voiced speech) or a noise source (unvoiced speech). The parameters that characterize the filter are encoded and then used by the decoder to synthesize speech segments. More advanced parametric coders also include the error signal resulting from the reconstruction using the extracted speech parameters. The error signal generally represents the excitation to the vocal tract filter, as implemented in Code-Excited Linear Predictive (CELP) coders.

On the other hand, waveform coders attempt to accurately replicate the waveform of the original signal. Such coders provide a more perceptually agreeable reconstruction of general audio signals than parametric coders. Efficient waveform coders remove redundancy within the coded signal by exploiting the correlation between signal components, either in time or transform domain. Perceptual waveform coders additionally remove information that is irrelevant to the perception of the signal. The block diagram of a generic perceptual audio coder is illustrated in Figure 1.1. The encoding of the input signal is performed in the upper branch of the diagram, whereas the lower branch determines the bit assignment per signal component.

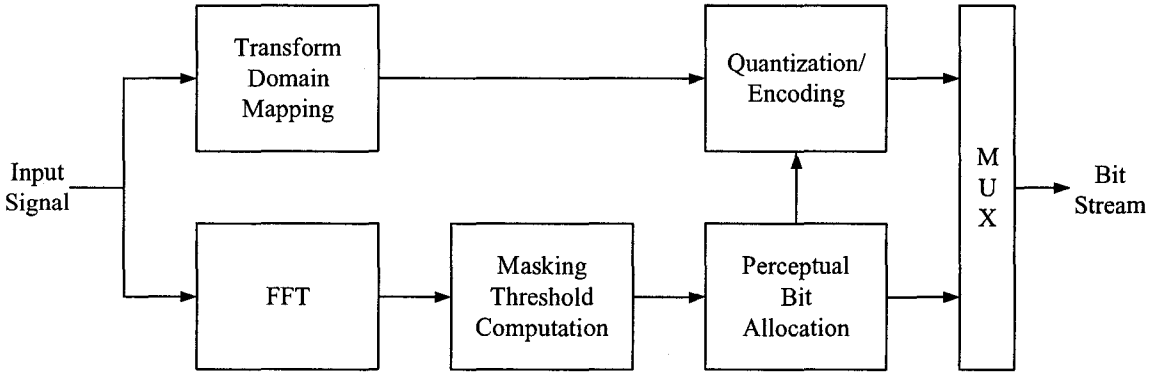


Fig. 1.1 Basic structure of a perceptual audio coder.

1.2.1 Transform Domain Mapping

A transformation is applied so as to obtain the spectral representation of the input signal. The transformation typically corresponds to a unitary transform or a bank of critically sampled bandpass filters. Several advantages result from encoding the input signal in a transform domain [2]. Firstly, effective transforms compact the information of the signal

into fewer coefficients, ensuing in a more efficient usage of quantizers (Section 1.2.3). Transform coefficients are less correlated than temporal samples of the input signal. Secondly, the desired frequency resolution is achievable through judicious selection of the transformation. Auditory masking effects are significantly influenced by the frequency composition of the input signal. As such, transform domain coding is ideal for the application of auditory perception characteristics.

The transformation is applied to temporal frames of the input signal during intervals for which the signal is considered stationary. Audio coders typically segment the input signal into frames ranging from 2 ms to 50 ms, depending on the desired temporal and frequency resolution [3].

1.2.2 Masking Threshold Computation

A masking threshold is computed based on the frequency representation of the signal. More specifically, the Discrete Fourier Transform (DFT) coefficients are used to evaluate the masking threshold. Audio signals have complex spectra, composed of multiple masking components. Masking components are extracted from the spectrum of the input signal and individual masking effects are combined to yield an overall masking threshold. Auditory models deliver a masking threshold along with the amount of allowable distortion in the frequency domain. Classical masking applications assume that signal energy lying below the masking threshold is inaudible. As many as 50% of transform coefficients are masked in transform coding of music and speech signals [2]. A frequently cited example of masking is the 13 dB miracle. Noise added to an audio signal, having a spectral structure that is adapted to that of the signal, is inaudible for signal-to-noise ratios as low as 13 dB [4]. A common output of the masking threshold computation stage is the *Signal-to-Mask Ratio* (SMR), which represents the ratio of the signal input to the amount of masking produced by the signal.

1.2.3 Quantization

Quantization is defined as the process of transforming the sample amplitude of a message signal into a discrete amplitude taken from a finite set of possible amplitudes [5]. In digital audio coding, an already quantized signal is further quantized by representing its samples using a smaller set of amplitudes. More specifically, signal components are represented

using fewer bits. The quantization process introduces an error in the representation of the signal. The resulting error is defined as the difference between the quantized signal and the original signal. The distribution of the quantization error depends on both the quantizer design and the distribution of the input signal. As such, the quantization process can be modelled as the addition of an error to the input signal, as depicted in Figure 1.2, where $e_q(n)$ represents the quantization error. In audio coding, distinct quantizers are employed to represent the different components or sub-bands of the input signal, depending on the applied transformation.

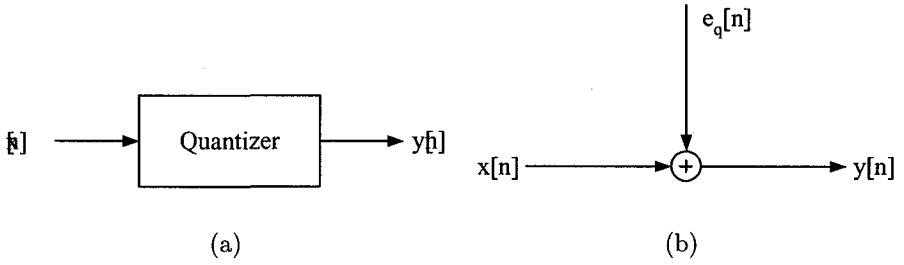


Fig. 1.2 Illustration of the quantization process.

1.2.4 Perceptual Bit Allocation

The allocation of information bits to the different quantizers is performed adaptively, based on the computed masking threshold. Firstly, spectral components lying beneath the masking threshold need not be represented. Such components do not contribute to the perception of the audio signal according to classical masking principles. Secondly, the noise that is introduced by the quantization process is shaped in frequency such that it becomes inaudible. For instance, more noise is allowed where the masking threshold is high, resulting in the allocation of fewer information bits to those regions. The process is referred to as *spectral noise shaping.*

1.3 Challenges in Low-Rate Audio Coding

The incorporation of perceptual effects has been extensively applied in medium to high rate audio coding. Standardized coders, such as MPEG-1 [6], MPEG-2 AAC [7] and Dolby

AC3 [8], arguably achieve CD quality at rates of 96 kb/s to 256 kb/s for wideband audio signals. Within this range, the noise that is introduced is shaped such that it lies considerably below the masking threshold. As a result, auditory models that are conservative in predicting the amount of masking are employed. The amount of masking is commonly underestimated so as to ensure that distortion remains inaudible.

More recently, it has been shown possible to reduce data rates to less than 10 kb/s for narrowband audio signals, while maintaining acceptable quality [2]. In the vicinity of 1 bit per sample, distortion is considerably higher, which entails the need for very accurate masking models. The level of introduced quantization noise is generally comparable to the masking threshold. The current work focuses on the application of auditory masking in low-rate coding of narrowband audio signals.

1.4 Other Applications of Auditory Masking

While audio coding is the most widespread, other applications of human auditory perception have been proposed in speech and audio processing. For instance, certain signal enhancement techniques and perceptual quality evaluation models consider auditory masking effects. A brief overview of these applications is provided in this section.

1.4.1 Audio Signal Enhancement

Several signal enhancement techniques have been proposed to improve the perceived quality of audio signals. These methods are commonly used in speech communications to remove background noise from a transmitted signal. However, the reduction of noise often introduces irritating artifacts due to the inaccuracy of noise estimates.

In conventional spectral subtraction, the estimated noise is removed from the short-term spectrum of the input signal. Over-subtraction inserts distortion in the reconstructed signal, causing such artifacts. Auditory masking has been introduced in order to restrict the amount of noise that is attenuated [9, 10, 11]. A masking threshold is computed from the spectrum of the input signal. Rather than subtracting all of the noise, only the part that is above the masking threshold is removed. Noise below the masking threshold is considered inaudible. This approach has been found to reduce the number of introduced artifacts since fewer modifications are performed to the signal. Similarly to spectral subtraction, auditory masking effects have been applied to signal subspace methods of speech enhancement [12].

1.4.2 Perceptual Quality Evaluation of Audio Signals

The performance of audio coding schemes is evaluated using objective and/or subjective quality measures that compare the coded signal with the input reference signal. Typical objective measures, such as *Signal-to-Noise Ratio* (SNR) or *Mean-Square Error* (MSE), do not accurately represent the perceived quality of the reconstructed signal. This inconsistency increases when considering low-rate coders that incorporate auditory masking effects. Rather than performing time-consuming and expensive subjective listening experiments, a model of human auditory perception is used to evaluate the perceptual difference between the reference signal and the signal under test.

The block diagram of a basic model for the perceptual quality evaluation of audio signals is illustrated in Figure 1.3. A masking threshold is computed from the frequency representation of the reference signal. The error signal is evaluated as the difference between the frequency representations of the reference signal and the signal under test. An audio quality measure is determined from the comparison between the error signal and the masking threshold. The described model evaluates audio quality based exclusively on auditory masking. More sophisticated models incorporate complex perceptive and cognitive representations of the human auditory system that consider additional psychoacoustic metrics.

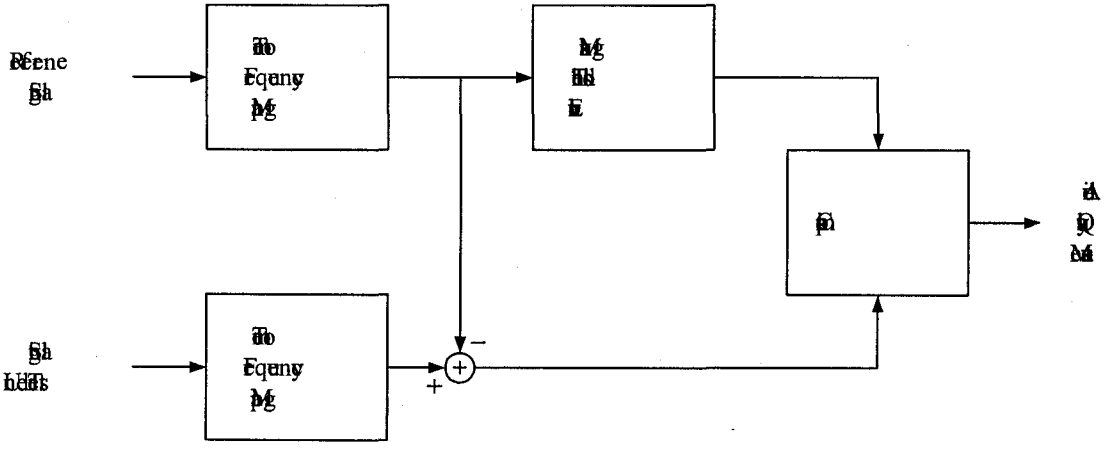


Fig. 1.3 Generic model for the perceptual evaluation of audio quality using auditory masking [13].

1.5 Thesis Contribution

Extensive research has been performed by audio coding specialists to incorporate human perception within medium to high rate coders. At low coding rates, some distortion is inevitable, which entails the need for a more accurate representation of perceptual effects. The current work is primarily concerned with research on auditory masking and its application to low-rate audio coding, with the aim of improving audio quality for bandwidth limited applications such as wireless communications and internet multimedia.

An assembly of psychoacoustic experimental results and auditory masking models are first presented and analyzed. Resulting from this study, a new algorithm is proposed for the prediction of auditory masking. Among other features, transient masking effects are modeled by tracking the temporal evolution of masking components. Moreover, it is suggested that psychoacoustic effects produced by quantization noise should be considered along with the masking threshold in the perceptual bit allocation algorithm. An adaptive bit allocation scheme is presented that considers the auditory excitation produced by the quantization noise. Quintessentially, this thesis proposes: (1) an innovative model for the prediction of auditory masking and (2) a novel perceptual bit allocation algorithm.

1.6 Thesis Synopsis

This thesis is structured into 6 chapters. Chapter 2 introduces concepts related to sound levels, the human auditory system and psychoacoustic processing that form a foundation for the proposed work. A collection of experimental results related to auditory masking and psychoacoustic processing are presented, along with associated models for their application.

Chapter 3 presents auditory models that have been developed for the prediction of masking thresholds in speech and audio processing. A thorough review of these models is provided along with the identification of their shortcomings. A novel auditory masking model is proposed that incorporates many of the psychoacoustic findings that are presented in Chapter 2.

Chapter 4 provides a review of adaptive bit allocation strategies for audio coding. A novel bit allocation scheme is presented that is based on a new criteria for the allocation of bits. The proposed allocation scheme collaborates with the masking threshold computation in the modelling of auditory effects.

Chapter 5 is dedicated to the evaluation of the proposed bit allocation scheme and the proposed auditory masking model. They are compared to previous work based on informal subjective listening experiments.

Finally, a complete summary of the proposed work is provided in Chapter 6, along with directions for future related research.

Chapter 2

Psychoacoustic Principles

The notion of auditory masking was introduced in the previous chapter in conjunction with an overview of its engineering applications. This chapter commences with the presentation of fundamental concepts related to sound levels and the human hearing system. Following, auditory masking is further described along with a presentation of a collection of related psychoacoustic results from various researchers. These psychoacoustic experiments examine the relation between sound stimuli and hearing perception. This relation includes the effects of both the physiology of the ear and the cognitive processing of auditory stimuli. Psychoacoustic models are then developed from the collected experimental data in order to predict effects such as auditory masking.

2.1 Temporal and Spectral Properties of Sound

Sound is generated through the mechanical vibration of objects. The vibrating motion travels through physical media, causing acoustic waves. In most cases, the physical medium corresponds to air while the sound waves represent the variations of atmospheric pressure. For example, the movement of the cone or dome of a speaker causes vibrations in the air.

The magnitude of sound is represented as a time-varying pressure, expressed in units of Pascal (Pa). Audible sound pressures can vary from 10^{-5} Pa (absolute threshold of hearing in the middle audible range) to 10^{2} Pa (threshold of pain) [14]. Given the extent of this range, sounds are more commonly characterized by their logarithmic level or *Sound Pressure Level* (SPL). The SPL expresses the pressure relative to some reference value on

a decibel scale,

$$L = 20\log_{10}(p/p_0) \quad \text{dB}, \tag{2.1}$$

where the reference pressure p_0 has a value of 10 μPa. An additional measure of sound magnitude is the sound intensity, which represents the sound energy transmitted per second through a unit area of a sound field [1]. Since sound intensity is proportional to the square of pressure, the level can also be expressed as a ratio of sound intensity levels,

$$L = 10\log_{10}(I/I_0) \quad \text{dB}, \tag{2.2}$$

where the reference intensity I_0 has a value of 10^{-12} W/m^2. Table 2.1 lists examples of levels (in dB SPL) along with their corresponding pressure and intensity ratios for typical sounds.

Table 2.1 Examples of sound pressure level, pressure and intensity ratio of typical sounds [1].

Sound Level dB SPL	Intensity Ratio I/I_0	Pressure Ratio P/P_0	Typical Example
120	10^{12}	10^6	Loud rock concert
100	10^{10}	10^5	Shouting at close range
70	10^7	3160	Normal conversation
50	10^5	316	Quiet conversation
30	10^3	31.6	Soft whisper
20	10^2	10	Country area at night
6.5	4.5	2.1	Absolute threshold at 1 KHz
0	1	1	Reference level

It is generally more convenient to evaluate the level of a sound from its frequency domain representation. For discrete spectra (*e.g.*, periodic signals), the overall level is calculated by summing the levels of individual spectral components. Individual component levels are directly related to the squared magnitude of the Fourier series coefficients of the signal. As for continuous spectra, the overall level is obtained by integrating the sound intensity density. The sound intensity density represents the sound intensity per Hertz. The density is calculated from the squared magnitude of the Fourier transform of the signal.

2.2 The Human Auditory System

Two distinct processing stages are recognized within the human auditory system [14]. In the first stage, commonly known as the ear, sound pressure waves are converted to mechanical vibrations. The ear converts processed mechanical oscillations to electrical impulses that are delivered to the auditory nerve. The general structure of the ear is displayed in Figure 2.1, including the outer ear, the middle ear and the inner ear. In the second stage, auditory nerve impulses are processed by the brain, resulting in auditory sensations. Interestingly, this auditory structure is shared amongst most animals.

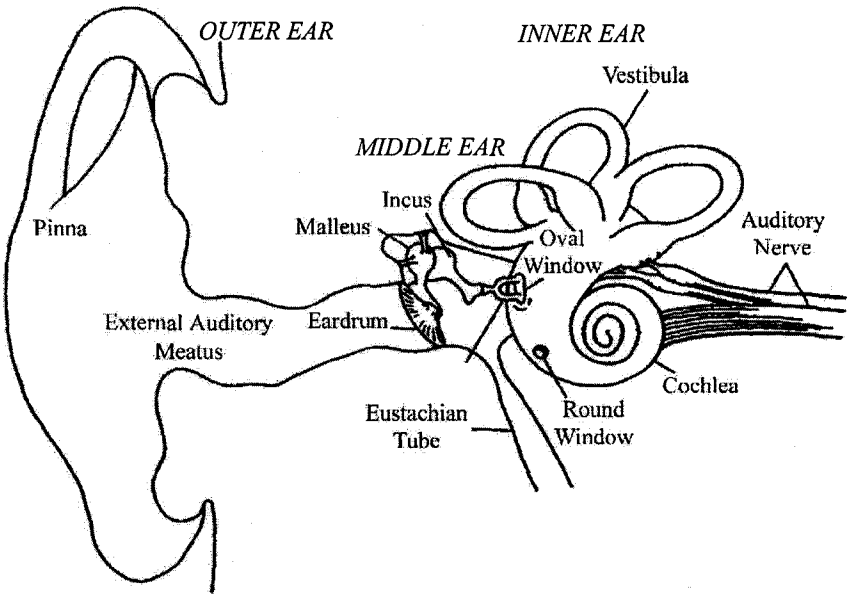


Fig. 2.1 Structure of the human ear, adapted from [15].

2.2.1 The Outer Ear

The outer ear is composed of the pinna and the auditory canal (also known as meatus). The pinna is the visible part of the ear that directs sound pressure waves towards the auditory canal. The pinna influences higher frequency sounds, contributing to the ability of localizing sounds [14]. Sound waves travel through the air-filled auditory canal from the

pinna all the way to the tympanic membrane. Open at one end and closed at the other, the meatus acts as a quarter-wavelength resonator, amplifying signals within the range of 3–5 kHz by as much as 15 dB [16]. As such, the outer ear principally accounts for the high sensitivity of the auditory system within this frequency range.

2.2.2 The Middle Ear

The middle ear originates at the tympanic membrane (also known as the eardrum) where sound pressure waves are converted to mechanical vibrations. The eardrum is connected to three small ossicular bones that lie within the air-filled cavity of the middle ear, as shown in Figure 2.2. Mechanical oscillations travel through the malleus, the incus and the stapes that connect to the inner ear via the oval window. These three bones are also commonly referred to as hammer, anvil and stirrup, and are noteworthy for being the smallest bones in the body [1].

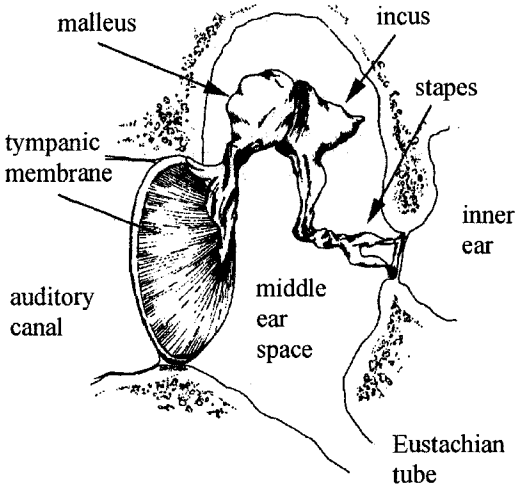


Fig. 2.2 Structure of the middle ear, adapted from [17].

The middle ear primarily acts as an impedance-matching stage between the air medium of the outer ear and the fluid of the inner ear. It accounts for the acoustical impedance mismatch between the eardrum and the oval window, reducing the amount of wave reflection. The closest impedance match occurs in the 1 kHz range, at which point sound pressure is

increased by as much as 20 dB [18]. Above this range, the middle ear resembles a low-pass filter having an attenuation of −15 dB/oct [15].

Finally, the Eustanchian tube equalizes the air pressure between the middle ear and the surrounding environment. Pressure differences between the auditory canal and the middle ear cavity greatly hinder the ability to vibrate of the tympanic membrane. Mismatches are commonly encountered in situations such as flying or diving. Normal hearing is resumed by swallowing as the upper throat end of the Eustanchian tube is opened, enabling pressure equalization [14].

2.2.3 The Inner Ear

The inner ear has the most significant role in perception within the auditory system. It includes the cochlea, from which mechanical vibrations emanating from the oval window are transformed into electrical impulses. The structure of the inner ear is detailed in Figure 2.3.

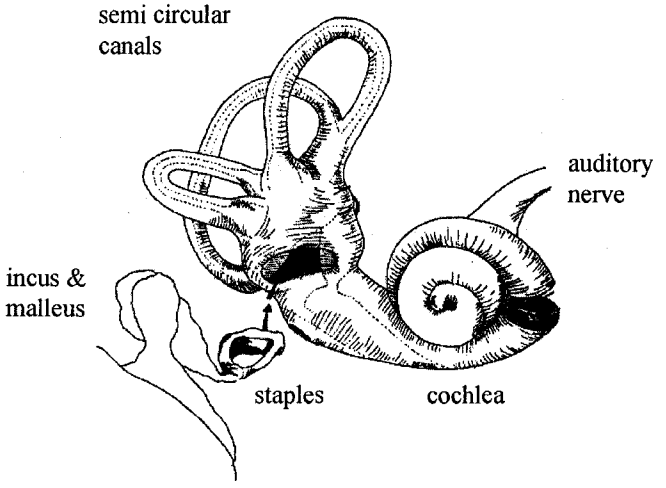


Fig. 2.3 Structure of the inner ear, adapted from [17].

The cochlea is a spiral-shaped tube of 35 mm length, coiled into approximately 2.5 turns [15]. Within its hard bony walls are two nearly incompressible lymphatic liquids. The region close to the oval window is recognized as the base, whereas the inner tip of the coil is known as the apex. Both Reissner's membrane and the basilar membrane partition

the cochlea along its length into three channels. A cross-sectional view of the cochlea is illustrated in Figure 2.4, showing the scala vestibuli, the scala media and the scala tympani. The helicotrema, a small opening near the apex, allows for pressure equalization between the scala vestibuli and the scala tympani. Pressure in the scala tympani is reduced through the round window of the basilar membrane, located near the base. Oval window vibrations are transmitted to the cochlear membranes through the incompressible fluids, setting them in motion.

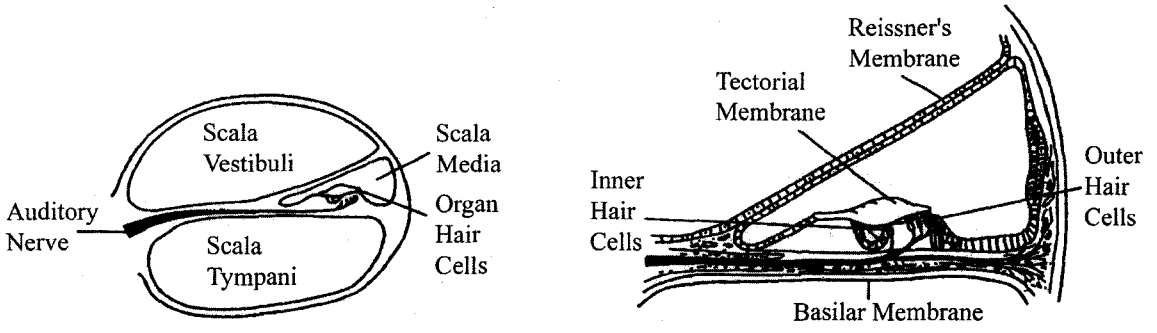


Fig. 2.4 Cross section of the cochlea, adapted from [15].

2.2.4 The Basilar Membrane

The basilar membrane extends throughout the cochlea, from the basal end to the apex. Profile and rigidity vary along its length and significantly influence the response along the basilar membrane to different frequencies. It is rigid and thin near the oval window (0.04 mm) while the apex is limp and vast (0.5 mm) [15]. Each point along the basilar membrane is associated with a *Characteristic Frequency* (CF) for which the amplitude of its vibrations is maximal. More specifically, travelling waves reach a maximum amplitude at the location along the basilar membrane where the characteristic frequency is equal to the frequency of the wave. Higher characteristic frequencies correspond to the base, whereas lower characteristic frequencies are near the apex.

2.2.5 Sensory Hair Cells

The basilar membrane supports the organ of Corti, where 30 000 sensory hair cells attach to the auditory nerve [15]. The sensory cells are arranged in one row of inner hair cells on the inner side of the organ of Corti, and three rows of outer hair cells near the middle of the organ of Corti [14]. The motion of the basilar membrane causes the bending of hair cells, leading to neural firings in the auditory nerve. Neural information propagates to the brain where it undergoes cognitive processing.

2.3 Absolute Threshold of Hearing

The absolute threshold of hearing (or audibility threshold) indicates the minimum sound pressure level that a sound must have for detection in the absence of other sounds. The threshold in quiet is easily measured through hearing experiments. A mean threshold is obtained by averaging the individual thresholds of numerous listeners. The audibility threshold shows a prominent dependence on frequency, as illustrated in Figure 2.5. Terhardt proposed an expression for the frequency dependent threshold [19] based on experimental data from an earlier study [20],

$$T_q(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5\exp\left[-0.6\left(\frac{f}{1000} - 3.3\right)^2\right] + 10^{-3}\left(\frac{f}{1000}\right)^4 \quad \text{dB}, \quad (2.3)$$

where f is expressed in Hz. The auditory threshold represents the combined effects of the outer and middle ear frequency responses along with the internal noise of the inner ear [10].

2.4 Critical Bands and Auditory Filters

As previously mentioned, a frequency-to-place conversion occurs within the inner ear that affects the frequency selectivity of the hearing system. Frequency selectivity is crucial to perception as it determines the ability of the auditory system to resolve frequency components. The concept of critical bands is introduced to define a frequency range within which changes in stimuli greatly affect perception. It is suggested that the ear integrates sound energy within a critical band. When two sounds have energy in the same critical band, the

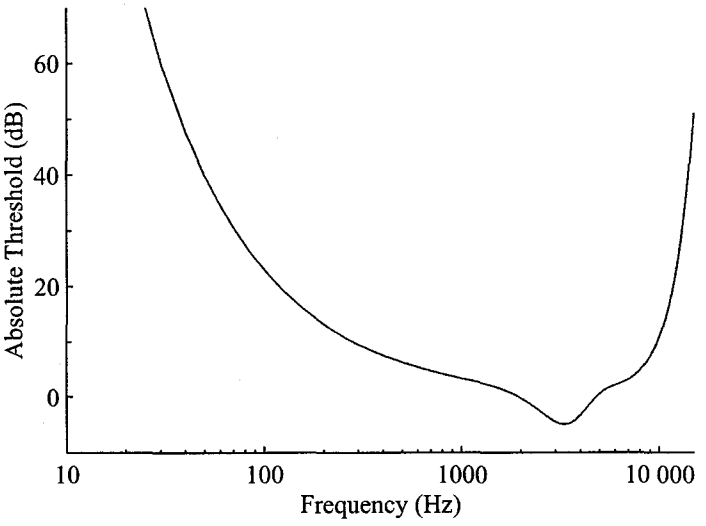


Fig. 2.5 Absolute threshold of hearing for normal listeners.

sound having the highest level dominates perception [15].

Fletcher first presented the concept of critical bands in 1940 [21]. He measured the audibility threshold of a sinusoid in the presence of narrowband noise, centred at the same frequency of the sinusoid. While maintaining its power density constant, the noise bandwidth was gradually increased. It was observed that the audibility threshold increased with the noise bandwidth up to a certain point, where further increases in bandwidth had minor effect on the threshold. Fletcher called this value the critical bandwidth. Following his experiments, he suggested that the peripheral auditory system behaves alike a bank of bandpass filters, where the bandwidth of each filter corresponds to the critical bandwidth. These filters are now commonly referred to as auditory filters. As discussed in Section 2.2.4, each point along the basilar membrane is associated with a characteristic frequency. Hence, the response at a given point along the basilar membrane corresponds to the output of the the auditory filter centred at its characteristic frequency.

Scharf measured the bandwidth of critical bands as a function of their centre frequency [22]. While attempting to represent the inner ear as a discrete set of non-overlapping auditory filters, he determined that 25 critical bands were sufficient to represent the audible frequency range of the ear. The bandwidth of the resulting critical bands are listed in Table 2.2, with centre frequencies spanning from 0 to 19 kHz. It is evident that auditory filter bandwidths are larger at lower centre frequencies than at higher centre frequencies.

According to Scharf's results, critical bands are constant below 500 Hz, while they steadily increase above 500 Hz. Additionally, the majority of critical bands lie below 5 kHz. The ability of the ear to resolve components is superior at low frequencies than at high frequencies. It has been suggested that each band corresponds to approximately 1.5 mm of spacing along the basilar membrane in such a discrete critical band structure [15].

Table 2.2 List of critical bands measured by Scharf [22].

Critical Band Number	Lower Frequency Hz	Centre Frequency Hz	Upper Frequency Hz	Bandwidth Hz
1	0	50	100	100
2	100	150	200	100
3	200	250	300	100
4	300	350	400	100
5	400	450	510	110
6	510	570	630	120
7	630	700	770	140
8	770	840	920	150
9	920	1000	1080	160
10	1080	1170	1270	190
11	1270	1370	1480	210
12	1480	1600	1720	240
13	1720	1850	2000	280
14	2000	2150	2320	320
15	2320	2500	2700	380
16	2700	2900	3150	450
17	3150	3400	3700	550
18	3700	4000	4400	700
19	4400	4800	5300	900
20	5300	5800	6400	1100
21	6400	7000	7700	1300
22	7700	8500	9500	1800
23	9500	10500	12000	2500
24	12000	13500	15500	3500
25	15500	19500		

Given the importance of the critical band concept, a perceptual scale has been based upon it. The critical-band rate is obtained by adding one critical band to the next in such a

way that the upper limit of the lower band corresponds to the lower limit of the next higher critical band [14]. Accordingly, there is a one-to-one mapping between frequency and the number of critical bands. The critical-band rate is expressed in units of Bark, where an increment of one Bark corresponds to one critical band. Zwicker suggested an analytical expression that characterizes the dependence of critical-band rate on frequency [14],

$$Z = 13 \arctan(760f) + 3.5 \arctan(f/7500)^2, \tag{2.4}$$

where f and Z are expressed respectively in Hz and Bark. The bandwidth of each critical band as a function of its centre frequency is approximately given by:

$$BW(f) = 25 + 75\left(1 + 1400\, f^2\right)^{0.69}, \tag{2.5}$$

where $BW(f)$ is expressed in Hz. Similarly, Schroeder proposed a relationship between frequency and critical-band rate that is mostly linear below 500 Hz and exponential above 1 kHz [23],

$$Z = 7 \operatorname{arcsinh}(f/650), \tag{2.6}$$

where f is expressed in Hz and Z in Bark. Schroeder suggested that this equation accurately matched experimentally measured critical bands for frequencies up to 5 kHz.

An alternative measure for the perceptual frequency of the ear was proposed by Moore and Glasberg [24]. Their novel scale is based on the *Equivalent Rectangular Bandwidth* (ERB) of the auditory filters of the inner ear. The ERB of a filter corresponds to the bandwidth of the rectangular filter which has the same peak transmission and passes the same power given a white noise input [1]. Moore and Glasberg proposed an equation relating the ERB to the centre frequency of an auditory filter,

$$\text{ERB} = 24.7(4370\, f + 1), \tag{2.7}$$

where f is expressed in Hz. Each ERB corresponds approximately to a 0.89 mm section along the basilar membrane [1]. When comparing ERB to the traditional critical bandwidths listed in Table 2.2, certain discrepancies are found at lower frequencies. Moore and Glasberg argued that the bandwidth of auditory filters steadily decreased below 500 Hz, whereas previous critical bandwidth measurements were relatively constant within the same

range. The proposed perceptual scale corresponds to the number of ERB's,

$$\text{Number of ERB's} = 21.4 \log_{10}(4370\, f + 1). \tag{2.8}$$

The three perceptual frequency relations mentioned above are illustrated in Figure 2.6. Zwicker's equation is represented by the solid line, Schroeder's equation is represented by the dotted line and the ERB scale is represented by the dashed line. It is apparent from the figure that ERB is smaller than the critical bandwidth for auditory filters having low centre frequencies. As a result, the number of ERBs grows more rapidly at low frequencies than the critical-band rate. Both Zwicker's equation and Schroeder's equation are very similar below 5 kHz, whereas the critical band rate grows faster for the latter at high frequencies. Schroeder's critical band relation yields greater perceptual resolution for wideband signals.

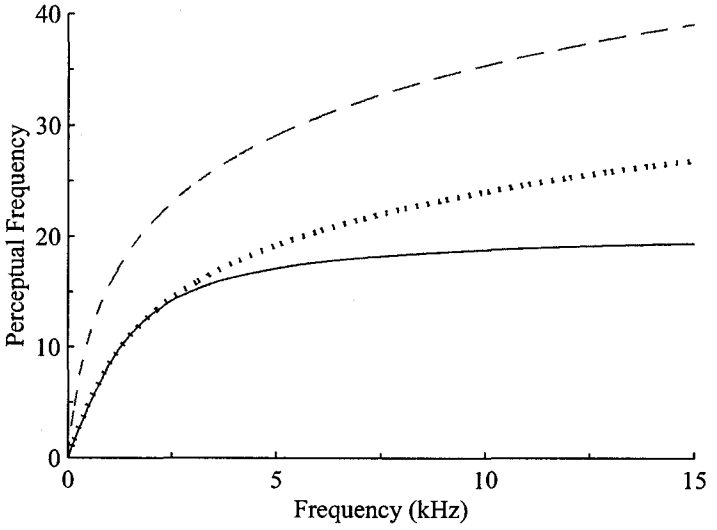


Fig. 2.6 Mapping between perceptual frequency and linear frequency. Solid line: Zwicker's equation in Barks. Dotted line: Schroeder's equation in Barks. Dashed line: Number of Equivalent Rectangular Bandwidths (ERB).

2.5 Auditory Masking

Essential mechanisms involved in the processing of auditory stimuli were presented in the previous sections. The current section focuses on the psychoacoustic phenomenon of mask-

ing. As discussed in Chapter 1, auditory masking is the process by which the perception of one sound is suppressed by another. Masking is characterized by an increase in the audibility threshold of a signal in the presence of a stronger signal. The amount of masking corresponds to the quantity by which the threshold is augmented above the threshold in quiet. The stronger masking signal is commonly referred to as the *masker*, whereas the signal being masked is identified as the *maskee*, *target* or *probe* signal.

Masking effects are generally categorized as one of two types: *simultaneous* or *temporal* masking. Simultaneous masking occurs when the masker and maskee are presented to the ear concurrently. Temporal masking, also termed non-simultaneous masking, occurs when the masker and target have a temporal offset with respect to each other. Accordingly, the target may be masked when presented prior to the masker onset or following its offset. The former scenario is known as backward masking while the latter is recognized as forward masking. The different masking types are illustrated in Figure 2.7, where the solid and dotted lines represent the masker and masking threshold respectively. Backward masking is observed prior to the masker onset. The masker is present from 0 ms to 200 ms, corresponding to simultaneous masking. The masker is removed at 200 ms, beyond which point masking continues as indicated by the decaying dashed line in the figure. Simultaneous and temporal masking, as illustrated in Figure 2.7, have been observed for a variety of masker and target types.

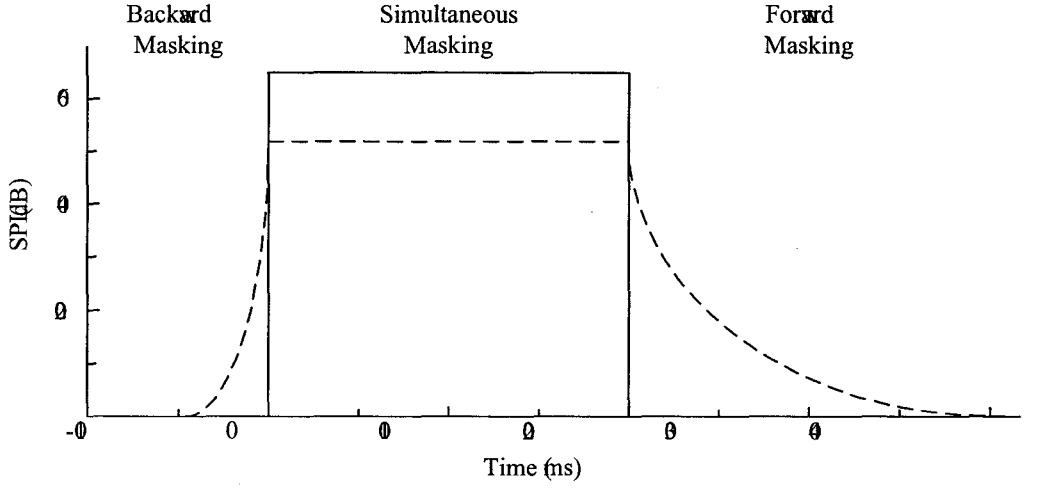


Fig. 2.7 Various types of masking obtained from a 200 ms masker burst, adapted from [3]. Solid line: masker signal. Dashed line: masking threshold.

A method is required to characterize masking with respect to frequency and time. A *masking pattern*, also known as *masked audiogram*, is a graph of the amount of masking produced by a given sound as a function of the centre frequency of the target sound or its temporal location [1]. The amount of masking produced by a signal may noticeably differ from one listener to another. As such, masking patterns are commonly obtained by averaging masking thresholds observed by numerous listeners. A more detailed discussion concerning masking patterns and their characteristics is presented in Section 2.7.

2.5.1 Simultaneous Masking

Simultaneous masking is the most significant since it produces the largest amount of masking. According to Moore, there are two distinct hypotheses that describe the origins of masking [1]. The first concept involves the swamping of neural activity by the stronger masker signal. Consider an example where a Bark-wide noise masker is presented at a sound pressure level of 40 dB. By adding a tone of 20 dB SPL within the same critical band, the increase in sound pressure level within the auditory filter is only 0.04 dB. It has been suggested that the ear cannot discriminate such small differences in sound pressure level. Consequently, the tone remains undetected in the presence of the more intense noise signal. This mechanism is in accordance with the concept that the ear integrates sound energy within each auditory filter.

The second mechanism suggested as a foundation for masking is suppression. In this case, the masker suppresses neural activity caused by the target signal. Neural activity is contained to the random and spontaneous hair cell firings that normally occur during silence periods. This results in a highly non-linear process that is difficult to predict. Although different in nature, Moore suggests that both mechanisms could contribute to simultaneous masking effects.

2.5.2 Temporal Masking

Temporal masking effects also provide a significant contribution to masking thresholds. Backward or pre-masking occurs when the target precedes the masker. Thiede proposed two distinct mechanisms that explain the incidence of backward masking [25]. Firstly, he suggested that intense signals are processed more rapidly than weak signals. The maskee can be overtaken by the masker during the processing of the signal, either within the

inner ear or at a cognitive level. Secondly, he suggested that backward masking results from the reduced temporal resolution of the ear. Average levels of maskee and masker determine masking thresholds rather than instantaneous levels. Backward masking has received considerably less attention in psychoacoustic research compared to other types of masking. It has been shown that backward masking only begins 20 ms prior to the masker onset [14]. Some experiments have also shown that a short tone ending 1 ms before the beginning of a noise burst can experience up to 60 dB of masking [15].

On the other hand, forward masking is significantly more effective in suppressing the perception of target signals. Forward, or post-masking, is observed when the target signal follows the masker. Its effects are observed up to 200 ms following the masker offset [15]. Moore reported the following characteristics of forward masking, determined from a series of psychoacoustic experiments [26]:

1. Forward masking is increased when the target approaches the masker offset in time. The amount of forward masking decreases linearly as the logarithm of the delay between the masker and the target increases.

2. Despite the initial amount of forward masking, it always decays to zero after 100 to 200 ms. This implies that the slope of forward masking decay is steeper for higher masker levels.

3. Increments in masker level do not result in equivalent increments in the amount of forward masking.

4. The amount of forward masking increases as the duration of masker increases.

In a later publication, Moore suggested that the following three factors contribute to forward masking [1]:

1. The basilar membrane vibrations continue for a certain amount of time after the masker offset. This effect, known as ringing, contributes to the masking of the target signal when temporally overlapped.

2. Fatigue in the auditory nerve or the time required for its adaptation following the masker offset.

3. The neural activity produced by the masker persists at higher processing levels than the auditory nerve, following the masker offset.

2.6 Excitation Pattern Model of Masking

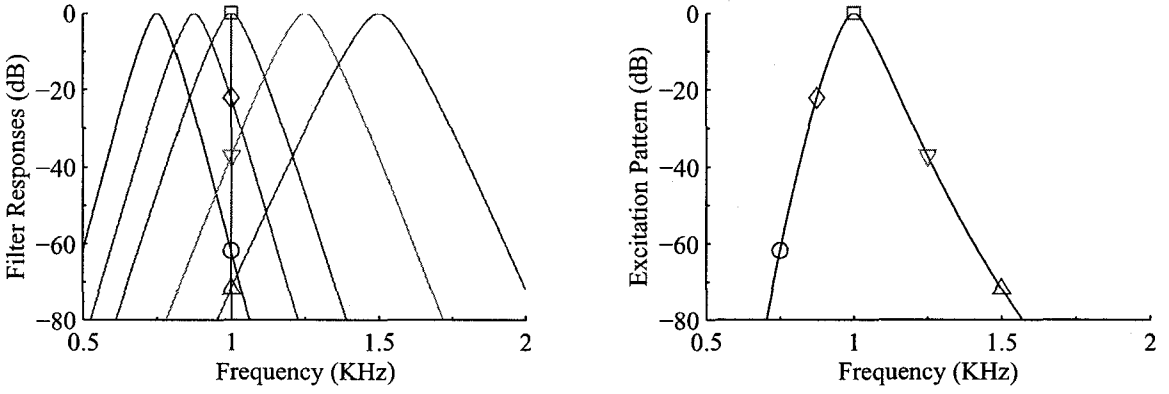
A variety of psychoacoustic studies have attempted to relate masking effects to the output of auditory filters. It has been shown that masking cannot be determined by simply examining the response of the auditory filter centred at the masker frequency. Masking patterns have different spectral characteristics above and below the masker signal. As such, it is instructive to examine the excitation that a sound produces throughout the audible spectrum. The excitation pattern of a sound is a representation of the activity or the excitation evoked by that sound as a function of characteristic frequency along the basilar membrane [27].

A method was suggested by Moore and Glasberg to predict the excitation patterns of sounds using the notion of auditory filters [28]. The excitation at a given frequency corresponds to the output of the auditory filter centred at that frequency. This method is illustrated in Figure 2.8 by deriving the excitation produced by a 1 kHz tone. Figure 2.8(a) displays the responses of five auditory filters having different centre frequencies. The roex filters described in [26] were employed in this example.[1] The input sinusoid is represented by the vertical line at 1 kHz. The response of each filter with respect to the tone is marked along the vertical line. Figure 2.8(b) displays the response of each auditory filter to the input sinusoid as a function of their centre frequency, yielding the excitation pattern.

Zwicker proposed a psychoacoustic model to predict the amount of masking produced by a sound that is based on its excitation [14]. The model exploits the just-noticeable level variations of the excitation pattern. The detection of a target signal is determined by comparing the excitation produced by the masker alone with that of the masker presented with the target signal. If the difference between the two excitation patterns is greater than a certain value at the output of any auditory filter, the target signal is detectable. The critical value represents the just-noticeable difference in level. Zwicker suggested that a value of 1 dB fit the observed experimental data.

The method described, commonly referred to as the excitation pattern model of masking, forms the basis for the models that are described later in Chapter 3. While it provides a straightforward approximation to the amount of masking produced by a signal, its accuracy remains uncertain. The model assumes that listeners use the overall level of the

[1] Roex filters are characterized by an asymmetric frequency response that forms an exponential with a rounded top. Such filters were derived using the notched-noise method for estimating auditory filter shapes [26].



(a) Frequency responses of five different auditory filters.

(b) Response of auditory filters to a 1 kHz tone as a function of their centre frequency.

Fig. 2.8 Derivation of the excitation pattern produced by a 1 kHz tone.

signal as its only detection cue. However, psychoacoustic experiments have revealed that listeners often utilize off-frequency listening or other more complex cues to improve signal detection. Masking patterns exhibit a dependence on the nature of both the masker and target signals.

2.7 Masking Patterns

Four basic combinations of masker and target types are commonly reported in literature, where each signal can be either a tone or noise band. The majority of reported psychoacoustic experiments involve masking of tonal target signals. The two combinations having noise-like targets, which are the most relevant in audio coding applications, have received less attention. This section presents simultaneous masking patterns for different masker-target combinations while noting their differences. Masking curves are typically approximated as triangular on a critical band scale, having different slopes above and beyond the centre frequency of the masker.

Veldhuis noted the lack of results concerning noise-like targets when he argued that most masking models are unsuitable for the masking of quantization noise targets in audio coding [29]. He conducted an informal experiment to measure the masking threshold produced by a tone for critical band noise targets at various frequencies. Although masking

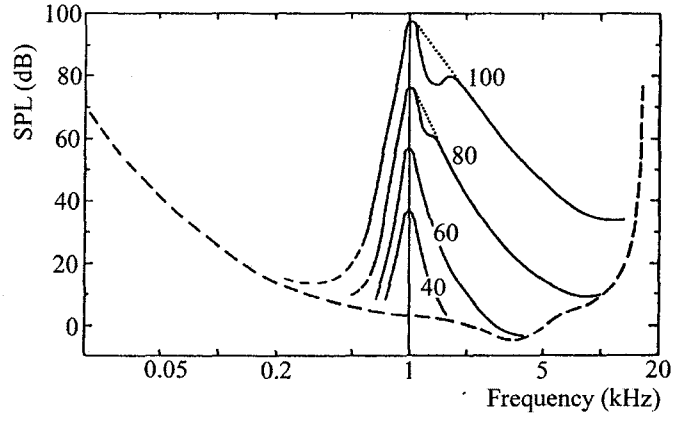
curves for the noise target had similar shapes to those obtained for tonal targets, some differences brought him to conclude that further study in masking of noise was necessary.

Similarly, previous work has shown some considerable differences in masking patterns produced by tonal and noise maskers. For instance, narrowband noise maskers are generally more effective than sinusoidal maskers for frequencies below the masker [30]. Furthermore, the slopes on the lower frequency side become less steep with decreasing masker level for tonal maskers, whereas they remain practically invariant for noise maskers [14]. On the higher frequency side, pure tones commonly produce more masking than narrow-band noise for masker levels in the neighbourhood of 80 dB SPL [31]. Traditional simultaneous masking patterns measured by Zwicker are reproduced in Figure 2.9 [14]. Figure 2.9(a) displays masking patterns of a sinusoidal target produced by a Bark-wide noise masker at various sound pressure levels. In Figure 2.9(b), the masker is a 1 kHz tone whereas the target signal is a noise band.

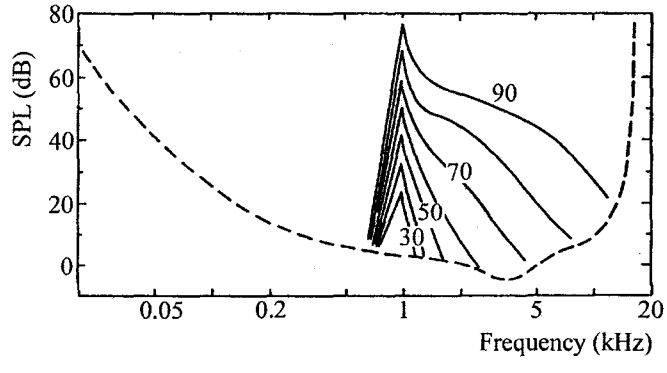
Moore *et al.* have recently investigated the four possible masker and signal type combinations in a series of five experiments [32]. The objective of their research was to show complex irregularities in masking patterns in order to disprove the idea that signal thresholds are directly related to the spread of masker excitation. They conducted a series of experiments involving different masker and signal characteristics to determine the roles of various cues in signal detection. Such detection cues are used to explain masking curves when the traditional excitation pattern models fail.

In their first experiment, they compared masking patterns for sinusoidal and noise maskers centred at 1 kHz for a variety of masker levels. Target signals were either sinusoidal or noise-like, comprising of all four masker-target combinations. All noise signals were narrowband with a bandwidth of approximately one critical band. The masking patterns obtained using the noise masker were similar for both types of target signals. Slight differences were observed for certain subjects, for whom the masking patterns for noise signal were more regular on the higher frequency side. Upper and lower frequency slopes were generally in accordance with previously reported results. Additionally, the threshold level was on average 0.7 dB higher for the noise target than for the tonal target. Thresholds surrounding the masker in frequency were only a few decibels below its level, which is consistent with the concept that the overall change in sound pressure level was used as the main detection cue.

The masking patterns observed from sinusoidal maskers were characterized by numerous



(a) Masking patterns produced by a Bark-wide noise masker with a sinusoidal target.



(b) Masking patterns produced by a 1 kHz sinusoidal masker with a noise target.

Fig. 2.9 Masking patterns obtained from maskers of various sound pressure level, centred at 1 kHz [14]. Masker levels are indicated next to the curves.

irregularities, as well as greater differences between individual listening subjects. The interaction between masker and target signals produce additional detection cues such as beats and combinations tones. These cues contribute to the irregular dips, peaks and shoulders observed throughout the masking patterns. For signal frequencies neighbouring

the masker, noise target thresholds were lower than those of the tonal signal by as much as 15 to 20 dB. While the change in level produced by adding the target was believed to be the only detection cue in the presence of a tonal target, an additional detection cue was available for the noise target. More specifically, the availability of a within-channel cue of fluctuation in level produced by the noise signal was very effective in the reduction of the audibility threshold. Moore compared his experimental results for tone-masking-noise to those previously reported by Greenwood in [33] and [34], and found them to be alike.

For frequencies above and below the masker, the two target types produced similar masking patterns when masked by the sinusoid. It is also important to note that Moore reported higher masking produced by the tonal masker than the noise masker for frequencies above 2 kHz. They summarized their first experiment by stating that the masker type mostly determined the masking characteristics above and below the masker frequency, except when the masker and target were centred. Apart from tone masking noise, the three other masker-signal combinations produced similar thresholds when the target signal and masker frequencies are equal.

In another experiment, Moore *et al.* measured masking patterns produced by sinusoidal and noise maskers using noise targets. The masking patterns were determined for a variety of masker levels and frequencies, while noise targets covered a wide range of frequencies below and above the masker. Experimental results were similar to those reported in experiment 1. Once again, they observed masking pattern slopes that were almost level invariant on the low frequency side, whereas patterns became steeper as the level increased on the high frequency side. Also, the tips of the masking patterns using the sinusoidal masker were flatter and well below the tips using the noise masker. The tonal masker produced more masking than the noise masker for higher masker levels when the target signal frequency was more than 500 Hz above the masker frequency.

2.8 The Temporal Course of Masking

In Section 2.5, masking effects were categorized according to the temporal relation between masker and target. Since they exhibit different properties, masking types should be accurately identified when analyzing signals. This represents a challenging requirement when considering complex signals such as speech or music.

All of the experimental results presented thus far were concerned with a specific type

of masking. Masker and target signal characteristics generally remain constant throughout the duration of experiments. For instance, the level and centre frequency of tonal maskers are maintained constant throughout the duration of an experiment. Moreover, the temporal relation between masker and target is clearly defined according to the desired masking type. In simultaneous masking experiments, maskers are considered stable as they have been present for a long duration with stable parameters when the target is presented. Similarly, in the temporal masking experiments, the masker is introduced or removed completely when the target is presented.

For general audio signals, such experimental assumptions regarding maskers and targets are invalid. Masking components in audio signals exhibit a complex temporal evolution. It is generally difficult to isolate the beginning or the end of a masker. In addition to their birth and death processes, masker characteristics, such as level and centre frequency, can vary from one analysis frame to another. It is thus important to consider the temporal evolution of masking effects rather than accounting for classical simultaneous and temporal masking individually. Changes in masking threshold should be determined from the temporal position of targets with respect to time-varying maskers.

2.8.1 Temporal Variation of Masker Spectral Properties

The auditory system is capable of tracking changes in masker characteristics as they vary in time. Zwicker examined temporal masking patterns produced by sounds periodically modulated in frequency [14]. In one experiment, the 1.5 kHz centre frequency of a tonal masker was sinusoidally modulated. A centre frequency swing of $\pm$700 Hz was used with a variety of modulation rates. A short test tone was presented for detection at different temporal positions during the period and at frequencies spanning the modulation range. The experimental results indicate that the ear is capable of tracking the temporal variations of the masking pattern for centre frequency modulation rates approaching 8 Hz. Above this rate, the deviation in masker frequency results in a flat masking threshold. Rapidly varying tones exhibit similar masking patterns to those produced by narrowband noise maskers. The critical modulation rate corresponds to a frequency variation of $\pm$112 Hz using an analysis window of 20 ms, which is approximately one critical band in the 1.5 kHz range.

The situation is rather different when considering noise maskers. Experimental results regarding the temporal variation of noise masker spectral characteristics are few. Further-

more, changes in noise maskers are difficult to identify when analyzing audio signals. A proper definition of the noise masker bandwidth is required, which is the topic of discussion in Section 2.10. Additionally, noise components have inherent random amplitude fluctuations. Frequency components within the masker bandwidth are intrinsically different from one analysis window to another.

2.8.2 Target Temporal Position

In addition to their time-varying spectral characteristics, masking components sporadically appear and disappear throughout the flow of an audio signal. As a result, masking patterns vary as a function of the temporal position of target signals with respect to the birth and death of sinusoidal maskers [35, 36, 37, 38]. Each of these experiments have shown a considerable elevation of the signal threshold near the masker onset and offset. The overshoot effect has been used to describe the increase in signal masked threshold when the signal is closer to the masker onset [35]. The overshoot effect is illustrated in Figure 2.10, where a tonal masker of 500 ms duration is presented to the listener. The audibility threshold of a shorter target signal (20 ms duration) is measured as a function of its temporal position with respect to the masker onset. Elevations in the order of 15 dB, decaying over a 100 ms second period, have been observed near the masker onset [35]. Although present, the effect is less significant near the masker offset.

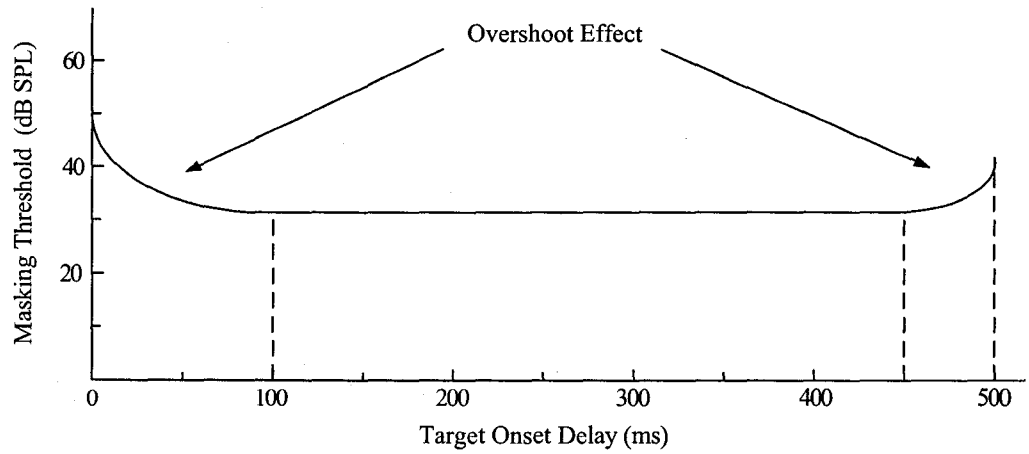


Fig. 2.10 Masking threshold of a tonal target as a function of its temporal position with respect to the onset of a 500 ms tonal masker, based on [35].

The overshoot effect is commonly considered to be the result of the spectral splatter produced at the masker onset, similarly to the effect of windowing in signal processing. When a signal is windowed, its frequency response is convolved with the frequency response of the window. As a result, the energy of the signal is spread into neighbouring frequency components. The frequency response of a tonal masker reassembles more that of a noise masker in such conditions. Following a series of experiments, Smith suggested that short-term neural adaptation might also contribute to overshoot effects [39]. In a later study [38], Bacon and Moore concluded that transient responses of the masker near its onset reduce the usefulness of the transient responses of the target, making the signal more difficult to detect. The process was referred to as *transient masking.*

As for narrowband maskers, conflicting results have been reported. According to Zwicker, the overshoot effect does not appear for narrowband maskers [14]. Zwicker also presented results that contradict those of Elliot, who observed an overshoot when the target signal is above the narrowband noise masker in frequency [40].

2.9 Additivity of Masking

Music and speech signals are composed of a multitude of complex spectral components. Accordingly, several masker components interact to yield an overall masking threshold. A variety of models have been suggested to represent the combination of masking effects from multiple maskers. As reported by Veldhuis in [29], Theile *et al.* proposed a conservative addition law for masking thresholds [41]. The total masking threshold at any given frequency was chosen as the maximum of all individual masking thresholds and the threshold in quiet.

Green first reported a study concerning the addition of masking in 1967 [42]. He measured masking thresholds for a sinusoidal masker and a narrowband noise masker when presented individually, calibrating them such that they produced an equal amount of masking. According to the traditional power spectrum model of masking[2], the output of the auditory filter centred on the target signal frequency should increase by 3 dB when the two maskers are combined, resulting in an equal amount of increase in masking. However, Green measured between 9 and 13 dB more masking when the two maskers were pre-

[2]In the power spectrum model of masking, the power of different maskers is linearly added at the output of the auditory filter.

sented simultaneously. This corresponds to as much as 10 dB of additional masking above the amount predicted by the power spectrum model. The amount of additional masking above 3 dB produced by combining maskers is commonly referred to as *excess masking* in literature [26].

Green explained excess masking by arguing that two different modes of sensory processing were employed for signal detection, depending on which masker type was presented. He described a system $\mathcal{N}$ that provided effective cues for signal detection when the noise masker was presented alone. Similarly, a sensory processing system $\mathcal{S}$ was identified for signal detection in the presence of a sinusoidal masker. When the two maskers were combined, the noise signal rendered system $\mathcal{S}$ ineffective while the sinusoidal signal rendered system $\mathcal{N}$ ineffective. Green attributed the increase in masking threshold to the inadequacy of individual detection cues, resulting from the blending of both masker types.

Lutfi also examined the additivity of simultaneous maskers in a more recent study [43]. He limited his experiments to spectrally non-overlapping maskers in order to restrain interaction between masking components. He measured excess masking in the range of 10 to 17 dB when two equally intense maskers were combined. His results were obtained for various masker combinations, using two sinusoidal maskers, two narrowband noise maskers or a sinusoidal masker with a narrowband noise masker. Moreover, Lutfi observed similar results whether the two maskers were on the same side or opposite sides in frequency of the target signal. Another interesting finding was that the amount of excess masking obtained by combination of two maskers was independent of the individual masker levels.

Lutfi argued through various experiments that neither different types of sensorial processing modes, as suggested by Green [42], nor off-frequency listening were adequate to explain these results. He rather suggested a model in which the effects of each masker are summed after undergoing independent compressive transforms. Let M_A and M_B represent the amount of masking produced by maskers A and B when presented individually. The combined masking threshold, M_{AB}, can be computed using a non-linear transform,

$$F(M_{AB}) = F(M_A) + F(M_B). \tag{2.9}$$

Lutfi found a transform that accurately matched his data, given by:

$$F(M_A) = \left(10^{\frac{M_A}{10}}\right)^p. \tag{2.10}$$

A value of the constant p ranging from 0.20 to 0.33 best fit the results of his experiments. Note that when p is equal to 1, the model reduces to the traditional power spectrum model of masking and no excess masking is predicted. Finally, he generalized the compression model for any arbitrary number K of spectrally non-overlapping maskers, given by:

$$F(M_{AB...K}) = F(M_A) + F(M_B) + \ldots + F(M_K). \tag{2.11}$$

Subsequently, Lutfi applied his model to predict results from previous studies that measured masking by sounds with various complex spectra [44]. He considered an extensive range of studies that included different masker combinations as well as a variety of individual masker levels. He applied his model to results from a study by Canahl [45] that combined four tonal maskers as well as studies by Nelson [46] and Zwicker [47] that combined two tonal maskers. He also considered the original experiment by Green [42] which employed a tonal and noise masker, along with a study by Patternson and Nimmo-Smith [48] where two noise maskers were presented simultaneously. Lutfi discovered that, in such experiments where there was minimal spectral overlap between maskers, setting $p = 0.33$ predicted masking that was in agreement with the collected data. Following, he applied his model to an experiment by Bilger [49] in which there was a considerable overlap between two noise maskers. He concluded that fixing $p = 0.5$ yielded a considerably more accurate prediction than the traditional linear model in such a situation.

Moore found it important to evaluate the data that Lutfi used in the derivation of the non-linear model of addition since it corresponded to a large divergence from the traditional power spectrum model of masking [50]. He presented two experiments that showed cases under which the compressive non-linearity model fell short. Moore argued that the excess masking measured by Lutfi was influenced by two factors: (1) combination-product detection and (2) the use of different detection cues for the single masker and masker pairs. The latter of these two factors was particularly important in an experiment where two narrowband noise maskers were combined. He demonstrated that excess masking occurred only if the two maskers had uncorrelated envelope fluctuations. This supports the idea that subjects use envelope fluctuations as a detection cue in the presence of a single noise masker. When two uncorrelated maskers are combined, the effectiveness of this detection cue is reduced and excess masking occurs. However, the cue remains when the noise bands have correlated envelope fluctuations and can occasionally lead to a release in masking. The

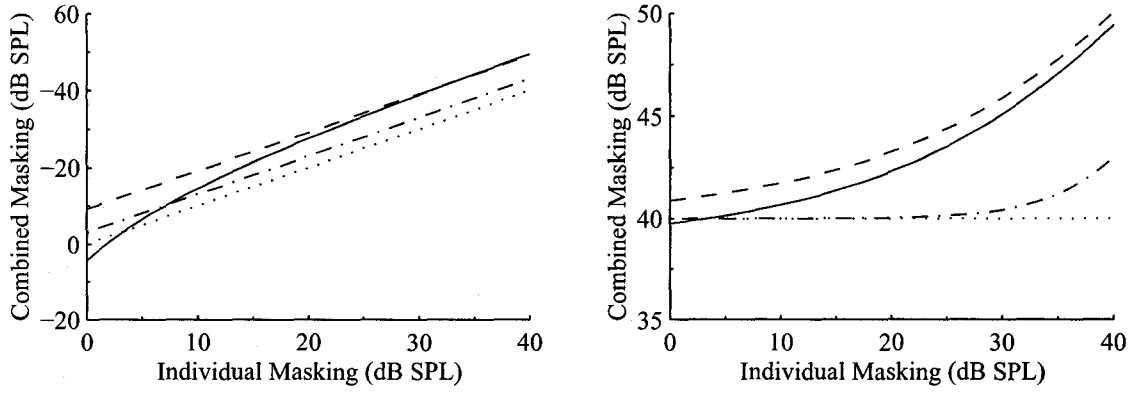
decrease in masking resulting from the combination of correlated noise maskers is referred to as *Comodulation Masking Release.*

More recently, Humes and Jesteadt investigated the additivity of simultaneous masking effects for a variety of masker conditions [51]. They reconsidered Lutfi's power-law model [43] as well as Humes *et al.*'s modified power-law model with compressed internal noise [52]. The latter regards the internal noise of the ear as a distinct masker that produces a continuous masking threshold, more commonly known as the threshold in quiet. Observed masking thresholds result from the combination of the internal noise masker with external maskers, which have all undergone a compressive transform. As such, the modified power-law model makes more sensible predictions for weak maskers. If M_A corresponds to the masking threshold produced by masker A and Q_T represents the threshold in quiet, the nonlinear transform is given by:

$$F(M_A) = \left(10^{\frac{M_A}{10}}\right)^p - \left(10^{\frac{Q_T}{10}}\right)^p. \tag{2.12}$$

In their study, Humes and Jesteadt recognized that interactions between maskers affect the contribution of each masker to the overall threshold. They suggested that inter-masker suppression might play a significant role in the combination of temporally overlapping maskers. Nevertheless, the modified power-law model accurately predicted masking in the context of Moore's experiment [50] which contradicted Lufti's additivity model. More generally, a parameter p ranging from 0.1 to 0.3 provided close estimates for masking additivity in all of their experiments.

The different masking additivity models are compared in Figure 2.11, where the predicted combined masking threshold produced by two maskers is illustrated. In Figure 2.11(a), both maskers A and B produce an equivalent amount of masking when presented individually. The combined threshold is shown as a function of the individual levels. In Figure 2.11(b), the individual masker threshold produced by masker A is fixed to 40 dB while the threshold produced by masker B is varied. Predictions are provided for Theile's model, the power spectrum model, the power-law model and the modified power-law model. It is evident that masking estimates widely vary depending on the additivity model, especially when individual masking thresholds are comparable in magnitude. As much as 9 dB in excess masking is predicted using the power-law and modified-power law models when individual masking thresholds are equal.



(a) Combined masking threshold when maskers A and B produce equal amounts of individual masking.

(b) Combined masking threshold when masker A individually produces a 40 dB threshold and masker B is varied.

Fig. 2.11 Predicted masking threshold produced by the combination of maskers A and B. Dash-dotted line: masking predicted using the power spectrum model. Dotted line: masking predicted using Theile's model. Dashed line: masking predicted using the power-law model. Solid line: masking predicted using the modified power-law model.

Masker signals are always clearly defined in experiments related to the addition of masking. However, this is not the case when considering complex audio signals such as speech and music. Veldhuis and Kohlrusch advised using a conservative criterion for masker additivity when applied to signals with a complex spectrum, since the interactions between maskers are unknown [53]. Given the compelling evidence in favour of excess masking, it is suggested to employ a compressive transform with a conservative parameter, *e.g.*, $p = 0.5$, for the addition of masking. However, a clear definition of masking components is required if the power-law model is to be applied. The problem of identifying maskers within a complex signal is addressed in the following section.

2.10 Masker Integration

The characteristics of masking components used in psychoacoustic experiments are generally very well defined within the experimental setup. Unfortunately, the distinction between maskers in the analysis of audio signals is much more ambiguous. Audio signals are com-

posed of a variety of complex elements that combine to form the spectrum of the signal. The question arises as to where should the spectral boundaries between masking components be positioned. For example, a portion of the spectrum can either be entirely integrated as a single noise masker, or partitioned and integrated as multiple noise maskers.

The importance of masker integration increases when considering non-linear models for the prediction of masking. For instance, estimates from level dependent masking patterns are influenced by masker spectral boundaries. Masker sound pressure levels are directly related to the bandwidth over which they are integrated. Thus, multiple noise components would not yield the same masking threshold as would a single larger masker within the same bandwidth. Similarly, the masking additivity models that were introduced in Section 2.9 are greatly influenced by the number of maskers as well as their sound pressure levels. Masking additivity is linear within the bandwidth over which maskers are integrated, whereas the additivity of distinct maskers is non-linear.

Humes and Lee performed a study that explored the spectral boundaries for the non-linear additivity of simultaneous masking [54]. In their experiment, they measured masking thresholds for a variety of configurations that combined two noise maskers. The amount of spectral overlap between the maskers was varied from no overlap to complete overlap. Their results showed that the masking effects from both components were added linearly when they overlapped within a critical band, centred at the signal frequency. On the other hand, the modified power law model [51] provided a good estimate to the additivity of masking for spectrally non-overlapping maskers.

These results suggest that the spectrum of a complex audio signal should be partitioned into a discrete set of non-overlapping critical bands. The sound pressure level of each masker is obtained by integrating the short term power spectral density over the bandwidth of the corresponding critical band.

2.11 Target Integration

Masking patterns are significantly influenced by the nature of the signals being masked. As previously mentioned, target signals result from the quantization process in audio compression. However, spectral boundaries between noise targets are not apparent since the quantization noise occupies the entire bandwidth of the coded signal. A clear definition of noise targets is necessary if masking is to be predicted.

Veldhuis seemingly addressed this issue after he identified that experimental results using tonal targets were inappropriate for quantization noise targets [29]. He proposed the definition of unit Bark noise targets since the ear integrates sound energy over critical bands. Although he provided little evidence, Veldhuis' ideas were consistent with other psychoacoustic findings. For instance, consider the excitation pattern model of masking that was introduced in Section 2.6. If the output of any auditory filter differs by more than 1 dB when the masker and target are combined, the target signal is audible. Since auditory filter bandwidths correspond to critical bandwidths, it appears reasonable to consider unit-Bark targets. Furthermore, the majority of reported experiments using noise targets employ critical band-wide noise bands.

2.12 Chapter Summary

This chapter introduced essential concepts concerning the perception of audio signals. The human auditory processing system was presented and the notion of masking described. Psychoacoustic results were subsequently introduced. These form the basis for the application of masking in audio coding.

Although numerous findings were presented, insufficient data is available to adequately model masking effects when considering complex signals. The majority of reported experiments involve relatively simple sound stimuli. The large number of masking components in audio signals exhibit complex interactions with each other and with the quantization noise targets. Such complex interactions have not yet been addressed in the literature and their detailed effects with respect to masking are unknown.

Nevertheless, some of the results that were presented in this chapter can be used to improve the accuracy of existing auditory models. For instance, the application of the excitation pattern model of masking (Section 2.6), the temporal course of masking (Section 2.8) and the discussions regarding masker and target integration (Section 2.10 and Section 2.11) represent potential contributions to the estimation of masking. Perceptual models that predict the amount of masking produced by audio signals are presented in the following chapter.

Chapter 3

Auditory Masking Models

Several of the auditory characteristics that were presented in Chapter 2 have been employed in audio coding to model masking effects of sound stimuli. As a result, various masking models have been proposed with different levels of accuracy and complexity. Four recognized auditory models that predict the amount of masking are presented in this chapter. All of these models are based on the excitation pattern model of masking, which was introduced in Section 2.6. Finally, a novel auditory model is presented that attempts to solve several of the inadequacies of current models.

3.1 Johnston's Model

Johnston proposed an auditory masking model in [55] that was largely based on the work of Schroeder [23]. Johnston's model was used to derive a short-term spectral masking threshold, from which quantization noise was shaped in a transform coder. The model operates on 64 ms frames of audio signals sampled at 32 kHz, yielding a spectral resolution of 15.625 Hz per frequency bin.

The first step in threshold calculation corresponds to critical band analysis. The short-term power spectrum is obtained from the complex spectrum of the input signal as follows:

$$P[k] = \mathrm{Re}^2\big(X[k]\big) + \mathrm{Im}^2\big(X[k]\big), \tag{3.1}$$

where $X[k]$ represent the DFT coefficients of the input signal[1]. The energy in each critical band is calculated by partitioning the power spectrum into non-overlapping maskers, each having unit-Bark width,

$$B[i] = \sum_{k=b_{li}}^{b_{hi}} P[k], \tag{3.2}$$

where b_{li} and b_{hi} are the lower and upper boundaries of critical band i and $B[i]$ is the band energy.

The critical band spectrum is spread in order to estimate masking effects between unit-Bark maskers. The spreading function, S, which has lower and upper slopes of 10 dB/Bark and -25 dB/Bark, is described analytically by:

$$S_{i,j} = 15.81 + 7.5\big((j-i) + 0.474\big) - 17.5(1 + ((j-i) + 0.474)^2)\big)^{1/2} \quad \text{dB}, \tag{3.3}$$

where i and j represent the Bark indices of the target and masker bands respectively. The spreading function is independent of both the centre frequency and level of the masking signal. The spread Bark spectrum is obtained by convolving the Bark spectrum with the spreading function,

$$C[i] = \sum_{j=1}^{Z} S_{i,j} B[j], \tag{3.4}$$

where $C[i]$ denotes the spread energy in band i and Z is the total number of critical bands. The convolution is carried out in the power spectrum domain, hence requiring the conversion of $S_{i,j}$ from its decibel representation. The convolution between the spreading function and the Bark spectrum is efficiently implemented as a matrix multiplication by forming the spreading matrix $\mathbf{S}_{\text{ij}}$, resulting in:

$$\mathbf{c} = \mathbf{S}\,\mathbf{b}. \tag{3.5}$$

The noise threshold is obtained from the spread Bark spectrum by subtracting from it an offset (in decibels), which is dependent on the nature of the masking signal. As tonal maskers and noise maskers generate different masking patterns, Johnston uses the Spectral Flatness Measure, SFM, to characterize the tonality of the signal. The Spectral Flatness

[1]For a frame of the input signal, the coefficients of a Discrete Fourier Transform (DFT) correspond to the samples at equally spaced frequencies of the Discrete-Time Fourier Transform (DTFT).

Measure is a ratio of the Geometric Mean (GM) to the Arithmetic Mean (AM) of the power spectrum,

$$\mathrm{SFM_{dB}} = 10 \log_{10} \frac{\mathrm{GM}}{\mathrm{AM}}. \tag{3.6}$$

The result is subsequently converted to a tonality coefficient α, according to:

$$\alpha = \min \left(\frac{\mathrm{SFM_{dB}}}{\mathrm{SFM_{dBmax}}}, 1 \right), \tag{3.7}$$

where $\mathrm{SFM_{dBmax}} = -60$ dB. A signal that is purely tonal would yield a coefficient $\alpha = 1$, whereas a noise-like signal has a coefficient α approaching zero.

The tonality index, α, is used to geometrically weight the two different threshold offsets produced by tonal maskers and noise maskers. For the former, Johnston estimates the noise threshold to be $14.5 + i$ dB below the spread Bark spectrum $C[i]$, where i is the Bark frequency. Noise maskers have a uniform masking index of 5.5 dB across the Bark spectrum. The resulting offset, $O[i]$, to be subtracted from $C[i]$, is given by:

$$O[i] = \alpha(14.5 + i) + 5.5\,(1 - \alpha) \quad \mathrm{dB}. \tag{3.8}$$

The spread threshold, $T[i]$, is computed as:

$$T[i] = 10^{\log_{10}(C[i]) - (O[i]/10)}. \tag{3.9}$$

The next step involves the inversion of the convolution operation that was required in the computation of the spread threshold, $T[i]$. The rationale underlying this process is related to the excitation pattern model of masking. A target signal is audible if the difference between the excitation patterns produced by the masker individually and the combination of masker and target is less than a critical value. Assuming linear summation of excitation patterns in the power domain, $T[i]$ represents the maximum acceptable excitation produced by the target signal. As a result, the noise power is obtained by deconvolving its excitation pattern, $T[i]$, with the spreading function. Johnston argued that this procedure is highly unstable, owing to the shape of the spreading function. He proposed a renormalization of the noise energy threshold rather than deconvolution. The renormalization multiplies each $T[i]$ by the inverse of the energy gain per band, assuming each band has unit energy. This compensates for the increase in critical band energy estimates that result from the

convolution with the spreading function. The normalized threshold is designated as $\tilde{T}[i]$.

Finally, the normalized threshold, $\tilde{T}[i]$, is compared to the absolute threshold of hearing. In view of the fact that the playback levels are unknown, absolute thresholds are established such that a signal of 4 kHz, having a peak amplitude of ± 1 least significant bit (assuming 16-bit coding per sample), is at the threshold of audibility. Equivalently, the assumed playback level of a full scale sinusoid is 86.9 dB SPL. The maximum value between $\tilde{T}[i]$ and the absolute threshold of hearing is chosen within each critical band, yielding the final masking threshold.

3.2 MPEG-1 Psychoacoustic Model 1

The Moving Pictures Expert Group (MPEG) draft [6] provides two informative psychoacoustic models that compute the just-noticeable level of noise for signal coding. This section describes the first of the auditory masking models, while the second is presented in Section 3.3. The output of the auditory model is a Signal-to-Mask Ratio for each coder sub-band. The model operates on different frame sizes, depending on the sampling rate of the input signal.

The masking threshold is computed from the short-term power spectral density estimate of the input signal. The power density spectrum is obtained from the FFT of the input signal, following multiplication by a Hann window. The magnitude of each spectral component is converted to a decibel scale, yielding the estimate $P[k]$. The power spectrum is normalized to an anticipated playback level of 96 dB SPL, such that the maximum spectral component corresponds to this value.

The following step involves the discrimination between tonal and noise maskers. This accounts for the dependence of masking thresholds on the nature of the maskers. Firstly, tonal components are identified through the detection of local maxima within the power spectrum. A component is labelled as a local maximum if $P[k] > P[k-1]$ and $P[k] \geq P[k+1]$. Components are declared as tonal if $P[k] - P[k+j] \geq 7$ dB, where j lies within a neighbourhood that is dependent on the centre frequency, k. The sound pressure level of the tonal masker, $P_{TM}(z)$, where z is the Bark value of the frequency line k, is computed as follows:

$$P_{TM}(z) = 10 \log_{10} \left(10^{P[k-1]/10} + 10^{P[k]/10} + 10^{P[k+1]/10} \right). \tag{3.10}$$

Tonal maskers are removed from the power spectrum, $P[k]$, by setting all frequency lines within the examined range to $-\infty$. The sound pressure levels of noise maskers are obtained by summing the energies of spectral lines within each critical band, yielding $P_{NM}(z)$.

Subsequently, the number of maskers considered for threshold computation is reduced. At first, only maskers having a level above the absolute threshold of hearing are retained. A decimation process then occurs between multiple tonal maskers that lie within half of a critical band. The tonal masker having the highest level is maintained while the other elements are removed from the directory of maskers.

The contribution of each masker to the overall masking threshold is evaluated. The spread of masking effects is modelled using the spreading function described below:

$$s(z_j, \Delta z, P_M(z_j)) = \begin{cases} 17\,\Delta z - 0.4\,P_M(z_j) + 11 & \text{for } -3 \le \Delta z < -1, \\ (0.4\,P_M(z_j) + 6)\Delta z & \text{for } -1 \le \Delta z < 0, \\ -17\,\Delta z & \text{for } 0 \le \Delta z < 1, \\ (0.15\,P_M(z_j) - 17)\Delta z - 0.15\,P_M(z_j) & \text{for } 1 \le \Delta z < 8 \\ -\infty & \text{otherwise,} \end{cases} \tag{3.11}$$

where z_j, Δz and $P_M(z_j)$ represent respectively the masker Bark frequency, the Bark frequency separation between the masker and target and the sound pressure level of the masker. The spread of masking is only considered within the range of $-3 \le \Delta z < 8$, for reasons of implementation complexity.

The masking indices for tonal maskers, $a_{TM}(z)$, and noise maskers, $a_{NM}(z)$, expressed below in Eq. (3.12), are both frequency dependent. They represent the offset to be subtracted from the excitation pattern of the masker to obtain the masking pattern.

$$a_{TM}(z) = -1.525 - 0.257\,z - 4.5 \quad \text{dB}, \tag{3.12a}$$

$$a_{NM}(z) = -1.525 - 0.175\,z - 0.5 \quad \text{dB}. \tag{3.12b}$$

The individual masking thresholds from each masker are calculated according to:

$$M_{TM}(z_j, \Delta z) = P_{TM}(z_j) + a_{TM}(z_j) + s\big(z_j, \Delta z, P_{TM}(z_j)\big) \quad \text{dB}, \tag{3.13a}$$

$$M_{NM}(z_j, \Delta z) = P_{NM}(z_j) + a_{NM}(z_j) + s\big(z_j, \Delta z, P_{NM}(z_j)\big) \quad \text{dB}, \tag{3.13b}$$

where z_j represents the masker critical band rate and Δz represents the Bark frequency separation between the masker and target. The individual masking thresholds are computed for each coder sub-band, using all maskers. The global masking threshold per coder sub-band is computed by summing the individual masking contributions from each masker along with the absolute threshold of hearing, $T_Q(z)$,

$$M_g(z_i) = 10\log_{10}\left(10^{T_Q(z_i)/10} + \sum_{j=1}^{m} 10^{M_{TM}(z_j,z_i)/10} + \sum_{j=1}^{n} 10^{M_{NM}(z_j,z_i)/10}\right). \tag{3.14}$$

Finally, the Signal-to-Mask Ratios are calculated by subtracting the global masking threshold, $M_g(z)$, from the signal power in each coder sub-band.

3.3 AAC Auditory Masking Model

An informative psychoacoustic model was given in the Advanced Audio Coding (AAC) standard [7] that is practically identical to the second psychoacoustic model presented in [6]. The model evaluates the maximum inaudible distortion energy for the coding of a frame of audio. The outputs are a Signal-to-Mask Ratio (SMR) and an energy threshold for each coder subband. The masking model is scalable in the sense that it accommodates a variety of input signal sampling rates and provides for two different frame lengths.

The first step in threshold calculation is the evaluation of the complex spectrum of the input signal. After multiplication by a Hann window, an FFT is computed, yielding $X[k]$. $X[k]$ is represented in terms of its magnitude, $r[k]$, and phase components, $\phi[k]$,

$$X[k] = r[k]\, e^{j\,\phi[k]}. \tag{3.15}$$

The energy in each coder partition is calculated by summing the energies of each component within a sub-band,

$$e[b] = \sum_{k=k_l}^{k_h} r^2[k], \tag{3.16}$$

where k_l and k_h denote the lower and upper boundaries of sub-band b.

The tonality of the input signal, which is used to estimate the amount of masking produced, is estimated using a method proposed in [56]. Rather than providing a global

value, this method evaluates a local tonality index for each sub-band that is estimated by means of a coherence measure. The coherence measure corresponds to a prediction of spectral components, in polar coordinates, from the spectrum of the two preceding frames,

$$r_{\text{pred}}[k] = r_{t-1}[k] + \big(r_{t-1}[k] - r_{t-2}[k]\big), \tag{3.17a}$$

$$\phi_{\text{pred}}[k] = \phi_{t-1}[k] + \big(\phi_{t-1}[k] - \phi_{t-2}[k]\big), \tag{3.17b}$$

where r_{pred} and ϕ_{pred} represent the predicted magnitude and phase. Each component prediction pair is transformed to an unpredictability measure, $c[k]$, by comparing with the actual spectral pair,

$$c[k] = \frac{\text{dist}\big(X[k], X_{\text{pred}}[k]\big)}{r[k] + \big|r_{\text{pred}}[k]\big|}, \tag{3.18}$$

where the distance operator is defined as:

$$\begin{aligned} \text{dist}\big(X[k], X_{\text{pred}}[k]\big) =& \big|X[k] - X_{\text{pred}}[k]\big| \\ =& \Big(\Big(r[k]\cos\big(\phi[k]\big) - r_{\text{pred}}[k]\cos\big(\phi_{\text{pred}}[k]\big)\Big)^2 + \\ & \Big(r[k]\sin\big(\phi[k]\big) - r_{\text{pred}}[k]\sin\big(\phi_{\text{pred}}[k]\big)\Big)^2\Big)^{1/2}. \end{aligned} \tag{3.19}$$

The partition unpredictability, $c[b]$, is computed by weighting each unpredictability measure by its component energy, as follows:

$$c[b] = \sum_{k=k_l}^{k_h} c[k]\, r^2[k]. \tag{3.20}$$

Subsequently, the partition energy, $e[b]$, and unpredictability measure, $c[b]$, are individually convolved with the spreading function. The resulting spread partition energy and spread unpredictability are respectively denoted $e_s[b]$ and $c_s[b]$. The spreading curve employed by the AAC model is referred to as the rounded modified function [57], as shown in Fig. 3.1. Following the convolution, the spread unpredictability, $c_s[b]$, is normalized by the spread energy, $e_s[b]$. This procedure is required because of the weighting by the signal energy that was involved in the computation of the unpredictability measure. The result of the normalization is designated as $\tilde{c}_s[b]$.

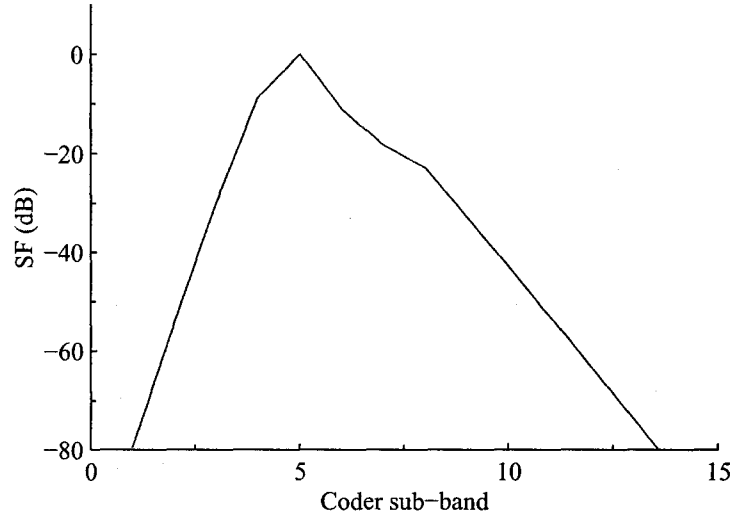


Fig. 3.1 AAC auditory masking model spreading function, centred at the 5^{th} coder sub-band.

The spread energy, $e_s[b]$, is normalized as well due to the nature of the spreading function. This accounts for the increase in energy estimates within each sub-band that results from the convolution operation. The normalized spread energy is denoted $\tilde{e}_s[b]$, as shown below:

$$\tilde{e}_s[b] = \frac{e_s[b]}{n[b]}. \tag{3.21}$$

The normalization factor, $n[b]$, is given by:

$$n[b] = \sum_{i=1}^{b_{\max}} s[i,b], \tag{3.22}$$

where $s[i,b]$ represents the spreading between masker i and target b.

Subsequently, the unpredictability, $\tilde{c}_s[b]$, is converted to a tonality index, $t[b]$. The tonality index is limited to the range $0 < t[b] < 1$, where a purely tonal signal would yield a unit value,

$$t[b] = -0.299 - 0.43 \log_e \left(\tilde{c}_s[b]\right). \tag{3.23}$$

The required Signal-to-Noise Ratio per band for noise inaudibility is computed from the tonality indices. The value for Noise-Masking-Tone (NMT) is selected as 6 dB for all

bands, while the value for Tone-Masking-Noise (TMN) is selected as 18 dB.

$$\mathrm{SNR}[b] = 18\,t[b] + 6\big(1 - t[b]\big) \tag{3.24}$$

Next, the noise energy threshold, $n[b]$, is calculated from the signal energy and the required SNR per band,

$$n[b] = \tilde{e}_s[b]\,10^{-\mathrm{SNR}[b]/10}. \tag{3.25}$$

The AAC auditory model avoids pre-echoes[2] in threshold calculation by comparing the current noise energy threshold with the noise energy threshold calculated in the previous frame, $n_{\mathrm{prev}}[b]$. The comparison is accomplished on a per band basis, as follows:

$$n[b] = \min\Big(n[b], \big(2\,n_{\mathrm{prev}}[b]\big)\Big). \tag{3.26}$$

This last step ensures that the masking threshold is not biased by a high energy onset occurring near the end of an analysis window. The resulting noise threshold is compared with the threshold in quiet. The maximum between the two values is retained within each band.

Finally, the Signal-to-Mask Ratio is calculate per coder sub-band, as shown below:

$$\mathrm{SMR}[b] = 10\log_{10}\left(\frac{e[b]}{n[b]}\right). \tag{3.27}$$

3.4 PEAQ Model

An auditory model was developed by the International Telecommunications Union (ITU) within the framework of the Perceptual Evaluation of Audio Quality (adopted as ITU-R BS.1387) [13]. PEAQ provides advanced metrics for the assessment of the perceptual quality of audio signals. Among other model output variables, a masking threshold is estimated from the auditory model. This section describes the various steps involved in the computation of the masking threshold within PEAQ. It is worthy of note that this model is the most advanced studied thus far.

The masking threshold is calculated from an FFT-based model of the ear. A 48 kHz

[2]Pre-echoes occur when a signal with a sharp onset begins in the later segment of an analysis window, following a period of low energy [3].

input signal is segmented into frames of 2048 samples having 50% overlap. Following the multiplication by a Hann window, a short-time discrete Fourier transform is computed. The frequency domain signal, $X[k]$ for $0 \leq k \leq (N_F - 1)$, is normalized to a recommended playback level of 92 dB SPL.

The frequency domain signal undergoes an attenuation process which combines the filtering effects of the outer and middle ear. An expression was adopted from [19] for the transfer function of the outer–middle ear system, given by:

$$\begin{aligned} A_{\text{dB}} &= -2.184(f/1000)^{-0.8} + 6.5\, e^{-0.6(f/1000-3.3)^2} - 10^{-3}(f/1000)^{3.6}, \\ W(f) &= 10^{A_{\text{dB}}(f)/20}, \end{aligned} \tag{3.28}$$

where f is represented in Hz. The vector weights to be applied to the normalized FFT outputs are given by:

$$W[k] = W\left(\frac{kF_S}{N_F}\right), \tag{3.29}$$

where F_S represents the input signal sampling frequency. The outer ear weighted outputs are referred to as $X_W[k]$,

$$\left|X_W[k]\right|^2 = G_L^2 W^2[k] \left|X[k]\right|^2, \tag{3.30}$$

where G_L represents the scaling factor for playback level normalization.

The frequency spectrum is subsequently partitioned into nonoverlapping bands according to a critical band scale. The pitch mapping is calculated from an approximation proposed in [23],

$$z = 7 \operatorname{arcsinh}\left(\frac{f}{650}\right), \tag{3.31}$$

where z represents the critical band rate in Bark units. The size of each frequency group corresponds to a resolution of 0.25 Bark, resulting in 109 bands. Each group is characterized by a lower frequency, $f_l[i]$, a centre frequency, $f_c[i]$ and an upper frequency, $f_u[i]$. The energies of the outer ear weighted outputs are summed within each frequency group, yielding $P_e[i]$. The final pitch patterns, $E[i]$, are obtained by adding a frequency dependent offset that represents the internal noise of the inner ear. The expression used to model the internal noise, E_{IN}, is detailed below in Equation (3.32). The result of this stage is a decibel

representation of the pitch patterns, $L[i]$.

$$\begin{aligned} E_{\mathrm{INdB}}(f) &= 1.456(f/1000)^{-0.8}, \\ E_{\mathrm{IN}}[i] &= 10^{E_{\mathrm{INdB}}(f_c[i])/10}. \end{aligned} \tag{3.32}$$

The excitation patterns are obtained through the spreading of the pitch pattern energies. The spreading function is adopted from an auditory model developed by Terhardt [19]. The upper-frequency slope is assumed to be dependent of masker sound pressure level and frequency, while the lower slope is independent of such factors. The spread of excitation is performed according to:

$$S_{\mathrm{dB}}(i,l,E) = \begin{cases} 27(i-l)\Delta z, & i \leq l, \\ \left(-24 - \frac{230}{f_c[l]} + 2\log_{10} E\right)(i-l)\Delta z, & i \geq l, \end{cases} \tag{3.33}$$

where i and l represent respectively the target and masker bands. The resulting spreading function is given by:

$$S(i,l,E) = \frac{1}{A(l,E)} 10^{S_{\mathrm{dB}}(i,l,E)/10}. \tag{3.34}$$

The denominator in the expression, $A(l,E)$, is used to normalize the spread of energy. Due to its shape, the spreading function increases the energy estimates in each band. The normalization regulates the energy gain to unity, as $A(l,E)$ is the sum over i of $S(i,l,E)$. The spread energy in band i is computed by summing the spread energy contributions from all bands,

$$E_S[i] = \frac{1}{B_S[i]} \left(\sum_{l=0}^{N_c-1} \left(E[l]S(i,l,E[l])\right)^{0.4} \right)^{\frac{1}{0.4}}. \tag{3.35}$$

An additional term is included to reverse the convolution operation required in the evaluation of the spread energy, as proposed by Johnston [55]. $B_S[i]$ is computed by summing the spread energies of all bands, assuming they have unit energy,

$$B_S[i] = \left(\sum_{l=0}^{N_c-1} \left(S(i,l,E_0)\right)^{0.4} \right)^{\frac{1}{0.4}}. \tag{3.36}$$

Forward masking effects are represented by smearing the band energies over time. A bank of first order low-pass filters is used to model the time domain spreading. The

frequency dependent time constants of the filters are computed as:

$$\tau[i] = \tau_{\min} + \frac{100}{f_c[i]}(\tau_{100} - \tau_{\min}), \tag{3.37}$$

where $\tau_{100} = 0.03$ s and $\tau_{\min} = 0.008$ s. The final excitation patterns are calculated according to:

$$\begin{aligned} E_f[i,n] &= \alpha[i]E_f[i,n-1] + (1-\alpha[i])E_S[i,n], \\ \tilde{E}_S[i,n] &= \max(E_f[i,n], E_S[i,n]), \end{aligned} \tag{3.38}$$

where $\alpha[i]$ is derived from the time constants by:

$$\alpha[i] = \exp\left(-\frac{4}{187.5\,\tau[i]}\right). \tag{3.39}$$

Finally, the masking threshold is computed by subtracting a frequency dependent offset, in decibels, from the excitation patterns. The masking offset, $m(k)$, is given by:

$$m[i] = \begin{cases} 3.0, & 0.25\,i \leq 12, \\ (0.25)^2\,i, & 0.25\,i > 12. \end{cases} \tag{3.40}$$

The resulting masking threshold per band, $M(z)$, is calculated as follows:

$$M[i] = \frac{\tilde{E}_S[i]}{10^{\frac{m[i]}{10}}}. \tag{3.41}$$

3.5 Current Model Inadequacies

Four of the most prominent auditory models have been presented, ranging from Johnston's low complexity model to the high complexity PEAQ model. However, these models do not represent certain psychoacoustic findings that were presented in Chapter 2, as discussed below.

3.5.1 Determination of Sound Pressure Level

The determination of sound pressure level is essential in auditory models where the absolute threshold of hearing and level dependent characteristics are considered. This requires a proper normalization of the spectrum of the audio signal to an expected playback level. All of the methods described thus far normalize the spectrum according to a single frequency component. Once the spectrum is normalized, the sound pressure level of a given frequency band is obtained by summing all of the components within that band. This approach is very sensitive to frequency resolution, which depends on both sampling frequency and frame size. For instance, the sound pressure level within a frequency band will decrease by 3 dB if the frequency resolution is halved. This suggests the necessity for a more robust level computation. The sound pressure level should accurately represent the level presented to the ear, independently of the frequency resolution of the auditory model.

3.5.2 Additivity of masking

Psychoacoustic results have revealed the presence of excess masking when combining the effects of individual maskers. Excluding PEAQ, auditory models generally combine separate masking thresholds using a power spectrum summation. This results in an underestimation of global masking. On the other hand, the PEAQ model employs Lutfi's power-law addition to combine masking effects. Although Lutfi's model yields more accurate predictions, there exist uncertainties regarding the correctness of its application. The root of the problem lies in masker integration, which is described in the following section.

3.5.3 Masker Integration

In Section 2.10, arguments were presented that suggest noise maskers should be integrated over a complete critical band. While all of the auditory models partition the spectrum according to a critical band scale, Johnston's model is the only one that integrates the spectrum over a complete Bark. The other models attempt to increase their resolution by using bands that are fractions of critical bands. The increase in resolution is undesirable when considering non-linear models as it has a significant effect on masking estimates.

3.5.4 Modelling Simultaneous and Temporal Masking

Masking threshold computation generally commences by estimating simultaneous masking effects. The more intricate models, such as AAC [7] and PEAQ [13], also estimate forward masking patterns. The effects of simultaneous and temporal masking are modelled individually and their corresponding thresholds are combined to yield the overall threshold.

Audio signals have much more complex temporal patterns than the masking stimuli that were used in masking experiments. Alluding to Section 2.5, simultaneous masking occurs when the masker and target signals are time aligned. Related masking experiments usually employ steady-state maskers, *i.e.*, components that have been audible for a considerable amount of time. Simultaneous masking models do not consider the duration of masking components. As a result, transient components are modelled as steady-state maskers.

On the other hand, temporal masking occurs prior to the onset of a masker or after its offset. Forward masking models typically use a low pass filter to estimate the temporal masking contribution. This method assumes that all of the components that were present in the previous frame are no longer present in the current frame. This is an incorrect assumption as most signal components last several frames. It is thus important to consider the temporal course of masking components as well as the temporal position of target signals with respect to these maskers, as suggested in Section 2.8.

3.5.5 Application of the Excitation Pattern Model of Masking

The four auditory models presented hitherto originate from the excitation pattern model of masking. In Section 3.1, the deconvolution problem was introduced as a result of the application of this model. Johnson suggested renormalization to approximate the result. This procedure was adopted as well by the PEAQ auditory model. As for the MPEG models, the need for deaspreading remained unacknowledged. The deconvolution issue is addressed later in Chapter 4, where a solution based on the bit allocation scheme is proposed. Rather than despreading the calculated threshold, the distribution of quantization noise takes into consideration its excitation pattern.

3.6 A Novel Auditory Model

The following section describes an innovative auditory model that was developed for the purpose of noise shaping in audio coders. This model predicts masking while incorporating the psychoacoustic results that were presented in Chapter 2.

3.6.1 Time-to-Frequency Mapping

The first step in threshold prediction is a time-to-frequency conversion. A standard FFT is computed from a frame of the input signal, resulting in the spectrum $X[k]$. The sound pressure density is obtained from the spectrum as follows:

$$P[k] = \frac{|X[k]|^2}{\xi}, \tag{3.42}$$

where ξ represents a normalization constant. The sound pressure level of a given masking component is obtained by integrating the density over its occupied bandwidth, as described in Section 2.1. Because of the discrete nature of the FFT, the density is composed of samples having Δf frequency separation. The continuous integration is approximated by summing the density samples, weighted by their frequency interval, Δf. The sound pressure level of the ith component is given by:

$$L[i] = \int_{f \epsilon \mathcal{F}_i} P(f) df \approx \sum_{k \epsilon \mathcal{K}_i} P[k] \Delta f, \tag{3.43}$$

where $\mathcal{F}_i$ represents the frequency range and $\mathcal{K}_i$ represents the corresponding range of discrete indices. The multiplication by Δf guarantees that the SPL accurately represents the level of the continuous-time signal presented to the ear, regardless of the spectral resolution of the model.

The normalization is required to ensure that components are considered at their playback levels. When unknown, the playback level is chosen such that a full scale sinusoid has an overall sound pressure level of 92 dB. This implies that the scaling factor corresponds to the inverse of the power of a full scale sinusoid. Although sinusoids have line spectra, temporal windowing spreads the power of the tone into surrounding components. The power of a sinusoid is calculated using the integral approximation described in Equa-

tion (3.43). The integration should be performed over all frequency components as a result of the window spreading. However, the majority of the power of the sinusoid is contained within three frequency components: the component closest to the centre frequency of the sinusoid and both of its neighbouring components. As a result, the integral approximation only considers these three components.

3.6.2 Masker Identification

Masking components must be identified from the spectrum of the audio signal. This task is accomplished by first locating tonal maskers within the sound pressure density, $P[k]$. Noise maskers are obtained from the remaining spectral components.

A variety of methods have been proposed for the extraction of tonal components from complex spectra. Of these, peak picking has been extensively used [19, 58, 59]. This technique estimates the frequencies of sinusoids as the locations of the peaks within the spectrum. Details underlying peak picking for the current work are similar to those developed by Terhardt [19]. The sound pressure density is scanned for local maxima, such that $P[k] > P[k-1]$ and $P[k] \geq P[k+1]$. Candidates found through this criterion are tested according to:

$$P_{\mathrm{dB}}[k] - P_{\mathrm{dB}}[k+j] \geq 7 \text{ dB}, \tag{3.44}$$

where $j = -3, -2, 2, 3$ and $P_{\mathrm{dB}}[k]$ is the decibel representation of $P[k]$. The latter condition ensures that tonal components lie above neighbouring spectral components by at least 7 dB. The sound pressure level of an uncovered sinusoid is obtained by integrating the density over the region which it is spread. Again, the integration is approximated by summing the central component with the two neighbouring components, weighted by the frequency resolution. The sound pressure level of the ith sinusoid, centred at the kth frequency bin, is given by:

$$L_T[i] = \big(P[k-1] + P[k] + P[k+1]\big)\Delta f. \tag{3.45}$$

Along with its level, $L_T[i]$, the centre frequency of the sinusoid, ω_i, is kept on record,

$$\omega_i = k\,\Delta f. \tag{3.46}$$

Detected sinusoids are removed from $P[k]$ by setting all three components within their summation range to zero. This ensures that noise masker estimation is not biased by tonal

components.

Noise maskers are obtained by partitioning the spectrum into non-overlapping critical bands, as suggested in Section 2.10. The level of a given noise masker, $L_N[i]$, is calculated by integrating $P[k]$ over its corresponding critical band. The frequency-to-Bark conversion is calculated according to the mapping given by Schroeder [23], expressed in Equation (3.31).

3.6.3 Masker Temporal Structure

As previously mentioned, the tracking of masker variations between consecutive frames is essential for modelling the temporal course of masking. In Section 2.8, results on the temporal course of tonal maskers were presented. Considerably fewer results pertaining to the temporal structure of noise masking patterns are available in literature. As such, the temporal structure of noise maskers is neglected. The assumption that masking patterns are invariant to temporal variations of noise maskers is reasonable. By nature, noise bands have inherent random amplitude fluctuations. The power of noise bands usually differ between frames, making it difficult to track their evolution.

The tracking of tonal maskers is accomplished using a method proposed by McAulay and Quatieri for sinusoidal coding [58]. In sinusoidal coding, speech signals are synthesized based on a sinusoidal representation. Sinusoidal parameters estimated in one frame are matched with those of the previous frame to allow a continuous evolution of sinusoids. The process of matching tones is based on the minimization of frequency deviation. The concept of birth and death of sinusoidal components is also introduced to account for rapid movements in spectral peaks. As a result, the duration of each tone is given in number of frames. The algorithm is described below.

Let ω_j^{n-1} and ω_i^n denote respectively the frequencies of the jth tone in frame $(n-1)$ and the ith tone in frame n. N and M represent the number of sinusoids detected in each frame, where $N \neq M$ in general . Further, assume that matches have been found for the first $(i-1)$ sinusoids in frame n. A set of candidate matches for frequency i is evaluated according to the criterion:

$$\left|\omega_i^n - \omega_j^{n-1}\right| < \Delta. \tag{3.47}$$

The matching interval Δ was selected as $0.1 \times \omega_i^n$, based on the work of Levine [60]. This implies that the frequency of a tone can vary by as much as 10% between consecutive frames. In the case where no candidates satisfy Equation (3.47), the ith sinusoid is considered a birth

having a duration of zero frames. This particular scenario is illustrated in Figure 3.2(a). When candidates are available, a tentative match is selected with component l within the set, for which the frequency deviation is smallest, *i.e.*,

$$\left|\omega_i^n - \omega_l^{n-1}\right| < \left|\omega_i^n - \omega_j^{n-1}\right| < \Delta, \quad \forall\, j \neq i. \tag{3.48}$$

A match is not yet declared since a better correspondence for ω_l^{n-1} might be available with another unmatched sinusoid in frame n. The match is verified by ensuring that,

$$\left|\omega_i^n - \omega_l^{n-1}\right| < \left|\omega_k^n - \omega_l^{n-1}\right|, \tag{3.49}$$

for $k > i$. If this condition is satisfied, the match is confirmed between ω_l^{n-1} and ω_i^n, as shown in Figure 3.2(b). If on the other hand Eq. 3.49 fails, two situations are possible. Firstly, no other frequency lies within the interval Δ and sinusoid i is declared a birth. In the second case, frequency ω_{l-1}^{n-1} is unmatched and lies within the interval Δ. In this case, a match is declared and the duration of sinusoid i is computed. Both circumstances are illustrated in Figure 3.2(c) and Figure 3.2(d). Once complete, the process is recommenced with sinusoid $(i+1)$ until all sinusoids in frame n have been considered.

After trajectories are formed by minimizing frequency deviation, the amplitudes of matched sinusoids are compared in the search for discontinuities. If sinusoid levels differ by more than 15 dB, they are considered as two different trajectories, resulting in a birth for the current frame and death in the previous. This additional criterion was proposed by Levine within the scope of his work on multiresolution sinusoidal analysis [60]. An example of the sinusoid tracking procedure is illustrated in Figure 3.3.

The output of this stage is the critical band rate and sound pressure level of tonal and noise maskers. Component age, expressed in number of frames, is available as well for sinusoids. Since sinusoidal masking effects are stable after 100 ms, a maximum duration corresponding to this value is considered. For example, a duration ranging from 0 to 4 frames is sufficient for a model employing 20 ms frames.

3.6.4 Excitation Patterns

Following the identification of maskers from the power spectrum of the input signal, masker excitation patterns are computed. Both tonal and noise maskers are spread using the same

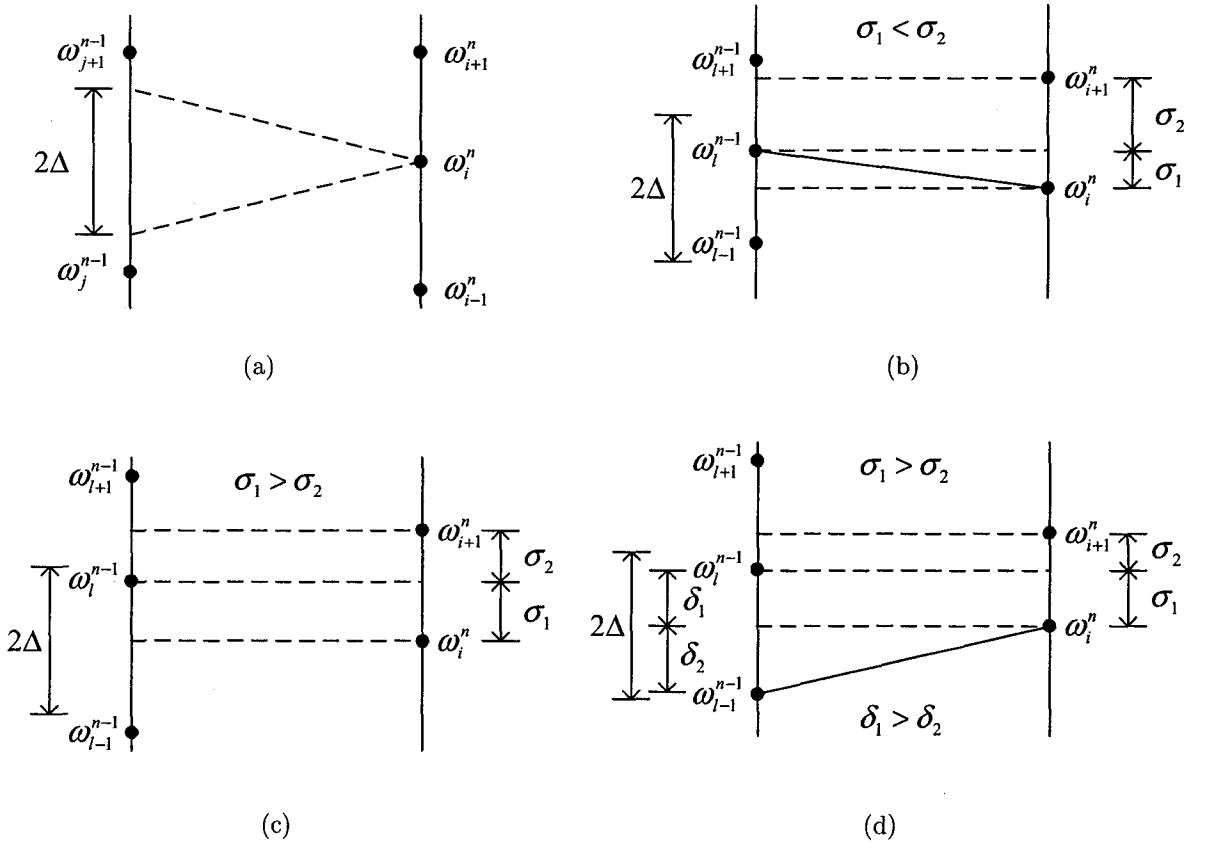


Fig. 3.2 Possible scenarios resulting from the sinusoid tracking procedure, adapted from [58].

upper and lower slopes proposed by Terhardt [19]:

$$S_{\mathrm{dB}}[i,l,L[i]) = \begin{cases} 27(i-l)\Delta z, & i \leq l, \\ \left(-24 - \frac{230}{f_c[l]} + 0.2L[i]\right)(i-l)\Delta z, & i \geq l, \end{cases} \tag{3.50}$$

where $f_c[l]$ and $L[l]$ correspond to the masker centre frequency, expressed in hertz, and sound pressure level. The spread energy of each masker is evaluated per target signal. Target signals are organized as a set of non-overlapping unit-Bark frequency bands that span the entire spectrum of the input signal, as suggested in Section 2.11. The resulting

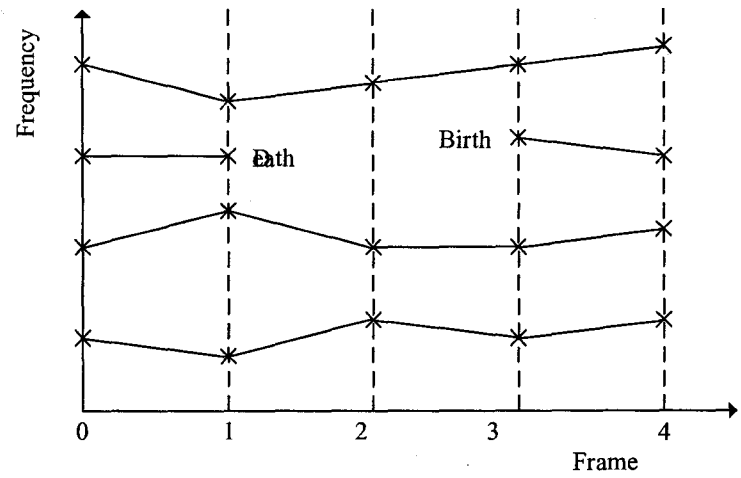


Fig. 3.3 Example of the sinusoid tracking procedure spanning a period of four frames, adapted from [58].

excitation patterns are expressed in matrix format as $\mathbf{E}_N$ and $\mathbf{E}_T$, produced by noise and tonal maskers respectively.

3.6.5 Masking Index

A masking index is subtracted from the individual excitation patterns in order to obtain the masking patterns, $\mathbf{M}_N$ and $\mathbf{M}_T$. The masking index for noise maskers, I_N, is set to 5.5 dB. The masking index for tonal maskers is given by:

$$I_T[i] = I_N + \frac{12.5}{\eta} D_T[i], \tag{3.51}$$

where η represents the maximum component duration and $D_T[i]$ represents the duration of the ith tonal component. As a result, the masking index of a tone birth is identical to that of a noise masker, which is consistent with the argument that energy splatter renders short tones equivalent to noise maskers. On the other hand, the masking index of stable tones corresponds to 18 dB.

3.6.6 Masking Threshold

The masking threshold is obtained by combining the individual masking patterns produced by noise and tonal maskers. Masking additivity is performed according to the power-law method that was described in Section 2.9. The resulting masking threshold per critical band, $M[i]$, is given by

$$M[i] = \left(\sum_{j=1}^{N_N} \mathrm{M}_N^p(i,j) + \sum_{j=1}^{N_T} \mathrm{M}_T^p(i,j) \right)^{\frac{1}{p}}, \tag{3.52}$$

where p was chosen to be 0.5. Coincidently, the selected value of p results in the addition of masking in the magnitude domain rather than in the power domain. Finally, the SMR is evaluated by dividing the signal power per critical band by the masking threshold.

According to the excitation pattern model of masking, the computed threshold, M, corresponds to the audibility threshold for the excitation produced by the target signal. The de-spreading of the masking threshold is addressed in the following chapter, where a solution is presented as part of the bit allocation scheme. As a result, the output of the novel auditory model remains the masking threshold for the excitation produced by the noise targets.

3.7 Chapter Summary

This chapter has presented four auditory masking models that have been developed for the purposes of audio coding and the evaluation of perceptual quality of audio signals. The masking models were described and their shortcomings identified. Finally, a novel auditory model was proposed that considers the psychoacoustic findings that were presented in Chapter 2. The deconvolution problem which results from the application of the excitation pattern model of masking was considered. The proposed model estimates a threshold that represents the maximum inaudible excitation produced by a target signal. Rather than de-spreading the evaluated threshold, a solution to the deconvolution problem was proposed within the adaptive bit allocation algorithm, which is presented in the following chapter.

Chapter 4

Perceptual Bit Allocation for Low-Rate Coding

Various models that predict the amount of masking produced by complex audio signals were presented in the previous chapter. According to the excitation pattern model of masking, such models evaluate the maximum inaudible level of the excitation produced by a quantization noise target. The deconvolution problem was identified whereby masking thresholds must be de-spread in order to evaluate a limit in noise power.

This chapter examines the application of perceptual models to audio coding. Different bit allocation strategies that consider auditory masking are presented. Following, a novel perceptual bit allocation scheme is presented that attempts to solve the deconvolution problem. The proposed algorithm considers the excitation produced by the quantization noise targets in the allocation of information bits to coder sub-bands.

4.1 Adaptive Bit Allocation

Spectral components of the audio signal are generally grouped into coder sub-bands. As such, coder sub-bands are quantized individually; a quantizer is associated with a sub-band. A fixed number of bits are available to represent all of the spectral components in every analysis frame. However, the number of bits allocated to each sub-band can vary from one frame to another. The underlying constraint is that the total number of allocated bits remains constant.

Information bits are allocated to coder sub-bands such that a distortion criterion is

optimized. The process is known as *spectral noise shaping*, where the spectrum of the quantization noise is shaped according to a certain criterion. Distortion measures can be categorized as either non-perceptual or perceptual. In either case, the distortion measure generally depends on the amount of quantization noise present in the reconstructed signal. *Signal-to-Noise Ratio*, among others, is an example of a non-perceptual distortion measure. The amount of noise above the masking threshold (*i.e.*, audible noise) is an example of a perceptual criterion. The allocation of bits to coder sub-bands is performed using the greedy algorithm, which is described in the following section.

4.1.1 Greedy Bit Allocation Algorithm

The greedy algorithm is a simple and intuitive method for achieving integer-constrained bit allocation [61]. The algorithm is performed iteratively, ensuring an integer assignment of bits to each quantizer. At each iteration, one bit is allocated to the quantizer for which the decrease in a distortion measure is largest. The algorithm is *greedy* since bit allocations are optimized per iteration rather than considering the final distortion. Segall argued that the greedy algorithm is optimal when the individual distortion functions are convex and monotonically decrease with the number of allocated bits and the total distortion is the sum of the individual distortions [62]. The algorithm, as described in [61], is summarized below.

Assume that B bits are available for N quantizers. Let $W_i(b)$ represent the distortion function associated with the ith quantizer having b bits. Additionally, let $b_i(m)$ represent the number of bits allocated to the ith quantizer after m iterations.

Step 0 – Initialize the number of bits assigned to each quantizer to zero such that $b_i(0) = 0$ for $i = 1 \ldots N$.

Step 1 – Find the index j such that:

$$j = \arg\max_i \Big(W_i\big(b_i(m-1)\big) - W_i\big(b_i(m)\big) \Big). \tag{4.1}$$

Step 2 – Set $b_j(m+1) = b_j(m) + 1$ and $b_i(m+1) = b_i(m)$ for all $i \neq j$.

Step 3 – Set $m = m + 1$. If $m \leq B$, return to step 1.

In another form of the greedy algorithm, a bit is assigned where the distortion measure is largest, which is not necessarily the band where the highest decrease in distortion occurs.

In this case, Equation (4.1) should be replaced by:

$$j = \arg\max_i W_i\big(b_i(m)\big). \tag{4.2}$$

4.1.2 Noise Power Update

The greedy algorithm, as described above, requires an update of the noise power at each iteration. Performing the quantization and computing the error in the representation of the signal is a standard method for updating the noise power. However, this approach can result in a large processing delay, particularly for complex quantizers.

As an alternative, the average noise reduction resulting from the allocation of a bit can be used to update the noise power. As described in [2], a large set of data vectors is used off-line to determine the average distortion produced by each quantizer. The rate-distortion relationship represents the average distortion as a function of the number of allocated bits per quantizer. The noise power is then updated using the average noise reduction, which is obtained as a function of the number of bits from the rate-distortion relationship. This method results in a lower computational complexity at the expense of a lower accuracy in noise update. An example of rate-distortion data is provided later in Section 5.3.4.

The average noise reduction is particularly appealing for the first implementation of the greedy algorithm, where allocation is dependent on the noise reduction rather than the initial noise power. The noise reduction must be computed for each quantizer, resulting in as many quantizations as there are quantizers per iteration. In this case, it is advantageous to use the expected decrease in noise power. A modified version of the greedy algorithm that employs the reduction in noise is presented later in Section 4.3.

4.2 Noise Energy-based Bit Allocation

This section describes bit allocation schemes for which the distortion function is based on the quantization noise energy per coder sub-band. Coder sub-bands are treated individually and noise targets are considered to be independent of each other. More specifically, noise perception in one sub-band is independent of quantization noise targets in neighbouring sub-bands. An exact application of the greedy algorithm is performed to achieve bit allocation. Three different distortion measures are presented.

4.2.1 Absolute Noise Energy

The distortion measure in this approach is the absolute noise energy. This is a purely objective measure that does not require any psychoacoustic processing (*i.e.*, masking threshold computation). The noise energy is initialized to the normalized signal energy prior to the first bit allocation iteration. At each iteration, a bit is allocated to the band which has the highest noise energy. Once a bit is assigned to a band, the quantization noise energy for that band is reduced.

4.2.2 Noise-to-mask ratio

In this approach, the ratio of the quantization noise to the masking threshold is used as a distortion criteria. It is assumed that noise is inaudible in one sub-band if it lies below the masking threshold. This method attempts to distribute equally the perception of noise in all coder sub-bands.

The *Noise-to-Mask Ratio* (NMR) is initially set to the *Signal-to-Noise Ratio* (SMR). More specifically, the quantization noise is equal to the signal energy in each band prior to bit assignment. Information bits are assigned at every iteration to the band having the largest updated NMR. The allocation of a bit to a band reduces the NMR associated with that band. Once the NMR in a band reaches zero, the noise energy lies below the masking threshold, at which point the noise is considered inaudible. No bits are assigned to bands for which the NMR is negative, unless excess bits are available.

4.2.3 Audible Noise Energy

The distortion criterion in this approach is the absolute level of noise above the masking threshold. This differs from the previous approach where the relative noise level with respect to the masking threshold was employed, *i.e.*, the ratio between the noise level and the masking threshold. In this case, the non-logarithmic difference between the noise level and the masking threshold within each coder sub-band is considered. As a result, information bits are distributed such that the overall amount of audible noise is minimized.

4.2.4 Performance

Najafzadeh and Kabal performed informal listening tests to evaluate subjective preferences of the noise energy-based methods [63]. A narrow-band low-rate audio coder was employed to compare the bit allocation schemes.

As expected, the quantized outputs resulting from the absolute energy approach had a higher SNR than that of any other method. Since audio signals typically exhibit more energy at low frequencies, more bits were allocated to this region. Higher frequency components were poorly represented using this criterion for bit allocation. This method yielded the least pleasant of reconstructed signals for listening subjects. The bit allocation schemes that employ a masking threshold provided similar perceptual quality. However, listening subjects favoured the NMR-based approach as it caused less high frequency distortion than the audible noise method. As a result of Najafzadeh's findings, only the NMR-based approach will be considered herein within the class of noise energy-based methods.

4.3 Noise Excitation-based Bit Allocation

Target signals from different coder sub-bands were considered to be independent in the previous section. However, owing to the properties of the ear, the quantization noise in one band is generally perceived in neighbouring bands. Noise targets excite hair cells associated with adjacent bands similarly to masker signals. The combined quantization noise targets exhibit a complex excitation pattern.

According to the excitation pattern model of masking, the threshold for noise audibility is given in the spread domain, *i.e.*, following convolution with the spreading function. However, bit allocation is performed in the non-spread domain. Johnston identified the need to despread the predicted masking threshold in order to apply it to noise shaping [55]. Many researchers have proposed models that have neglected or simplified this step (*e.g.*, renormalization of the masking threshold). The current work suggests employing the excitation of the quantization noise in the bit allocation rather than attempting the deconvolution of the masking threshold. In other words, the bit allocation is performed by taking into account the spread of the excitation of the quantization noise.

Let $E_S(z)$ and $E_R(z)$ represent respectively the excitation patterns of the clean and reconstructed audio signals. The excitation pattern model of masking states that the noise

is inaudible if:

$$|E_S(z) - E_R(z)| < M_{Th}(z), \tag{4.3}$$

where M_{Th} represents the masking threshold. Assuming power domain addition of the excitation patterns produced by the clean signal and the quantization error, Equation (4.3) can be rewritten as:

$$E_Q(z) < M_{TH}(z), \tag{4.4}$$

where $E_Q(z)$ represents the excitation pattern produced by the quantization noise. As such, the objective of noise excitation-based bit allocation is to minimize $E_Q(z)$ with respect to $M_{TH}(z)$.

A difficulty emerges with respect to the application of the greedy allocation algorithm when considering noise excitation. The distortion measure is not localized to coder sub-bands. More specifically, assigning an information bit to one band affects the noise excitation in other bands. The greedy bit allocation algorithm must be adapted accordingly. Rather than optimizing the distortion on per band basis, the objective is to obtain the maximum auditory gain at each allocation iteration.

4.3.1 Previous Noise-Excitation-Based Methods

Perreau-Guimaraes *et al.* proposed a bit allocation scheme that is based on the excitation produced by the quantization noise [64]. Firstly, they argued that bit allocation should be performed in the Bark frequency scale. Information bits are allocated to basilar sub-bands, where basilar sub-bands are equivalent to critical bands. Once the allocation is complete, bits are distributed to the coder sub-bands contained within each basilar sub-band.

Perreau-Guimaraes suggested that the majority of perceived noise results from the basilar sub-band where the ratio of noise excitation to masking threshold is highest. Accordingly, the objective at each iteration of the algorithm is the reduction of noise excitation in band i_0 for which:

$$i_0 = \arg\max_i \left(\frac{E_Q(i)}{M_{TH}(i)} \right). \tag{4.5}$$

As a result of spectral spreading, the excitation $E_Q(i_0)$ is influenced by an array of bands lying in the vicinity of i_0. The allocation of a bit to an adjacent band may perhaps further decrease $E_Q(i_0)$ compared to the allocation to i_0. For instance, the addition of a bit to i_0 scarcely reduces $E_Q(i_0)$ when the primary contributor to the excitation is a neighbouring

band. The band to which a bit is allocated is selected as:

$$j_0 = \arg\max_j \left(SF(i_0, j)\, S_Q(j)\right), \tag{4.6}$$

where $SF(i_0, j)$ represents the spreading of band j into band i_0 and $S_Q(j)$ represents the noise power in band j. Due to the rapid decrease away from the peak of the spreading function, it is suggested to only consider a restricted group of candidate bands that lie near i_0. Once selected, the noise power in band j_0 is updated along with its contribution to the overall noise excitation.

Although more accurate than noise energy-based approaches, this bit allocation scheme optimizes a single band at each iteration. Bit allocation is not performed based on the overall auditory benefit. A novel bit allocation scheme that considers overall noise audibility rather than localized noise audibility is presented in the following section.

4.3.2 A Novel Bit Allocation Scheme

In accordance with Perreau-Guimaraes, bit allocation should be performed to basilar subbands. This follows from the discussion in Section 2.11 where it was suggested that noise targets should be defined over bands of unit-Bark width. Furthermore, the auditory masking model proposed in Section 3.6 provides masking threshold estimates on a critical band basis.

At each iteration of the greedy algorithm, a bit is assigned to the basilar band which yields the highest overall auditory gain. A clear definition of auditory gain is required along with the corresponding distortion criteria that will be optimized in the allocation. Firstly, the *Noise-Excitation-to-Mask Ratio*, or NEMR, is formally introduced within the current work,

$$\text{NEMR}(z) = \frac{E_Q(i)}{M_{TH}(i)}. \tag{4.7}$$

The NEMRs in each band are amalgamated to yield a distortion criteria. The current work proposes the sum of all the NEMRs above 1 as the perceptual distortion, D_P,

$$D_P = \sum_{\text{NEMR}(z)>1} \text{NEMR}(z). \tag{4.8}$$

More specifically, only bands for which the noise excitation is audible are considered in the

perceptual distortion. The overall auditory gain resulting from the allocation of a bit is defined as the reduction in perceptual distortion. In addition to the perceptual distortion, the total distortion, D_T, is defined as the sum of all NEMRs,

$$D_T = \sum_{z=1}^{Z} \text{NEMR}(z). \tag{4.9}$$

The perceptual distortion resulting from each possible allocation must be evaluated so as to determine which band receives the additional bit. The following steps are performed independently for each band at each iteration:

Step 1 – Estimate the reduction in noise power resulting from the allocation of a bit to the ith band. The noise reduction is obtained using the rate-distortion curve associated with the ith band.

Step 2 – Compute the reduction in noise excitation produced by the ith band. This is performed by multiplying the reduction in noise excitation by the spreading function.

Step 3 – Subtract the reduction in noise excitation to the overall noise excitation estimate. In order to simplify computational complexity, it is assumed that the excitation produced by different noise targets is added linearly.

Step 4 – Compute the perceptual distortion and the total distortion resulting from the allocation of a bit to the ith band.

The information bit is allocated to the band yielding the lowest perceptual distortion. In the case where bits are abundant, it is possible that the updated perceptual distortion be null. The remaining bits are allocated using the total distortion criteria.

4.4 Chapter Summary

This chapter has presented traditional bit allocation methods that are based on the quantization noise power. From this family of algorithms, the NMR-based approach was deemed most appropriate and has been selected for the remainder of this thesis. Secondly, noise-excitation based bit allocation schemes were presented. A novel algorithm that incorporates psychoacoustic processing along with a new optimization criterion was proposed, whereby each iteration of the algorithm attempts to minimize overall audible excitation produced by quantization noise targets.

Chapter 5

Performance Evaluation

This chapter presents the evaluation of the auditory masking model and the bit allocation scheme that have been proposed. Firstly, a discussion is provided concerning the validity of objective and subjective performance metrics for the evaluation of the current work. Secondly, the experimental setup for subjective evaluation is presented. Performance results for the adaptive bit allocation scheme are provided, followed by results for the auditory masking model.

5.1 Objective Evaluation

Various objective metrics have been proposed for the quality evaluation of audio signals. Among these, *Signal-to-Noise Ratio* and *Segmental Signal-to-noise Ratio* are the most prevalent. Such metrics measure distortion by simply comparing the original and coded signals, without considering human perception. Accordingly, they are inappropriate for the evaluation of perceptual-based coding algorithms. Metrics based exclusively on noise energy favour coding paradigms that minimize overall noise rather than audible noise. Discrepancies in quality evaluation resulting from the use of SNR are generally larger at low coding rates, where the amount of coding noise is significant.

Perceptual-based models for quality evaluation have also been proposed, as discussed in Section 1.4.2. Such methods are designed to represent the quality evaluation of human listeners. However, these metrics favour the performance of coding schemes that employ similar perceptual models, rendering them inappropriate for the performance evaluation of dissimilar auditory masking models. Additionally, the auditory models employed by

these methods do not accurately represent certain perceptual characteristics, as suggested in Section 3.5. In essence, objective quality measures are inappropriate for the evaluation of the current work.

5.2 Subjective Evaluation

Subjective evaluation is the ultimate method for assessing the quality of audio signals. Sound files are presented to listeners that grade and/or compare them according to perceived quality. When presented speech signals, listeners generally seek out intelligibility and naturalness. Overall audible distortion is a common target for the evaluation of music and general audio signals. Although highly indicative of audio quality, subjective listening tests are expensive with regard to time and resources. Regardless, subjective performance evaluation is the ideal candidate for assessing the performance of perceptual-based coding algorithms.

Several informal listening tests were performed in order to validate the concepts introduced within the current work. Four untrained listeners were presented with the original sound file, followed by sound files that were processed using different algorithms. Listeners were asked to compare and rank the processed files from best to worst, in terms of perceptual quality. When undecided, listeners were asked to tie the files that were equally preferred. The generation of experimental data is described in the following section.

5.3 Experimental Data

The experimental data employed for the performance evaluation of the proposed algorithms was generated using four different audio segments. The selected signals included a female speech segment, a male speech segment and two music segments.

The original sound files were processed using the test bed that is depicted in Figure 5.1. All original segments were sampled at 16 kHz and represented using *Pulse Code Modulation* (PCM) with 16 bits per sample. The test bed implements many features of a perceptual audio coder. More specifically, it is based on the perceptual audio coders developed by Johnston [55] and Najafzadeh-Azghandi [2]. While it does not implement the entire audio coding process, the test bed is sufficient for the evaluation of differences between the algorithms under comparison. The development of a complete audio coder is out of the scope

of the current work.

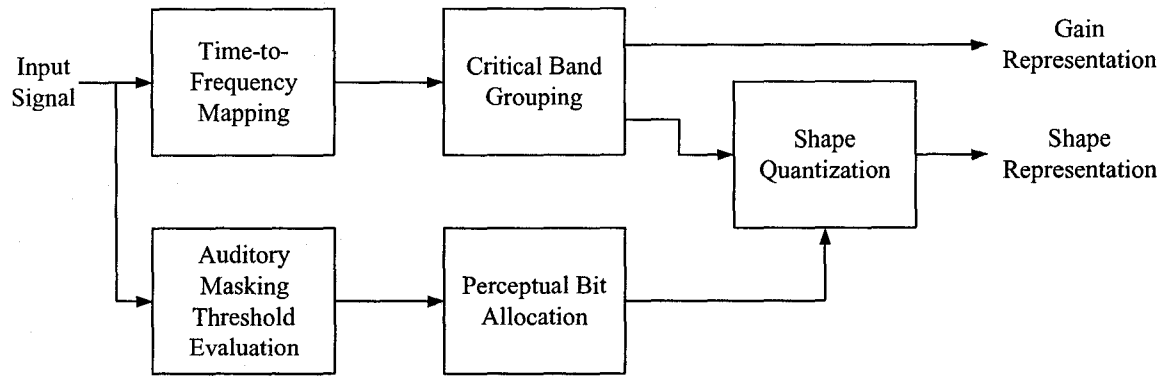


Fig. 5.1 Functional block diagram of the proposed test bed.

5.3.1 Time-to-Frequency Mapping and Critical Band Grouping

The time-to-frequency mapping is achieved by first segmenting the input signal into blocks of 32 ms length, corresponding to 512 samples at a rate of 16 kHz, followed by the multiplication with a Hanning window. A 1/16 th frame overlap (i.e. 32 samples) is introduced in order to reduce block-edge distortion effects incurred in the reconstruction. The reconstruction of the audio signal is achieved using the overlap-add method, as described in [65]. A 512-point FFT is performed on the windowed input signal, yielding the desired frequency response.

Each coder sub-band includes the frequency components lying within a critical band, *i.e.*, there exists a one-to-one correspondence between coder sub-bands and critical bands. This is based on the auditory models under investigation, which report masking thresholds on a critical band basis. The frequency-to-Bark conversion is calculated according to the mapping given by Schroeder [23], which is expressed in Equation (2.6). The output of the critical band grouping stage is represented using a gain scalar and shape vector per critical band, as described in Section 5.3.3 and Section 5.3.4.

5.3.2 Auditory Masking Model and Perceptual Bit Allocation

The auditory masking threshold evaluation and the perceptual bit allocation stages represent the algorithms under test. The former computes the masking threshold according to

the desired model, while the latter determines the number of bits allocated per signal component. The bit allocation is performed using the average rate-distortion curve that was described in Section 4.1.2 for all of the algorithms under study. The rate-distortion curve, which depends on the quantization process, is explained in more detail in Section 5.3.4.

5.3.3 Gain Representation

A scalar gain factor is extracted from each coder sub-band. The gain is computed as the average energy of all components within the sub-band. In a real audio coder, the gain factors are quantized and transmitted to the decoder. In the test bed, the gain factors are left unquantized. This simplification is appropriate for the comparison of auditory masking models and bit allocation schemes, which are only applied to the shape representation.

5.3.4 Shape Quantization

A shape vector represents the contour of the spectral components within a coder sub-band. More specifically, FFT bins are grouped within a sub-band and normalized by the associated gain factor to generate the shape vector. The assignment resulting from the bit allocation stage is applied to the representation of the shape vector.

The objective of this stage is the introduction of quantization noise into the audio signal spectrum. The quantization of shape vector components is achieved using scalar product quantizers. Real and imaginary parts of complex components are quantized as separate elements. Although vector quantizers are generally more efficient in representing signals [61], their design is too complex for the current purpose. Sub-optimal scalar product quantizers are used, as the aim of the subjective experiment involves the evaluation of relative performance rather than absolute performance.

Following the gain factor normalization, individual signal components have a "bell-curve" distribution with zero mean and unit variance. Scalar quantizers, optimized for a Gaussian source, were designed using the Lloyd-Max algorithm [61]. Provided that the amount of noise introduced by the quantization process is known at the bit allocation stage, the approximation of a Gaussian distribution should not jeopardize the comparison between the auditory models and bit allocation schemes under study.

The amount of quantization noise as a function of the number of allocated bits per coder sub-band is conveyed to the bit allocation stage through the rate-distortion relationship,

as described in Section 4.1.2. The rate-distortion relationship represents the average noise energy as a function of the number of allocated bits. Sample rate-distortion curves for four coder sub-bands are illustrated in Figure 5.2, given the quantization scheme described above. The rate-distortion curves can be accurately approximated using a linear function.

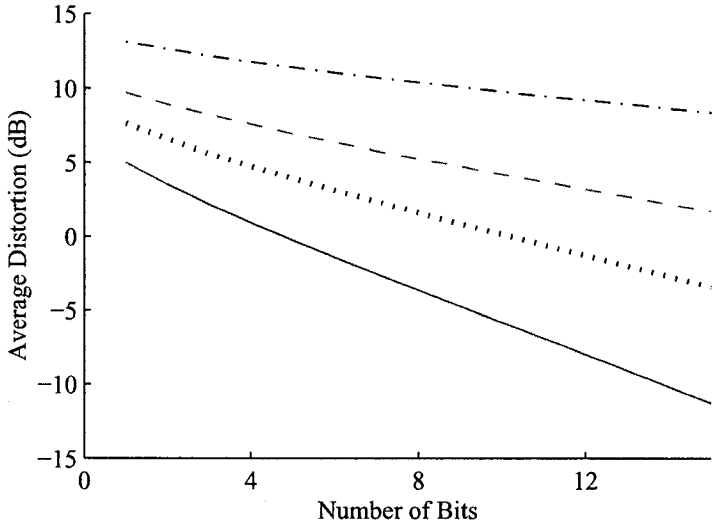


Fig. 5.2 Average distortion as a function of the number of bits assigned to a coder sub-band. From bottom to top, the curves represent the rate-distortion relationship for the 1^{st} sub-band (5 components), the 5^{th} sub-band (8 components), the 10^{th} sub-band (12 components) and the 15^{th} sub-band (24 components).

The estimated slopes are given in Table 5.1 along with the number of components within each of the 23 coder sub-bands. As the number of components within a band increases, the gain resulting from the allocation of a bit is smaller. As such, greater reductions in quantization noise energy are obtained for lower coder sub-bands.

Inputs to the shape quantization stage are the unquantized shape vectors along with the number of allocated bits per component within each vector. The quantization process is modelled by adding noise to the individual components, according to the number of allocated bits. Components for which bits were not allocated in the bit allocation stage are represented by the mean value of the signal, which corresponds to zero. The following sections present subjective performance results for the adaptive bit allocation scheme and auditory model.

Table 5.1 Rate-distortion slopes for 23 coder sub-bands.

Coder Sub-band	Centre Frequency (Hz)	Vector Size (Number of Components)	Slope (dB/bit)
1	47	5	−1.12
2	140	8	−0.72
3	237	6	−0.95
4	339	6	−0.95
5	447	8	−0.72
6	565	8	−0.72
7	694	8	−0.72
8	838	10	−0.58
9	998	10	−0.58
10	1179	12	−0.48
11	1384	14	−0.42
12	1617	16	−0.37
13	1883	18	−0.33
14	2189	22	−0.27
15	2538	24	−0.25
16	2940	28	−0.21
17	3402	32	−0.18
18	3932	36	−0.17
19	4544	42	−0.14
20	5249	48	−0.12
21	6060	56	−0.11
22	6996	64	−0.09
23	7762	31	−0.20

5.4 Evaluation of the Adaptive Bit Allocation Scheme

The novel bit allocation algorithm was conceived for use with the auditory model that is described in Section 3.6. However, it is recognized that the algorithm is applicable with any auditory model that is based on the excitation pattern model of masking (Section 2.6). Accordingly, the first phase in the validation of the current work involves the evaluation of the proposed adaptive bit allocation scheme, combined with an impartial reference auditory model. Johnston's auditory masking model, which is described in Section 3.1, was used as the reference model. Johnston's model is well recognized in the field of audio coding as it

has been used to determine the bit allocation for perceptual-based audio coders [55].

The perceptual quality of the new bit allocation algorithm was compared to the NMR-based approach and the method proposed by Perreau-Guimaraes *et al.*, described in Section 4.2.2 and Section 4.3.1 respectively. A series of listening tests were performed for the three bit allocation algorithms. When combined with the noise excitation-based methods (*i.e.*, the novel algorithm and Perreau-Guimaraes' algorithm), Johnston's masking model was modified by removing the renormalization step that is suggested in the later stage of the masking threshold computation. The modification was not required when combined with the NMR-based approach in the test bed.

The four original sound files were processed using the three different bit assignment algorithms, combined with Johnston's auditory model, at coding rates of 125, 250, 500 and 1000 bits per frame for shape quantization. The selected range of bit rates encompassed scenarios having significant audible distortion (125 bits per frame) and minute audible distortion (1000 bits per frame). As a result, 16 different auditory tests were performed per listener, *i.e.*, one test per original sound file per coding rate. Each auditory test commenced with the playback of the original uncoded file, followed by the three segments that had been processed with the bit allocation algorithms under study. As previously mentioned, listeners were asked to rank the files from best to worst with respect to perceptual quality. When undecided, listeners were invited to tie the two or three segments having similar quality.

All four listeners unanimously ranked the novel bit allocation algorithm in first place for the lowest three bit rates with all sound files. In second place was the NMR-based approach, followed by the method proposed by Perreau-Guimaraes in the last position. The differences in perceived quality between the three algorithms were greater for the lowest bit rate, and decreased with increasing coding rate. For the 1000 bits per frame test case, all three methods yielded similar quality, with a slight bias towards the novel algorithm.

The poor quality resulting from Perreau-Guimaraes' method operating at lower bit rates was of particular interest. An investigation showed that, at such rates, the majority of bits were allocated to higher frequency critical bands, while low frequency bands received few or none. Johnston's auditory model computes the masking threshold by subtracting a frequency dependent offset from the spread Bark power spectrum. The offset generally increases with frequency, depending on the overall signal tonality. In the bit allocation stage, the noise excitation is initialized to the spread Bark power spectrum of the input

signal prior to the first iteration of the greedy bit algorithm. The initial NEMR is equivalent to the frequency dependent offset that is employed in Johnston's model, which is larger at high frequencies. Information bits are allocated to high frequency bands since Perreau-Guimaraes' method targets the band having the highest NEMR at each iteration. Moreover, small reductions in noise energy are obtained for high frequency allocations (as shown in Table 5.1), requiring multiple bits to achieve a considerable decrease in local NEMR. The phenomenon was not observed when operating at the highest coding rate, where the number of bits was sufficient to reduce high frequency NEMRs and encode other perceptually relevant bands. This observation clearly demonstrates the danger involved with considering the noise excitation of a single band at each iteration of the bit allocation. Instead, the overall noise excitation should be considered at each iteration, as discussed in Section 4.3.2.

In summary, the novel adaptive bit allocation algorithm outperformed the NMR-based approach and Perreau-Guimaraes' method with regard to perceived quality for all listeners. A significant gain was obtained when applying the novel algorithm, particularly at low coding rates. Accordingly, the novel algorithm will be employed as the reference bit allocation scheme for the evaluation of the auditory masking model, which is presented in the following section.

5.5 Evaluation of the Auditory Masking Model

The second phase required for the validation of the current work is the evaluation of the new auditory model that was proposed in Section 3.6. A subjective experiment was conducted where perceptual quality using the new auditory model was compared to that of Johnston's auditory model and the PEAQ model, described in Section 3.1 and Section 3.4 respectively. The novel bit allocation algorithm (Section 4.3.2) was selected as the reference bit allocation scheme in the test bed, as suggested in the previous section.

The four original sound files were processed using the three auditory masking models under study, combined with the new bit allocation algorithm in the test bed, at coding rates of 250 and 500 bits per frame for shape quantization. The subjective experiment was performed in a similar manner to that described in the previous section for the evaluation of the bit allocation scheme. Subjective results were more ambiguous for this set of listening tests than for the previous experiment. The PEAQ model was generally preferred over the other two auditory models when applied to the female and male speech segments at

both coding rates. The novel auditory model warranted the second position, followed by Johnston's model. Nevertheless, differences in perceived quality between the latter two models were minute. In the case of music segments, all three auditory models generated similar quality reconstructed signals. Some listeners perceived slightly inferior quality using the PEAQ model over the other two.

The observed results demonstrated that the proposed auditory model performed similarly to the other models. The PEAQ masking model yielded the best overall performance. The novel auditory model and the PEAQ model differ mainly in the way that the spectrum of the input signal is decomposed into masking components. The former clearly distinguishes between tonal and noise maskers, whereas the PEAQ model performs a high resolution decomposition of the spectrum without identifying the nature of maskers. Additionally, the new auditory model considers transient masking effects by tracking the temporal evolution of tonal maskers, while the PEAQ model individually accounts for simultaneous and forward masking.

Given the limited number of audio files that were used (2 speech segments and 2 music segments), it is difficult to draw any meaningful conclusions from the experiment. Further subjective testing is required, including additional sound files and a larger number of listeners. Extensive testing was not performed in light of the uncertainties related to auditory masking. As discussed in Section 2.12, insufficient data is available regarding auditory masking for its application towards complex audio signals. A more in depth understanding of masking effects is required, which is out of the scope of the current work.

5.6 Chapter Summary

Performance evaluations of the proposed adaptive bit allocation scheme and auditory model were presented in this chapter. In both cases, results from informal subjective listening tests were provided by comparison to other well-recognized paradigms. A significant gain was observed when considering the adaptive bit allocation algorithm, while average performance was observed for the novel auditory model.

Chapter 6

Conclusion

This thesis has focussed on the application of human perception to low-rate audio coding. The phenomenon of auditory masking was studied along with its use in perceptual coding. A review of reported psychoacoustic experiments pertaining to auditory masking and the analysis of existing auditory models allowed for the development of a new auditory masking model. The proposed model predicts masking from a complex input signal, while considering the temporal course of masking components as opposed to individually accounting for simultaneous and forward masking effects. Moreover, a correct application of the excitation pattern model of masking was achieved by taking into account the spread of excitation of the quantization noise in the allocation of information bits. A novel bit allocation scheme was proposed that solves the deconvolution problem, which is applicable to any auditory masking model that is based on the excitation pattern model of masking.

Through various subjective experiments, the proposed bit allocation algorithm considerably outperformed other methods with respect to perceptual quality. Improvements were particularly noticeable for low coding rates where quantization noise is audible. As for the proposed auditory model, performance more or less similar to other auditory masking models was observed through limited tests. It was concluded that insufficient knowledge relating to auditory masking is available for accurate estimation of masking thresholds produced by complex audio signals, such as speech or music.

This chapter provides a summary of the current work and presents directions for future research in the field of perceptual audio coding.

6.1 Thesis Summary

Firstly, the motivation for low-rate audio compression was presented in Chapter 1. The basic concept of auditory masking was introduced and its application to audio coding was described through the presentation of a generic perceptual audio coder in Section 1.2. The challenges involved in low-rate audio coding were discussed and the need for a more accurate auditory masking model was rationalized. At low coding rates, the amount of distortion that is introduced by the quantization process is comparable to the amount of masking produced by the input signal, giving reason for a more accurate prediction of masking. Additionally, an overview of other signal processing applications involving auditory masking was presented in Section 1.4.

Chapter 2 introduced the basic theory of sound levels, along with a description of the physiology of the human auditory system. The concept of critical bands was presented in order to explain the frequency resolution of the ear. Subsequently, auditory masking was thoroughly detailed in Section 2.5, highlighting the differences between simultaneous, backward and forward masking.

Section 2.6 presented the excitation pattern model of masking, which forms the basis for the prediction of masking in the proposed work. The model states that a target signal is inaudible if its presence does not change the output of any auditory filter by an amount greater than 1 dB. Assuming power domain addition, a target signal should remain inaudible if the excitation it produces is 1 dB below the excitation produced by the masker at any given frequency.

The following sections in Chapter 2 reported various characteristics of auditory masking. Namely, the temporal course of masking (or transient masking) was described along with the additivity of masking with multiple maskers. It was suggested that noise maskers and noise targets spanning wide frequency ranges should be integrated over critical bands when predicting masking thresholds. Finally, the insufficiency of reported experimental results relating to auditory masking was discussed. The lack of data results in inaccurate modelling of the complex interactions that exist between multiple maskers and targets.

Chapter 3 commenced with the presentation of four well-known auditory models that predict the amount of masking produced by a complex audio signal. The models under study spanned various levels of complexity, ranging from the simple Johnston model (Section 3.1) to the more detailed PEAQ model (Section 3.4). Following their description, a

discussion identifying the shortcomings of these four models was presented in Section 3.5. Among other issues, the treatment of different types of masking and the correctness of the application of the excitation pattern model of masking were noted.

A novel auditory model that predicts masking was presented in Section 3.6. The proposed model includes the psychoacoustic findings that were presented in Chapter 2. Transient masking effects are modelled rather than individually representing simultaneous and forward masking effects. The temporal evolution of tonal maskers is tracked, influencing the amount of masking produced by tones.

The estimated masking threshold represents an upper bound for the excitation produced by the quantization noise. The need for a deconvolution of the masking threshold prior to its use for noise shaping has been acknowledged in previous work. However, this process was neglected or simplified in the four auditory models studied in Chapter 3. The current work proposed a solution to the deconvolution problem within the bit allocation scheme. Rather than attempting to despread the masking threshold, bit allocation is performed by taking into consideration the excitation produced by the quantization noise. The proposed bit allocation algorithm was presented in Chapter 4

Chapter 4 examined various bit allocation strategies that shape quantization noise in frequency according to a perceptual criterion. Noise energy-based methods, such as the SMR approach, were first described, followed by noise excitation-based methods. The NEMR was formally introduced in this thesis, along with a new distortion criterion for the allocation of information bits. The new criterion minimizes the overall audible distortion at each iteration of the bit allocation, rather than minimizing local distortion.

The performance assessment of the proposed work was presented in Chapter 5. Firstly, arguments supporting the inadequacy of objective performance metrics for the evaluation of perceptual models were discussed. The need for subjective listening experiments was justified, followed by a description of the experimental setup in Section 5.3. The evaluation of the proposed bit allocation algorithm was performed in Section 5.4 by comparison with the NMR-based approach and Perreau-Guimaraes' method. Although few listeners were used, a unanimous preference for reconstructed signals that had been coded with the new algorithm was observed. A significant improvement in audio quality was perceived, particularly at low coding rates.

The performance evaluation of the proposed auditory masking model was presented in Section 5.5. Little or no performance improvements were observed when compared to

Johnston's model and the PEAQ model. This concurs with the idea that insufficient data is available regarding auditory masking for its application with complex audio signals, as discussed in Chapter 2.

6.2 Future Research Directions

This section provides guidance for future research in the field of perceptual audio coding. The predominant issue that remains to be addressed is the insufficiency of psychoacoustic results and hence, the accuracy of auditory masking models. Additionally, the application of the proposed bit allocation algorithm to a well-recognized audio coder should be studied.

The majority of reported psychoacoustic experiments that relate to auditory masking have been performed with the aim of understanding human perception. These experiments generally isolate a specific aspect of auditory perception, enabling the use of relatively simple sound stimuli. Further auditory masking experiments are required, which aim at understanding the masking of broadband quantization noise targets by complex audio signals.

Interactions between masking components in complex audio signals are not well understood. For instance, the decomposition of the spectrum of an audio signal into multiple masking components is oversimplified. Current auditory models generally segment the spectrum of the input signal into a discrete set of non-overlapping maskers, on a critical band scale. However, the human ear is composed of a continuum of overlapping auditory filters. The means by which masking contributions should be combined in such a high resolution model remains unclear.

Similarly, interactions between maskers and noise targets are not considered in current auditory models. In audio coding, the quantization noise targets occupy the entire bandwidth of the masker signal. The masking threshold is computed without consideration for the composition of target signals. This results from the application of the excitation pattern model of masking, where the overall level of the signal is considered as the primary detection cue. However, quantization noise targets might contribute to the masking threshold by altering detection cues when presented to the ear. Additional psychoacoustic research must be performed with sound stimuli that are relevant to audio coding.

The current work considered the temporal evolution of tonal maskers when evaluating their masking effects. Similarly, tracking the temporal course of noise maskers should be

investigated. Transient masking effects (*e.g.*, the overshoot and undershoot effects) of noise maskers should be evaluated and a method for tracking the evolution of noise components in an audio signal should be developed.

The amount of masking produced by a signal may noticeably differ from one listener to another. Reported masking patterns are generally obtained by averaging masking thresholds observed by different listeners. The auditory model should be extended to represent the variability in masking thresholds for different listeners.

Reductions in masking levels have been observed when listening with two ears rather than listening with one [1]. Binaural processing involves the combination and/or comparison of sounds received by one ear with those received by the other ear. Binaural masking effects have received little attention in auditory masking models thus far. Moreover, such effects are difficult to model as they depend on the physical location of the listener with respect to the sound source.

Finally, the proposed bit allocation algorithm should be integrated within a low rate audio coder. The current work did not consider details such as a method for informing the decoder of the bit allocation. For instance, certain audio coders explicitly inform the decoder of the bit allocation by means of side information. In this case, the amount of side information is often overly high for low-rate coders. Other coding schemes dedicate just enough side information such that the decoder can autonomously determine the bit allocation. However, this often results in a sub-optimal assignment since the encoder must also allocate bits according to the limited side information. The integration of the proposed bit allocation algorithm within various coding schemes should be studied.

This thesis has studied the work of psychoacoustic researchers and audio coding specialists, and applied human perception to low rate audio coding. Hopefully, the current work can contribute to the improvement of perceptual quality in low rate audio coders.

References

[1] B. C. J. Moore, *An Introduction to the Psychology of Hearing.* Academic Press, fourth ed., 1997.

[2] H. Najafzadeh-Azghandi, *Perceptual Coding of Narrowband Audio Signals.* PhD thesis, McGill University, Montreal, Canada, Apr. 2000.

[3] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, pp. 451–513, Apr. 2000.

[4] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ – the ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, pp. 2–29, Jan. 2000.

[5] S. Haykin, *Communication Systems.* John Wiley and Sons Inc., third ed., 1994.

[6] International Standards Organization, *Coding of Moving Pictures and Associatied Audio*, Apr. 1993. ISO/IEC JTC/SC29/WG 11.

[7] International Standards Organization, *Generic Coding of Moving Pictures and Associated Audio Information (Part 7)-Advanced Audio Coding (AAC)*, 1996. ISO/IEC DIS 13818-7.

[8] Advanced Television Systems Committee (ATSC), *Digital Audio Compression Standard (AC-3)*, Dec. 1995.

[9] D. Tsoukalas, J. H. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech and Audio Processing*, vol. 5, pp. 497–514, Nov. 1997.

[10] G. A. Souloudre, *Adaptive Methods for Removing Camera Noise from Film Soundtracks.* PhD thesis, McGill University, Montreal, Canada, Nov. 1998.

[11] J. Thiemann, "Acoustic noise suppression for speech signals using auditory masking effects," Master's thesis, McGill University, Montreal, Canada, May 2001.

[12] M. Klein, "Signal subspace speech enhancement with perceptual post-filtering," Master's thesis, McGill University, Montreal, Canada, Jan. 2002.

[13] International Telecommunication Union, *Method for Objective Measurements of Perceived Audio Quality*, July 1999. ITU-R Recommendation BS.1387.

[14] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models.* Springer-Verlag, second ed., 1999.

[15] D. O'Shaughnessy, *Speech Communications: Human and Machine.* IEEE Press, second ed., 2000.

[16] C. Giguere and P. Woodland, "A computation model of the auditory periphery for speech and hearing science," *J. Acoust. Soc. Am.*, vol. 95, pp. 331–349, Jan. 1994.

[17] "Anatomy and function of the ear." [online] `http://depts.washington.edu/otoweb/ear_anatomy.htm`.

[18] S. Puria, W. Peake, and J. Rosowski, "Sound pressure measurements in the cochlear vestibule of human-cadaver ears," *J. Acoust. Soc. Am.*, vol. 101, pp. 2754–2770, May 1997.

[19] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Am.*, vol. 71, pp. 679–688, Mar. 1982.

[20] E. Terhardt, "Calculating virtual pitch," *Hear. Res.*, vol. 1, pp. 155–182, 1979.

[21] H. Fletcher, "Auditory patterns," *Revs. Modern Phys.*, vol. 12, pp. 47–66, Jan. 1940.

[22] J. Tobias, *Foundations of Modern Auditory Theory.* Academic Press, 1970.

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

[24] B. R. Glasberg and B. C. J. Moore, "Derviation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.

[25] T. Thiede, *Perceptual Audio Quality Assessment Using a Non-Linear Filter Bank.* PhD thesis, Technical University of Berlin, Berlin, Germany, Apr. 1990.

[26] B. C. J. Moore, ed., *Hearing.* Academic Press, second ed., 1995.

[27] B. C. J. Moore, "Masking in the human auditory system," in *Collected Papers on Digital Audio Bit-Rate Reduction* (N. Gilchrist and G. C., eds.), Audio Engineering Society, 1996.

[28] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, Sept. 1983.

[29] R. N. J. Veldhuis, "Bit rates in audio coding," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 86–96, Jan. 1992.

[30] B. R. Glasberg and B. C. J. Moore, "Growth-of-masking functions for several types of maskers," *J. Acoust. Soc. Am.*, vol. 96, pp. 134–144, July 1994.

[31] J. P. Egan and H. W. Wake, "On the masking pattern of a simple auditory stimulus," *J. Acoust. Soc. Am.*, vol. 22, pp. 622–630, Sept. 1950.

[32] B. C. J. Moore, J. I. Alcantára, and T. Dau, "Masking patterns for sinusoidal and narrow-band noise maskers," *J. Acoust. Soc. Am.*, vol. 104, pp. 1023–1038, Aug. 1998.

[33] D. D. Greenwood, "Auditory masking and the critical band," *J. Acoust. Soc. Am.*, vol. 33, pp. 484–501, Apr. 1961.

[34] D. D. Greenwood, "Aural combination tone and auditory masking," *J. Acoust. Soc. Am.*, vol. 50, pp. 502–543, Aug. 1971.

[35] S. P. Bacon and N. F. Viemeister, "The temporal course of simultaneous tone-on-tone masking," *J. Acoust. Soc. Am.*, vol. 78, pp. 1231–1235, Oct. 1985.

[36] D. M. Green, "Masking with continuous and pulsed sinusoids," *J. Acoust. Soc. Am.*, vol. 46, pp. 939–946, Oct. 1969.

[37] B. Leshowitz and E. Cudahy, "Masking of continuous and gated sinusoids," *J. Acoust. Soc. Am.*, vol. 51, pp. 1921–1929, June 1972.

[38] S. P. Bacon and B. C. Moore, "Transient masking and the temporal course of simultaneous tone-on-tone masking," *J. Acoust. Soc. Am.*, vol. 81, pp. 1073–1077, Apr. 1987.

[39] R. L. Smith, "Adaptation, saturation, and physiological masking in single auditory-nerve fibers," *J. Acoust. Soc. Am.*, vol. 65, pp. 166–178, Jan. 1979.

[40] L. L. Elliot, "Development of auditory narrow-band frequency contours," *J. Acoust. Soc. Am.*, vol. 42, pp. 143–153, July 1967.

[41] G. Theile, M. Link, and G. Stoll, "Low bit rate coding of high quality audio signals," in *82nd AES Conv. Preprint 2431*, Mar. 1987.

[42] D. M. Green, "Additivity of masking," *J. Acoust. Soc. Am.*, vol. 41, pp. 1517–1525, Apr. 1967.

[43] R. A. Lutfi, "Additivity of simultaneous masking," *J. Acoust. Soc. Am.*, vol. 73, pp. 262–267, Jan. 1983.

[44] R. A. Lutfi, "A power-law transformation for predicting masking by sounds with complex spectra," *J. Acoust. Soc. Am.*, vol. 77, pp. 2128–2136, June 1985.

[45] J. A. Canahl, "Two versus four tone masking at 1000 hz," *J. Acoust. Soc. Am.*, vol. 50, pp. 471–474, Aug. 1971.

[46] D. A. Nelson, "Two tone masking and auditory critical bandwidth," *Audiology*, vol. 18, pp. 279–301, 1979.

[47] E. Zwicker, "Die von schmallbandgerauschen durch sinustone," *Acustica*, vol. 4, pp. 415–420, 1954.

[48] R. D. Patternson and I. Nimmo-Smith, "Off-frequency listening and auditory-filter asymmetry," *J. Acoust. Soc. Am.*, vol. 67, pp. 229–245, Jan. 1980.

[49] R. C. Bilger, "Additivity of different types of masking," *J. Acoust. Soc. Am.*, vol. 31, pp. 1107–1109, Aug. 1959.

[50] B. C. J. Moore, "Additivity of simultaneous masking, revisited," *J. Acoust. Soc. Am.*, vol. 78, pp. 488–494, Aug. 1985.

[51] L. E. Humes and W. Jesteadt, "Models of the additivity of masking," *J. Acoust. Soc. Am.*, vol. 85, pp. 1285–1294, Mar. 1989.

[52] L. E. Humes, B. Espinoza-Vara, and C. S. Watson, "Modeling sensorineural hearing loss. i. model and retrospective evaluation," *J. Acoust. Soc. Am.*, vol. 83, pp. 188–202, Jan. 1988.

[53] W. B. Kleijn and K. K. Paliwal, eds., *Speech Coding and Synthesis.* Elsevier, 1995.

[54] L. E. Humes and L. W. Lee, "Two experiments on the spectral boundary conditions for nonlinear additivity of simultaneous masking," *J. Acoust. Soc. Am.*, vol. 92, pp. 2598–2606, Nov. 1992.

[55] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.

[56] K. Brandenburg and J. D. Johnston, "Second generation perceptual audio coding: The hybrid coder," in *88th AES Conv. Preprint*, (Montreaux), pp. 1–10, Mar. 1990.

[57] S. Voran, "Observations of audiotry excitation and masking patterns," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, (Boulder), pp. 206–209, Oct. 1995.

[58] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on sinusoidal representation," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, Aug. 1986.

[59] X. Serra and J. Smith, "Spectral modeling synthesis: A sound analysis/synthesis based on a deterministic plus stochastic decomposition," *Computer Music J.*, vol. 14, pp. 12–24, Winter 1990.

[60] S. N. Levine, *Audio Representations for Data Compression and Compressed Domain Processing.* PhD thesis, Stanford University, Stanford, USA, Dec. 1998.

[61] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, 1992.

[62] A. Segall, "Bit allocation and encoding for vector sources," *IEEE Trans. Information Theory*, vol. 22, pp. 162–169, Mar. 1976.

[63] H. Najafzadeh-Azghandi and P. Kabal, "Perceptual bit allocation for low rate coding of narrowband audio," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, (Istanbul), pp. 893–896, June 2000.

[64] M. Perreau-Guimaraes, M. Bonnet, and N. Moreau, "Low complexity bit allocation algorithm with psychoacoustical optimization," in *European Conf. on Speech Commu. and Tech.*, (Budapest), Sept. 1999.

[65] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications.* Prentice Hall, third ed., 1996.