Chapter 14
MPEG Audio Compression

14.1 Psychoacoustics

- The range of human hearing is about 20 Hz to about 20 kHz
- The frequency range of the voice is typically only from about 500 Hz to 4 kHz
- The dynamic range, the ratio of the maximum sound amplitude to the quietest sound that humans can hear, is on the order of about 120 dB

Equal-Loudness Relations

- **Fletcher-Munson Curves**
  - Equal loudness curves that display the relationship between perceived loudness ("Phons", in dB) for a given stimulus sound volume ("Sound Pressure Level", also in dB), as a function of frequency

- Fig. 14.1 shows the ear's perception of equal loudness:
  - The bottom curve shows what level of pure tone stimulus is required to produce the perception of a 10 dB sound
  - All the curves are arranged so that the perceived loudness level gives the same loudness as for that loudness level of a pure tone at 1 kHz

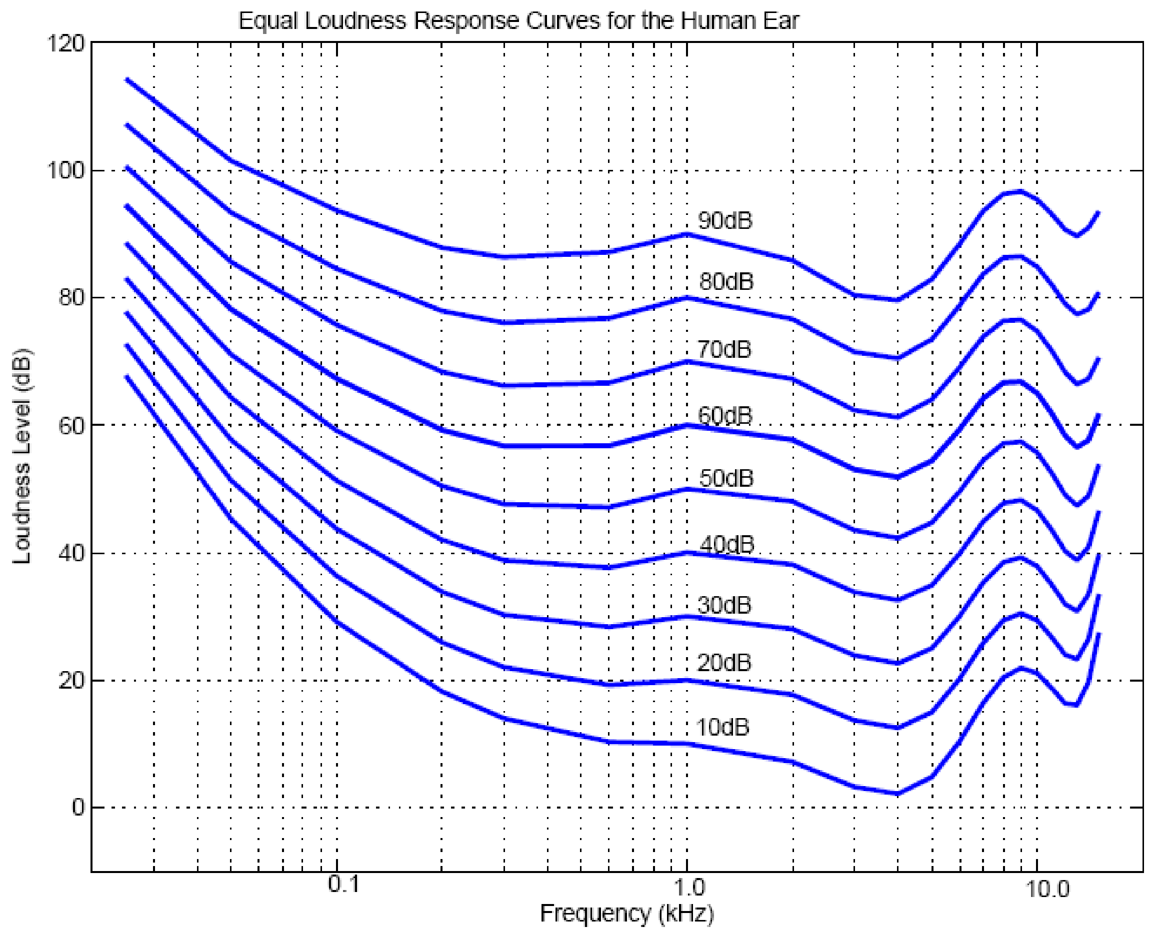

Fig. 14.1: Fletcher-Munson Curves (re-measured by Robinson and Dadson)

Frequency Masking

- Lossy audio data compression methods, such as MPEG/Audio encoding, remove some sounds which are masked anyway
- The general situation in regard to masking is as follows:

1. A lower tone can effectively mask (make us unable to hear) a higher tone
2. The reverse is not true – a higher tone does not mask a lower tone well
3. The greater the power in the masking tone, the wider is its influence – the broader the range of frequencies it can mask.
4. As a consequence, if two tones are widely separated in frequency then little masking occurs

Threshold of Hearing

- A plot of the threshold of human hearing for a pure tone

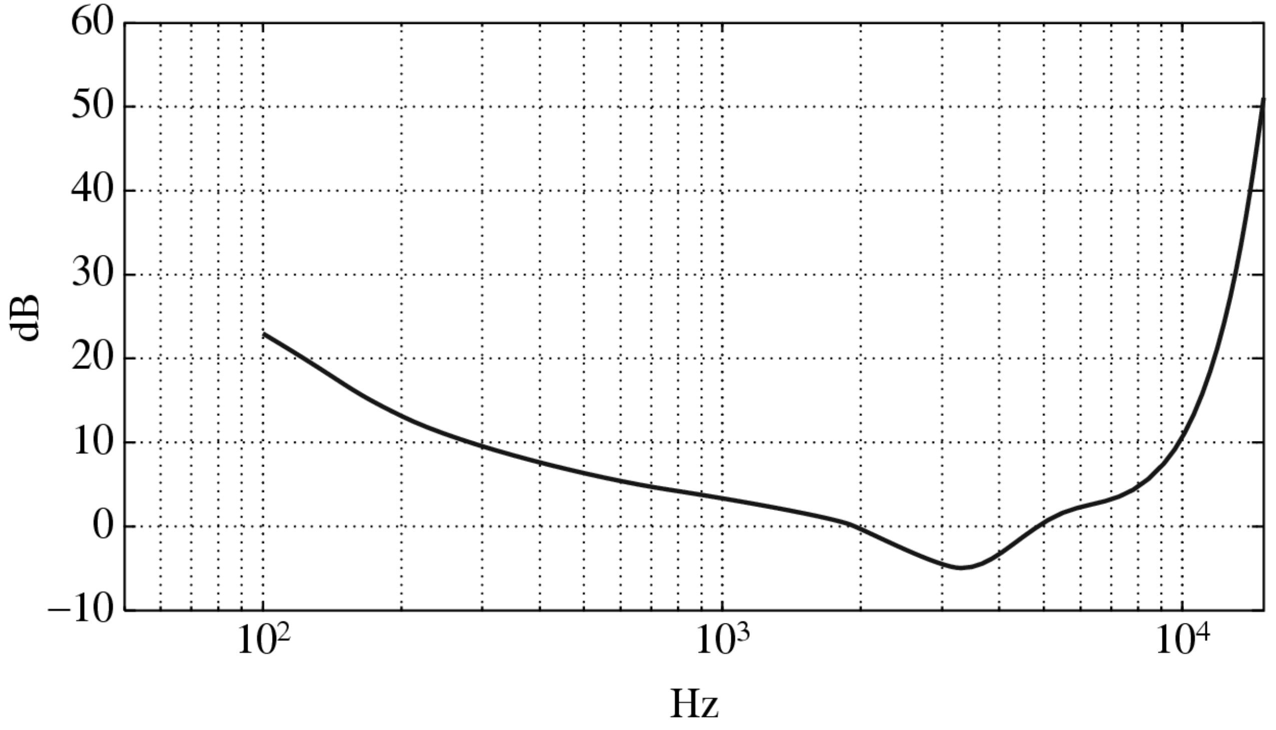

Fig. 14.2: Threshold of human hearing, for pure tones

Threshold of Hearing (Cont'd)

- The threshold of hearing curve: if a sound is above the dB level shown then the sound is audible
- Turning up a tone so that it equals or surpasses the curve means that we can then distinguish the sound
- An approximate formula exists for this curve:

$$\text{Threshold}(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \qquad (14.1)$$

The threshold units are dB; the frequency for the origin (0,0) in formula (14.1) is 2,000 Hz: Threshold(f) = 0 at f =2 kHz

Frequency Masking Curves

- Frequency masking is studied by playing a particular pure tone, say 1 kHz again, at a loud volume, and determining how this tone affects our ability to hear tones nearby in frequency
    - one would generate a 1 kHz *masking* tone, at a fixed sound level of 60 dB, and then raise the level of a nearby tone, e.g., 1.1 kHz, until it is just audible
- The threshold in Fig. 14.3 plots the audible level for a single masking tone (1 kHz)
- Fig. 14.4 shows how the plot changes if other masking tones are used

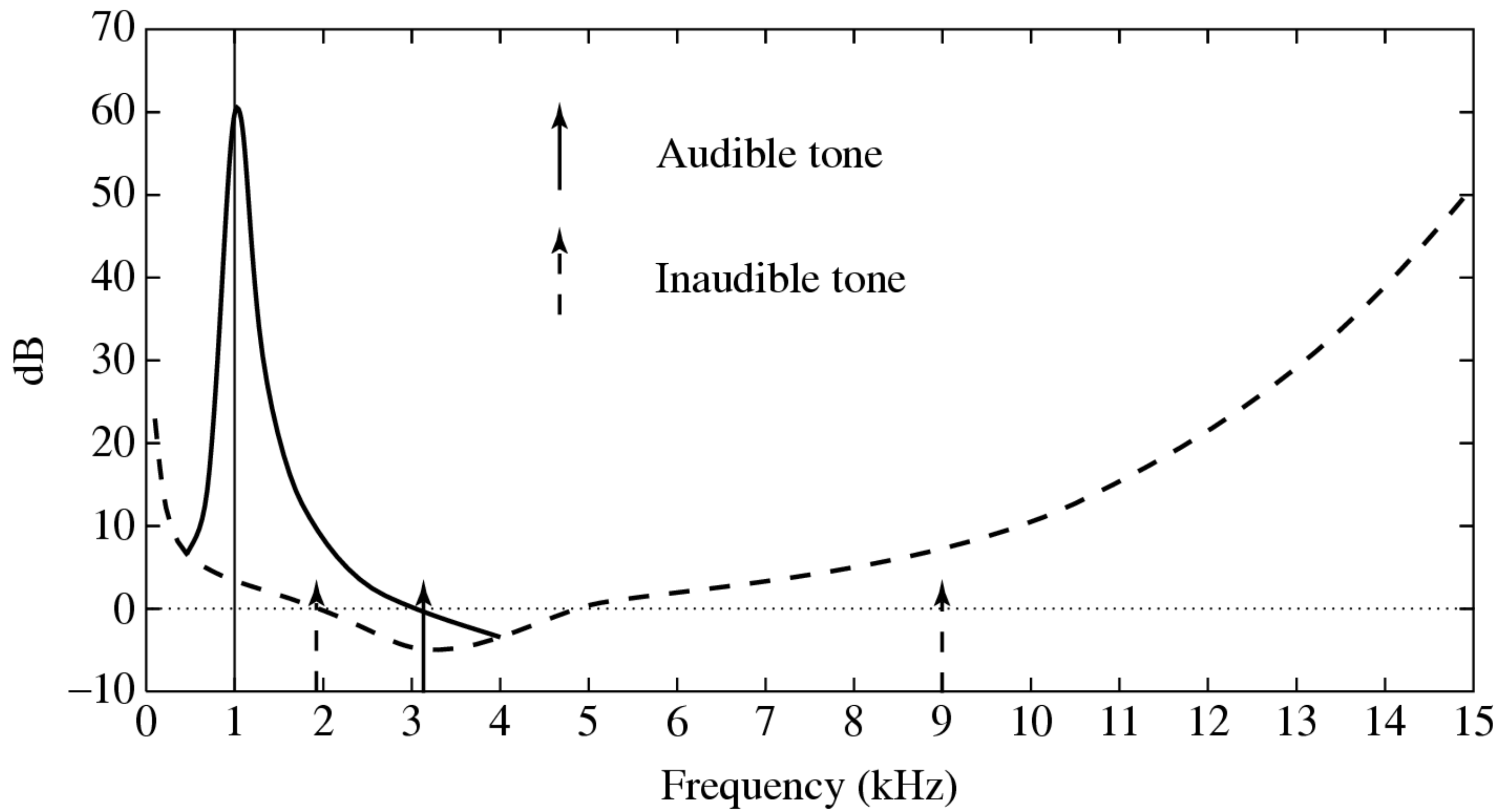

Fig. 14.3: Effect on threshold for 1 kHz masking tone

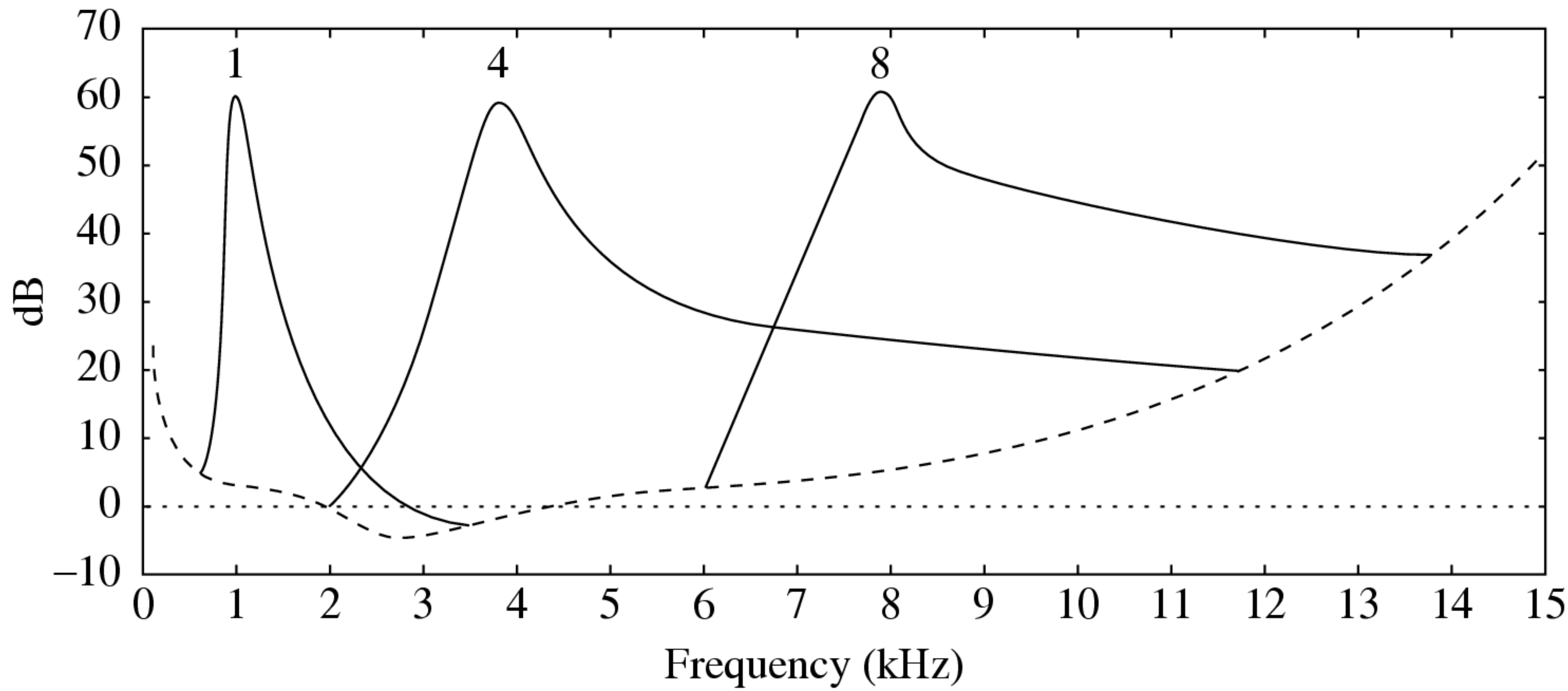

Fig. 14.4: Effect of masking tone at three different frequencies

Critical Bands

- **Critical bandwidth** represents the ear's resolving power for simultaneous tones or partials
  - At the low-frequency end, a critical band is less than 100 Hz wide, while for high frequencies the width can be greater than 4 kHz
- Experiments indicate that the critical bandwidth:
  - for masking frequencies <500Hz: remains approximately constant in width (about 100 Hz)
  - For masking frequencies >500Hz: increases approximately linearly with frequency

Table 14.1: 25-Critical Bands and Bandwidth

Band #	Lower Bound (Hz)	Center (Hz)	Upper Bound (Hz)	Bandwidth (Hz)
1	-	50	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

Band #	Lower Bound (Hz)	Center (Hz)	Upper Bound (Hz)	Bandwidth (Hz)
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	18775	22050	6550

Bark Unit

- **Bark unit** is defined as the width of one critical band, for any masking frequency
- The idea of the Bark unit: every critical band width is roughly equal in terms of Barks (refer to Fig. 14.5)

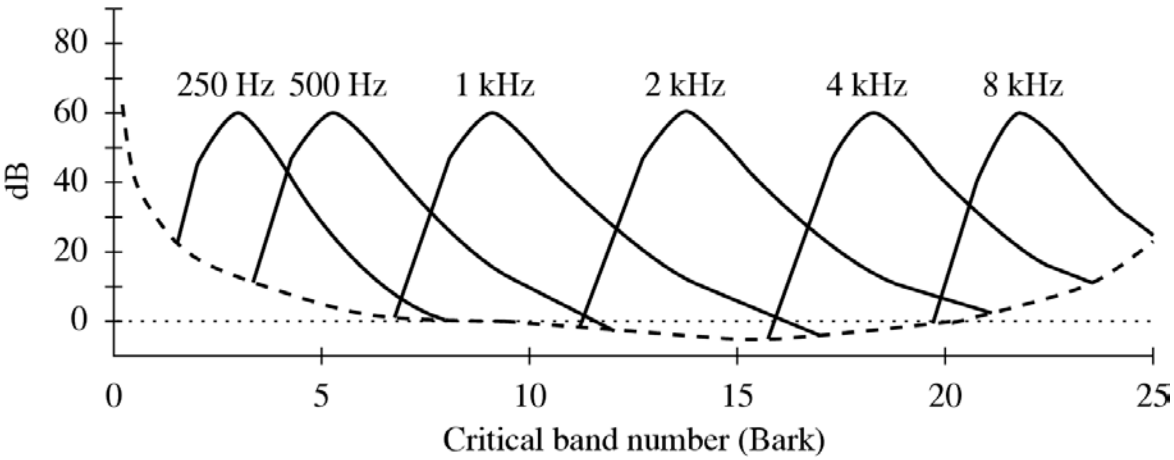

Fig. 14.5: Effect of masking tones, expressed in Bark units

Conversion: Frequency & Critical Band Number

- Conversion expressed in the Bark unit:

$$\text{Critical band number (Bark)} = \begin{cases} f/100, & \text{for } f < 500 \\ 9 + 4\log_2(f/1000), & \text{for } f \geq 500 \end{cases} \tag{14.2}$$

- Another formula used for the Bark scale:

$$b = 13.0\arctan(0.76 f) + 3.5\arctan(f^2/56.25) \tag{14.3}$$

where *f* is in kHz and *b* is in Barks (the same applies to all below)

- The inverse equation:

$$f = [(\exp(0.219 * b)/352) + 0.1] * b - 0.032 * \exp[-0.15 * (b-5)2] \tag{14.4}$$

- The critical bandwidth (*df*) for a given center frequency *f* can also be approximated by:

$$df = 25 + 75 \times [1 + 1.4(f^2)]^{0.69} \tag{14.5}$$

Temporal Masking

- **Phenomenon**: any loud tone will cause the hearing receptors in the inner ear to become *saturated* and require time to recover
- The following figures show the results of masking experiments:

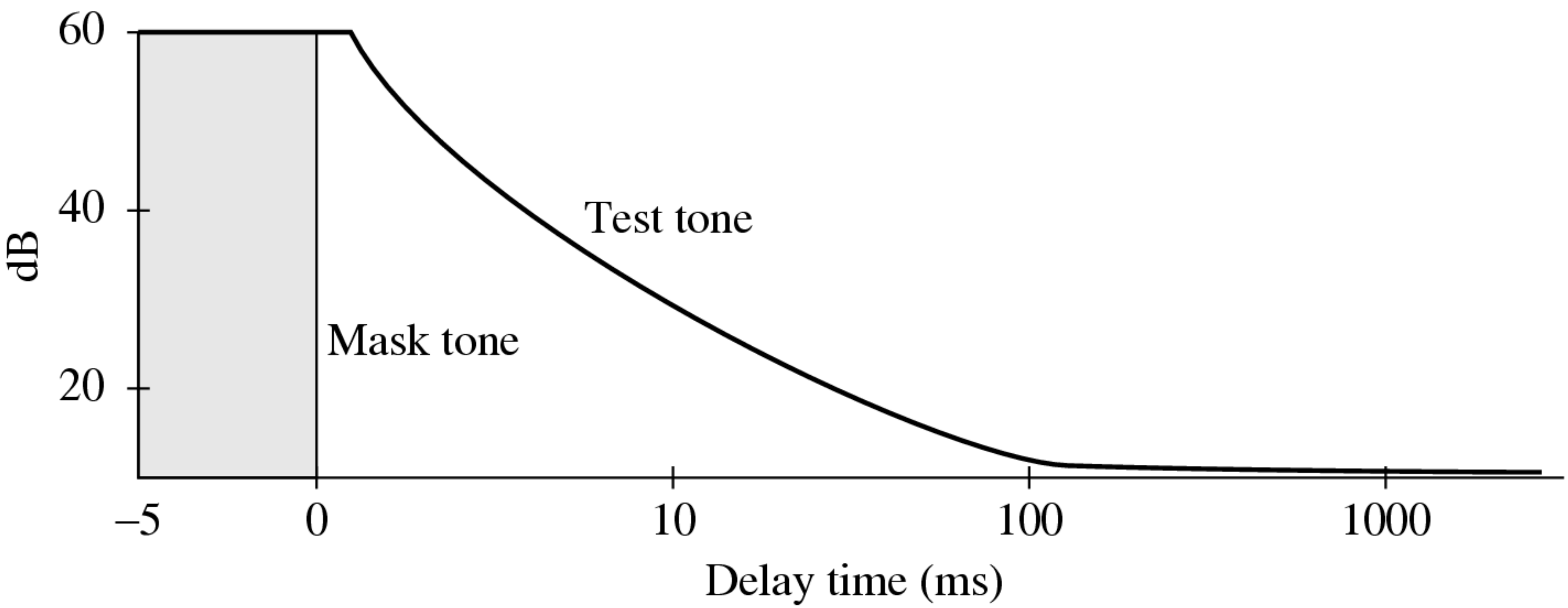

Fig. 14.6: The louder the test tone, the shorter the amount of time required before the test tone is audible once the masking tone is removed.

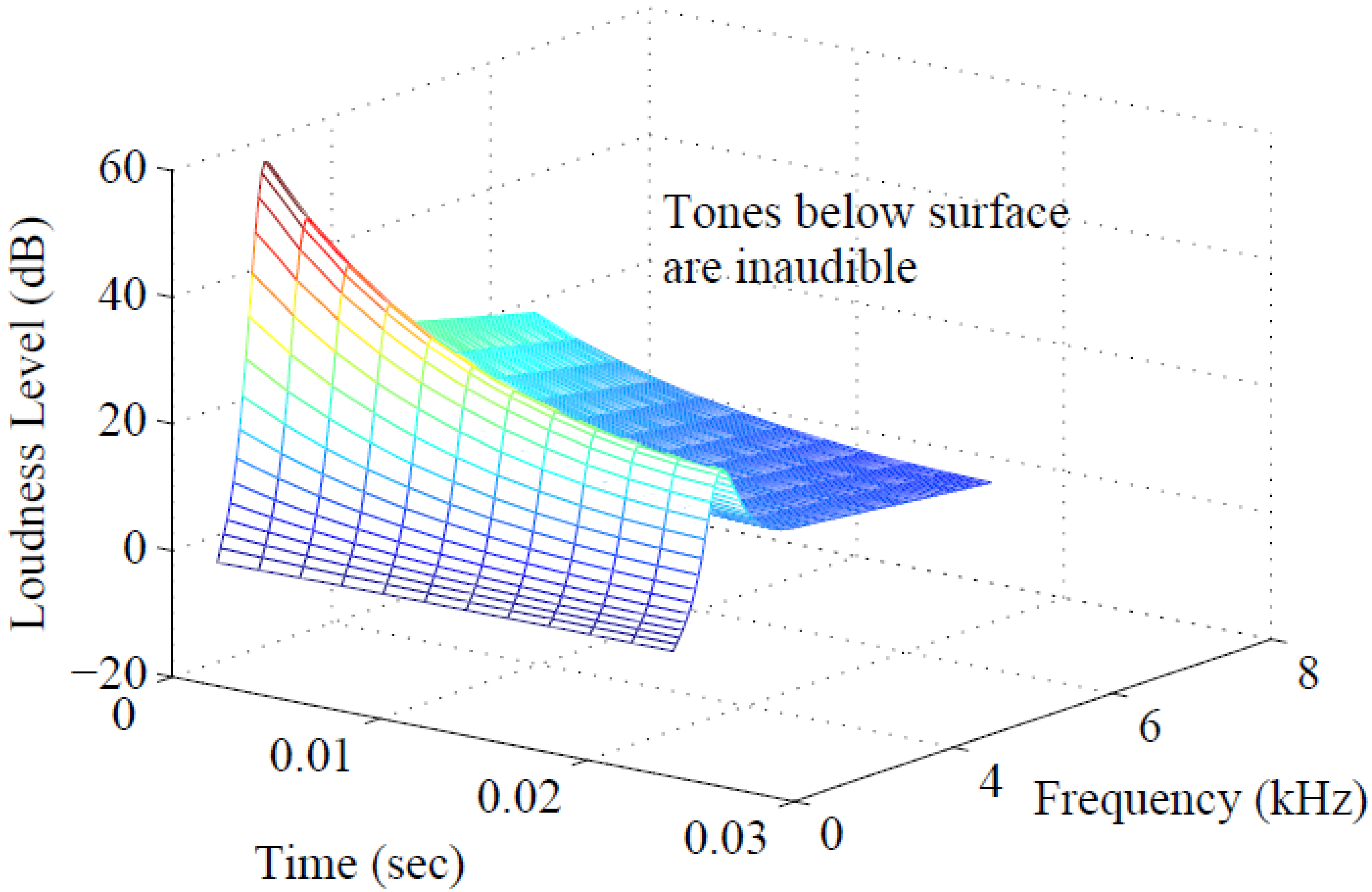

Fig. 14.7: Effect of temporal masking depends on both time and closeness in frequency.

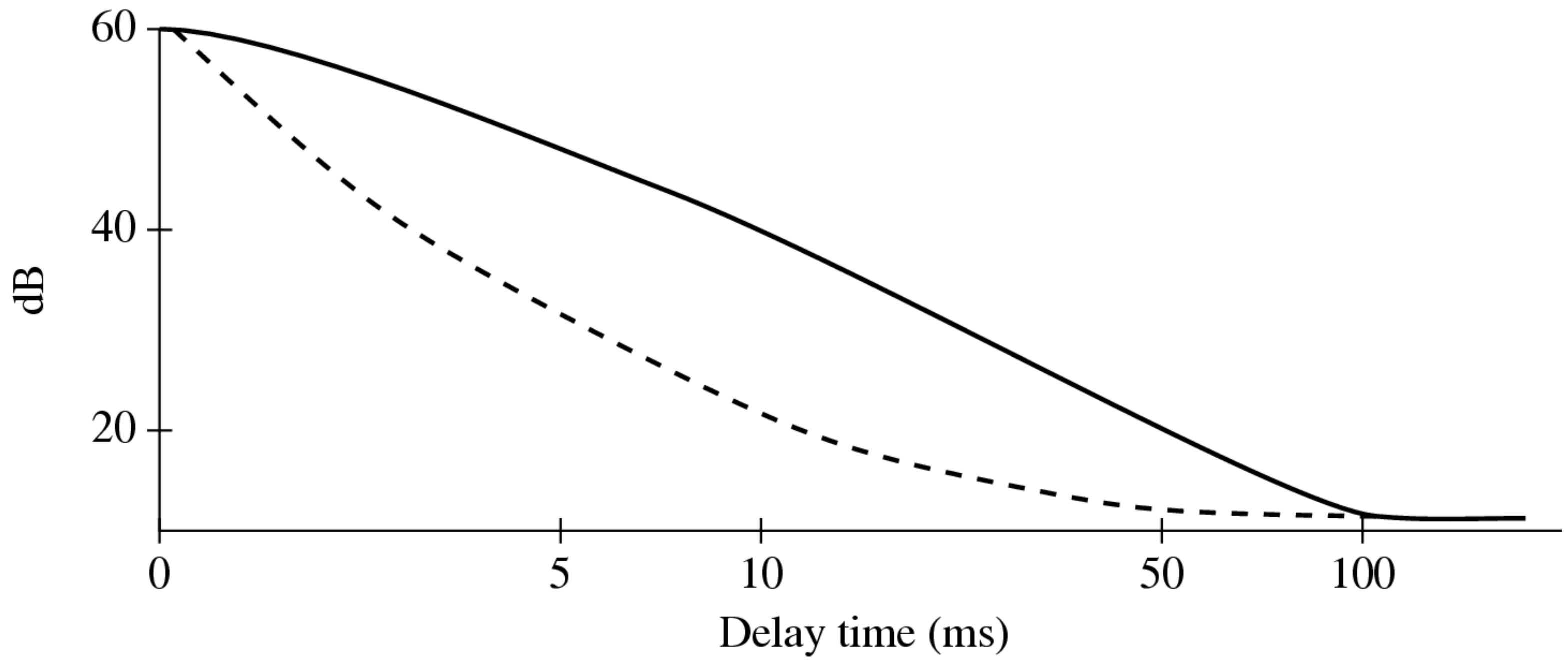

Fig. 14.8: For a masking tone that is played for a longer time, it takes longer before a test tone can be heard. Solid curve: masking tone played for 200 msec; dashed curve: masking tone played for 100 msec.

14.2 MPEG Audio

- **MPEG audio compression** takes advantage of psychoacoustic models, constructing a large multi-dimensional lookup table to transmit masked frequency components using fewer bits

- **MPEG Audio Overview**

1. Applies a filter bank to the input to break it into its frequency components

2. In parallel, a psychoacoustic model is applied to the data for bit allocation block

3. The number of bits allocated are used to quantize the info from the filter bank – providing the compression

MPEG Layers

- MPEG audio offers three compatible *layers*:
  - Each succeeding layer able to understand the lower layers
  - Each succeeding layer offering more complexity in the psychoacoustic model and better compression for a given level of audio quality
  - each succeeding layer, with increased compression effectiveness, accompanied by extra delay
- The objective of MPEG layers: a good tradeoff between quality and bit-rate

MPEG Layers (Cont'd)

- Layer 1 quality can be quite good provided a comparatively high bit-rate is available
  - Digital Audio Tape typically uses Layer 1 at around 192 kbps
- Layer 2 has more complexity; was proposed for use in Digital Audio Broadcasting
- Layer 3 (MP3) is most complex, and was originally aimed at audio transmission over ISDN lines
- Most of the complexity increase is at the encoder, not the decoder – accounting for the popularity of MP3 players

MPEG Audio Strategy

- **MPEG approach to compression** relies on:
  - Quantization
  - Inaccuracy of human auditory system within the width of a critical band

- **MPEG encoder** employs a bank of filters to:
  - Analyze the frequency ("spectral") components of the audio signal by calculating a frequency transform of a window of signal values
  - Decompose the signal into subbands by using a bank of filters (Layer 1 & 2: "quadrature-mirror"; Layer 3: adds a DCT; psychoacoustic model: Fourier transform)

MPEG Audio Strategy (Cont'd)

- **Frequency masking**: by using a psychoacoustic model to estimate the just noticeable noise level:
    - Encoder balances the masking behavior and the available number of bits by discarding inaudible frequencies
    - Scaling quantization according to the sound level that is left over, above masking levels
- May take into account the actual width of the critical bands:
    - For practical purposes, audible frequencies are divided into 25 main critical bands (Table 14.1)
    - To keep simplicity, adopts a *uniform* width for all frequency analysis filters, using 32 overlapping subbands

MPEG Audio Compression Algorithm

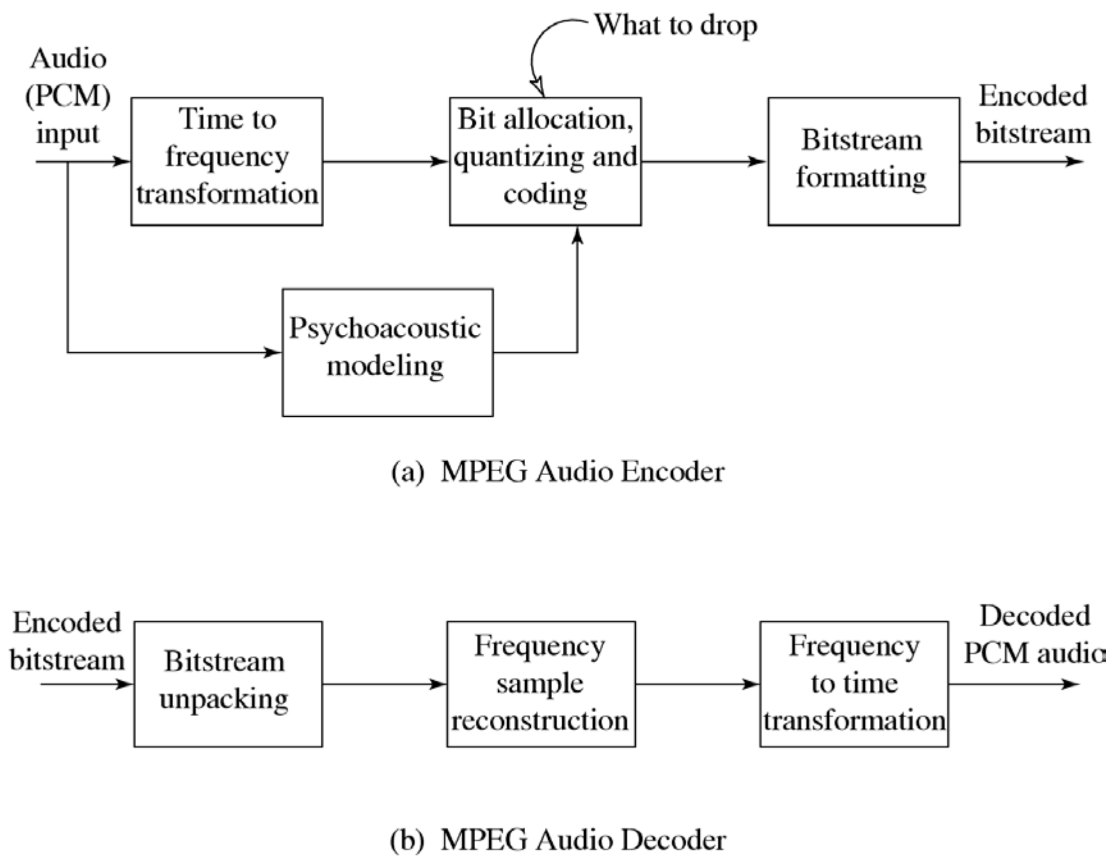

Fig. 14.9: Basic MPEG Audio encoder and decoder.

Basic Algorithm (Cont'd)

- The algorithm proceeds by dividing the input into 32 frequency subbands, via a filter bank
  - A linear operation taking 32 PCM samples, sampled in time; output is 32 frequency coefficients
- In the Layer 1 encoder, the sets of 32 PCM values are first assembled into a set of 12 groups of 32s
  - an inherent time lag in the coder, equal to the time to accumulate 384 (i.e., 12×32) samples
- Fig.14.11 shows how samples are organized
  - A Layer 2 or Layer 3, frame actually accumulates more than 12 samples for each subband: a frame includes 1,152 samples

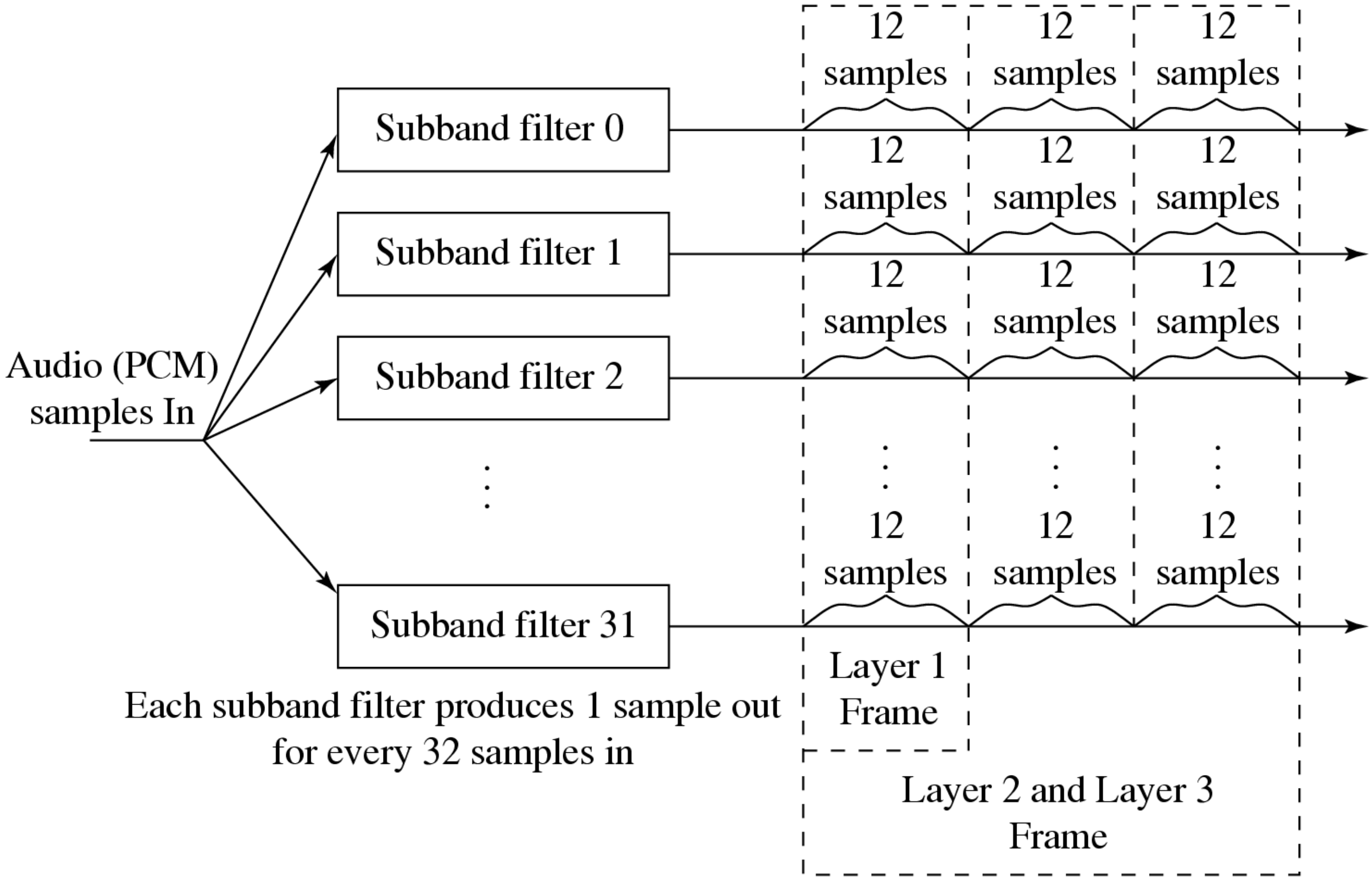

Fig. 14.11: MPEG Audio Frame Sizes

Bit Allocation Algorithm

- **Aim**: ensure that all of the quantization noise is below the masking thresholds
- **One common scheme**:
  - For each subband, the psychoacoustic model calculates the *Signal-to-Mask Ratio* (SMR)in dB
  - Then the "Mask-to-Noise Ratio" (MNR) is defined as the difference (as shown in Fig.14.12):

$$\mathrm{MNR_{dB}} \equiv \mathrm{SNR_{dB}} - \mathrm{SMR_{dB}} \qquad (14.6)$$

  - The lowest MNR is determined, and the number of code-bits allocated to this subband is incremented
  - Then a new estimate of the SNR is made, and the process iterates until there are no more bits to allocate

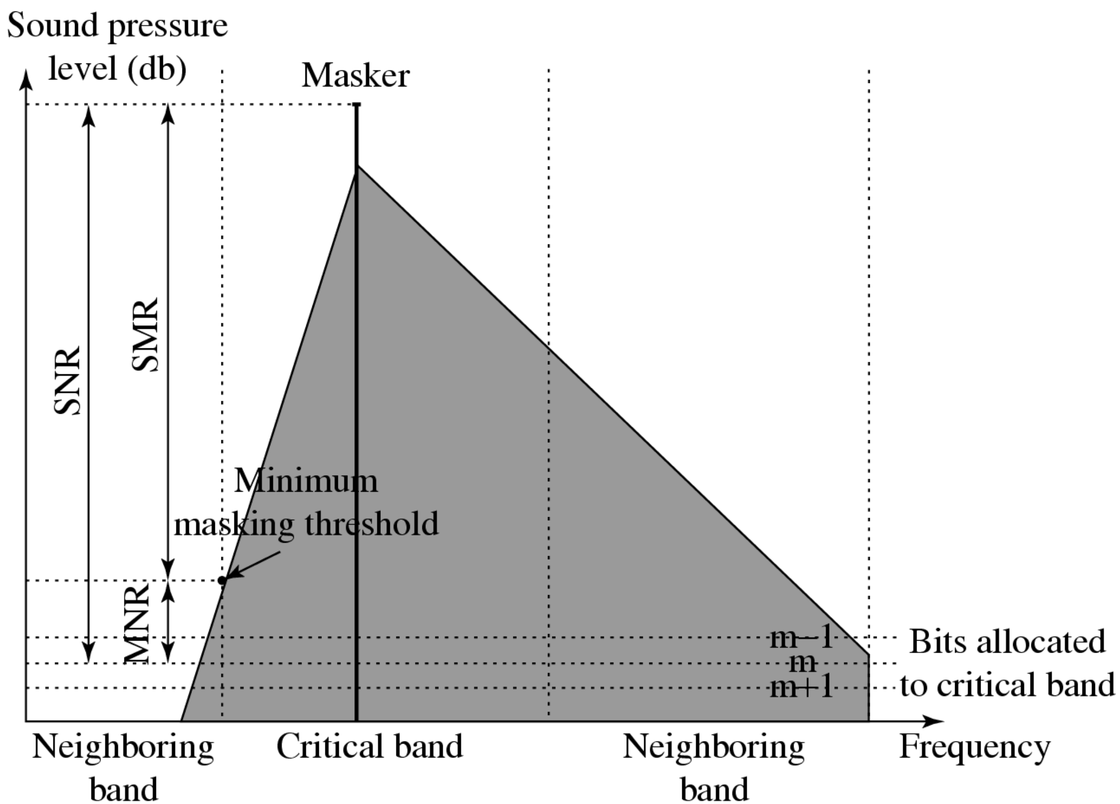

Fig. 14.12: MNR and SMR. A qualitative view of SNR, SMR and MNR are shown, with one dominate masker and m bits allocated to a particular critical band.

- Mask calculations are performed in parallel with subband filtering, as in Fig. 4.13:

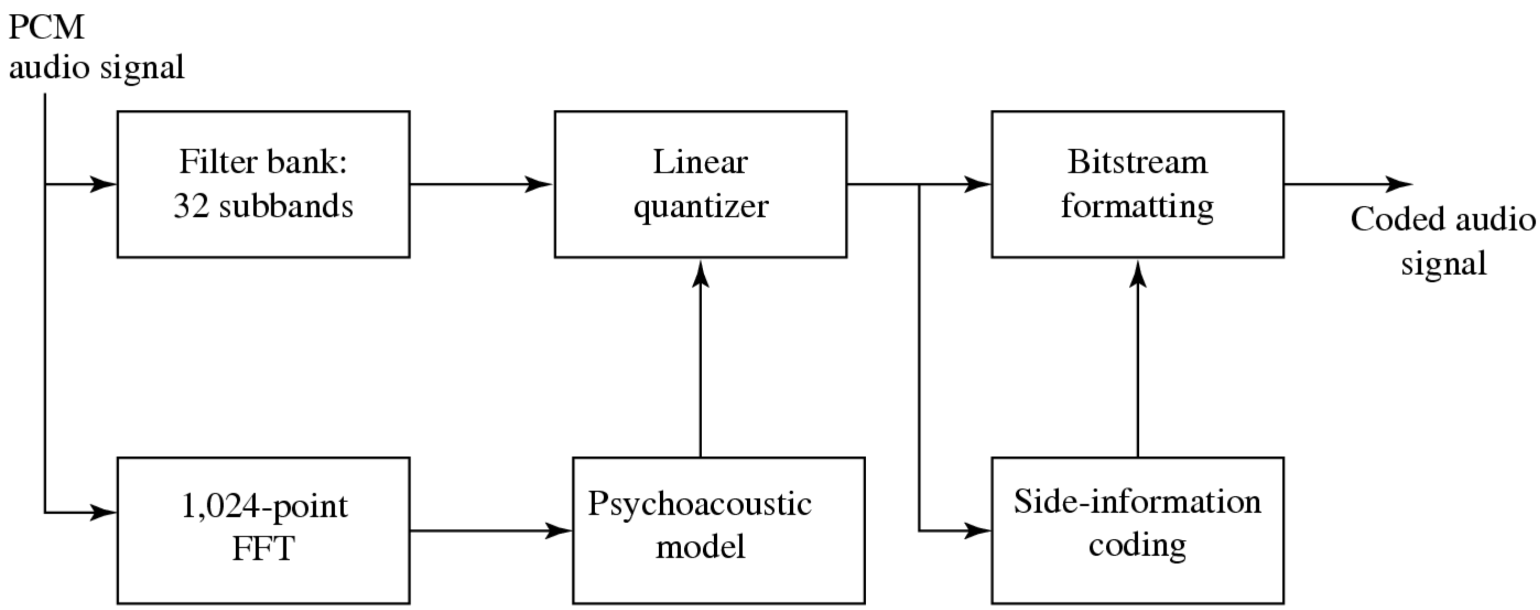

Fig. 14.13: MPEG-1 Audio Layers 1 and 2.

Layer 2 of MPEG-1 Audio

- **Main difference**:
  - Three groups of 12 samples are encoded in each frame and temporal masking is brought into play, as well as frequency masking
  - Bit allocation is applied to window lengths of 36 samples instead of 12
  - The resolution of the quantizers is increased from 15 bits to 16
- **Advantage**:
  - a single scaling factor can be used for all three groups

Layer 3 of MPEG-1 Audio

- **Main difference**:
    - Employs a similar filter bank to that used in Layer 2, except using a set of filters with non-equal frequencies
    - Takes into account stereo redundancy
    - Uses Modified Discrete Cosine Transform (MDCT) — addresses problems that the DCT has at boundaries of the window used by overlapping frames by 50%:

$$F(u) = 2\sum_{i=0}^{N-1} f(i)\cos\left[\frac{2\pi}{N}\left(i + \frac{N/2+1}{2}\right)(u+1/2)\right], u = 0,..,N/2-1 \qquad (14.7)$$

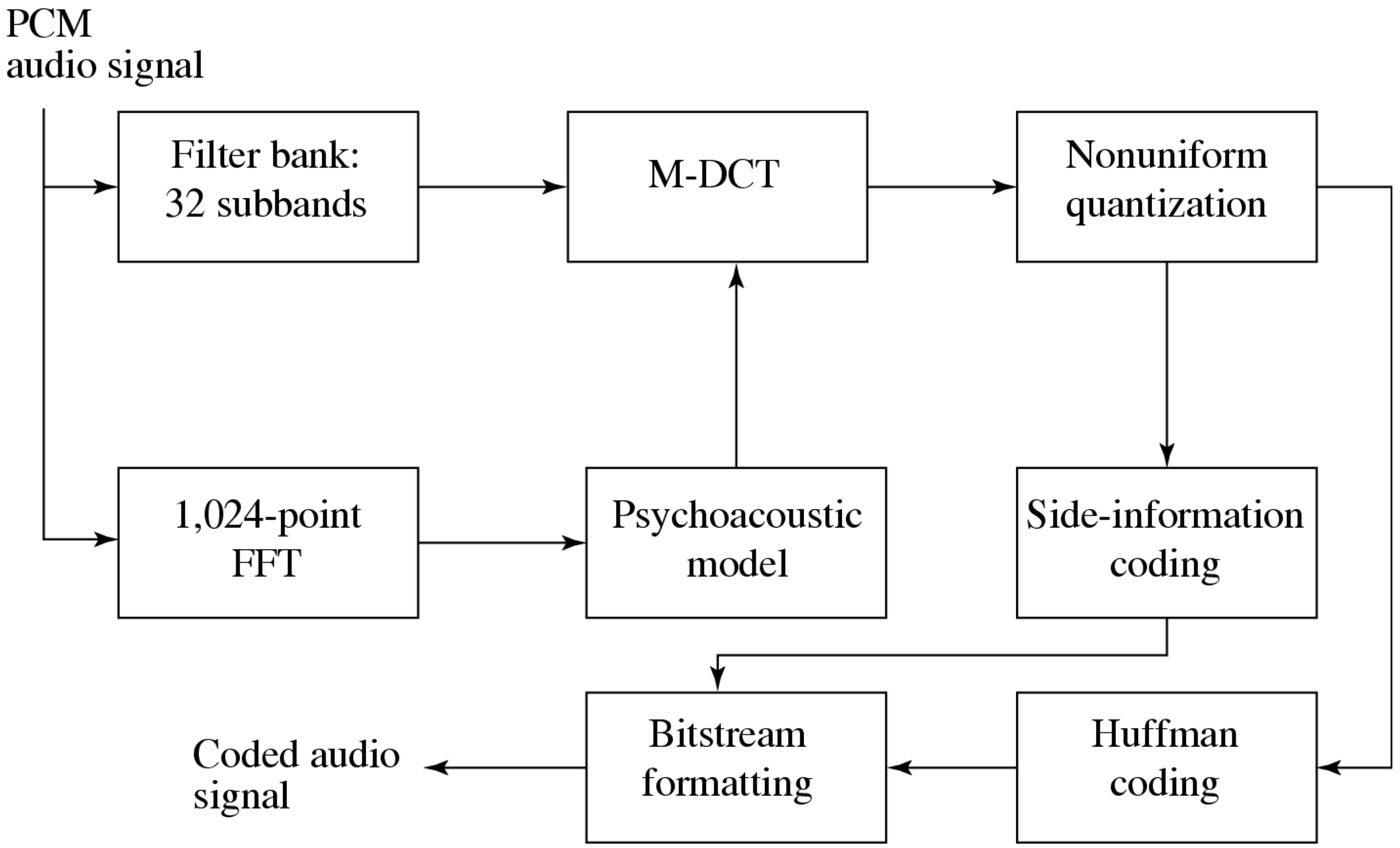

Fig 14.14: MPEG-Audio Layer 3 Coding.

- Table 14.2 shows various achievable MP3 compression ratios:

Table 14.2: MP3 compression performance

Sound Quality	Bandwidth	Mode	Compression Ratio
Telephony	3.0 kHz	Mono	96:1
Better than Short-wave	4.5 kHz	Mono	48:1
Better than AM radio	7.5 kHz	Mono	24:1
Similar to FM radio	11 kHz	Stereo	26 - 24:1
Near-CD	15 kHz	Stereo	16:1
CD	$>$ 15 kHz	Stereo	14 - 12:1

MPEG-2 AAC (Advanced Audio Coding)

- AAC was developed to succeed MP3 for digital audio; delivers better sound quality than MP3 for the same bitrate
- Default audio format for YouTube, iPhone, iTunes, Nintendo, and PlayStation
- The standard vehicle for DVDs:
  - Audio coding technology for the DVD-Audio Recordable (DVD-AR) format, also adopted by XM Radio
- Aimed at transparent sound reproduction for theaters
  - Can deliver this at 320 kbps for five channels so that sound can be played from 5 different directions: Left, Right, Center, Left-Surround, and Right-Surround
  - 5.1 channel systems also include a low-frequency enhancement (LFE) channel (a “woofer”)
- Also capable of delivering high-quality stereo sound at bit-rates below 128 kbps

MPEG-2 AAC (Cont'd)

- Support up to 48 channels, sampling rates between 8 kHz and 96 kHz, and bit-rates up to 576 kbps per channel
- Like MPEG-1, MPEG-2, supports three different "profiles", but with a different purpose:
  - *Main* profile
  - *Low Complexity*(LC) profile
  - *Scalable Sampling Rate* (SSR) profile

MPEG-4 Audio

- Integrates several different audio components into one standard: speech compression, perceptually based coders, text-to-speech, and MIDI
- *MPEG-4 AAC* (*Advanced Audio Coding*), is similar to the MPEG-2 AAC standard, with some minor changes
- **Perceptual Coders**
  - Incorporate a *Perceptual Noise Substitution* module
  - Include a *Bit-Sliced Arithmetic Coding* (BSAC) module
  - Also include a second perceptual audio coder, a vector-quantization method entitled TwinVQ

MPEG-4 Audio (Cont'd)

- **Structured Coders**
  - Takes "Synthetic/Natural Hybrid Coding" (SNHC) in order to have very low bit-rate delivery an option
  - **Objective**: integrate both "natural" multimedia sequences, both video and audio, with those arising synthetically – "structured" audio
  - Takes a "toolbox" approach and allows specification of many such models.
  - E.g., *Text-To-Speech* (TTS) is an ultra-low bit-rate method, and actually works, provided one need not care what the speaker actually sounds like

14.3 Other Commercial Audio Codecs

- Table 14.3 summarizes the target bit-rate range and main features of other modern general audio codecs

Table 14.3: Comparison of MP3, MPEG-4 AAC, and Ogg vorbis

	MP3	MPEG-4 AAC	Ogg vorbis
File extension	.mp3	.aac, .mp4, .3gp	.ogg
Original name	MPEG-1 Audio Layer 3	Advanced Audio Coding	Ogg
Developer	CCETT, IRT, Fraunhofer Society	Fraunhofer IIS, AT&T Bell Labs, Dolby, Sony Corp., and Nokia	Xiph.org Foundation
Released	1994	1997	v1.0 frozen May 2000
Algorithm	lossy compression	lossy compression	lossy comrpession
Quality	Lower quality than AAC and Ogg	Better quality at same bit rate as MP3	Better quality and smaller file size than MP3 at same bit rates
Used in	Default standard for audio files	iTunes raised its popularity	Open-source platform

- Table 14.4 summarizes the target bitrate range and main features of other modern general audio codecs

Table 14.4: Comparison of audio coding systems

Codec	Bitrate kbps/channel	Complexity	Main application
Dolby AC-2	128–192	Low (encoder/decoder)	Point-to-point, cable
Dolby AC-3	32–640	Low (decoder)	HDTV, cable, DVD
Dolby Digital (Enhanced AC-3)	32–6144	Low (decoder)	HDTV, cable, DVD
DTS: Digital Surround	8–512	Low (for lossless audio extension)	DVD, entertainment, professional
WMA: Windows Media Audio	128–768	Low (low-bit-rate streaming)	Many applications
MPEG SAOC	As low as 48	Low (decoder/rendering)	Many applications

14.4 MPEG-7 Audio and Beyond

- **Difference** from current standards:
- MPEG-4 is aimed at compression using objects.
- MPEG-7 is mainly aimed at "search": How can we find objects, assuming that multimedia is indeed coded in terms of objects
  - The objective, in terms of audio, is to facilitate the representation and search for sound content, perhaps through the tune or other descriptors
  - An example application supported by MPEG-7 is automatic speech recognition (ASR)
  - Further standards in the MPEG sequence are mostly not aimed at further audio compression standardization. For example, MPEG-DASH (Dynamic Adaptive Streaming over HTTP) is aimed at streaming of multimedia