CMPT 365 Multimedia Systems

Media Compression
- Audio

Spring 2017

Edited from slides by Dr. Jiangchuan Liu

Approximate file sizes for **1 second** of audio

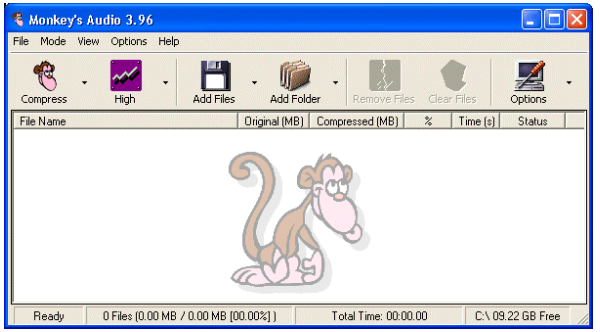
Channels	Resolution	Fs	File Size
Mono	8bit	8Khz	64Kb
Stereo	8bit	8Khz	128Kb
Mono	16bit	8Khz	128Kb
Stereo	16bit	16Khz	256Kb
Stereo	16bit	44.1Khz	1441Kb*
Stereo	24bit	44.1Khz	2116Kb

1CD 700M 70-80 mins

Outline

- Lossless Audio Coding
- Lossy Audio Coding
- MPEG audio (MP3)

Lossless coding



- Monkey's Audio – APE format
  - Slightly better than FLAC (but time-consuming decoding)
  - Officially only for Windows
- Algorithm
  - Linear prediction: estimate what the value a sample will have, based on previous samples (recall AR(1))

$$x'(n) = \sum_{i=1}^{p} a_i x(n-i)$$

  - Channel coupling - mid/side-coding:
    - calculates a "mid"-channel by addition of left and right channel (l+r)/2 and a "side"-channel (l-r)/2.
  - Range coding (an entropy coder):
    - Similar to Arithmetic Coding
    - Build a table of frequencies and then allocate certain ranges of numbers to a certain value

Lossless coding

- **FLAC (Free Lossless Audio Codec)**
  - Linear prediction
  - Golomb-Rice coding
    - closely related to Huffman/Arithmetic
  - Run-length coding

Performance of Lossless Coding

- http://members.home.nl/w.speek/comparison.htm

Compressor	Options	Ratio*	Encoding Speed**	Decoding Speed**
		%	x realtime	x realtime
La 0.4b	default	55.5	2.1	2.7
OptimFROG 4.509	highnew	55.8	1.1	1.5
Monkey's Audio 3.99	extra high	56.4	8.8	8.7
OptimFROG 4.509	default	56.7	6.0	8.9
Monkey's Audio 3.99	high	56.9	15.8	14.3
Monkey's Audio 3.99	normal	57.3	18.1	16.0
WavPack 4.0	high	58.0	16.0	16.1
WMA 9	default	58.0	9.4	10.8
RKAU 1.07	fast (l1)	58.4	8.0	9.6
LPAC Archiver 1.41	medium, JS, random access	58.8	13.5	24.6
LPAC Archiver 1.41	extra high, JS, random access	58.8	8.1	20.2
TTA 3.2	default	59.0	26.6	23.4
WavPack 4.0	normal	59.4	26.3	28.4
FLAC 1.1.2	8	59.6	4.2	44.7
FLAC 1.1.2	default (5)	59.8	19.9	44.7
Apple Lossless (iTunes 4.7)	automatic	60.0	15.6	35.0
Shorten 3.6.0	default	63.7	33.7	70.9

Outline

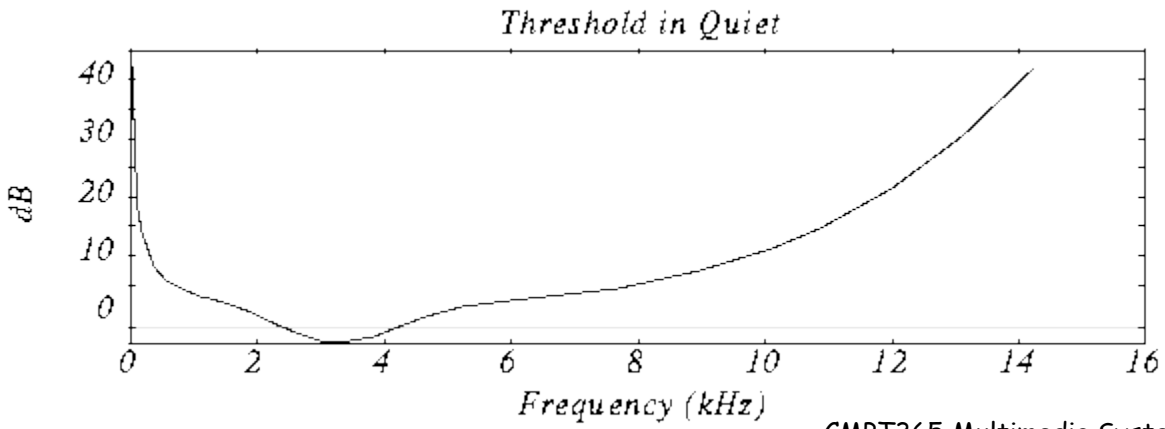
- Lossless Audio Coding
- Lossy Audio Coding
- MPEG audio (MP3)

Lossy coding: Perceptual Coding

- Hide errors where humans will not see or hear it
  - Study hearing and vision system to understand how we see/hear
  - Masking refers to one signal overwhelming/hiding another (e.g., loud siren or bright flash)
- Natural Bandlimitng
  - Audio perception is 20-20 kHz but most sounds in low frequencies (e.g., 2 kHz to 4 kHz)
  - Low frequencies may be encoded as single channel

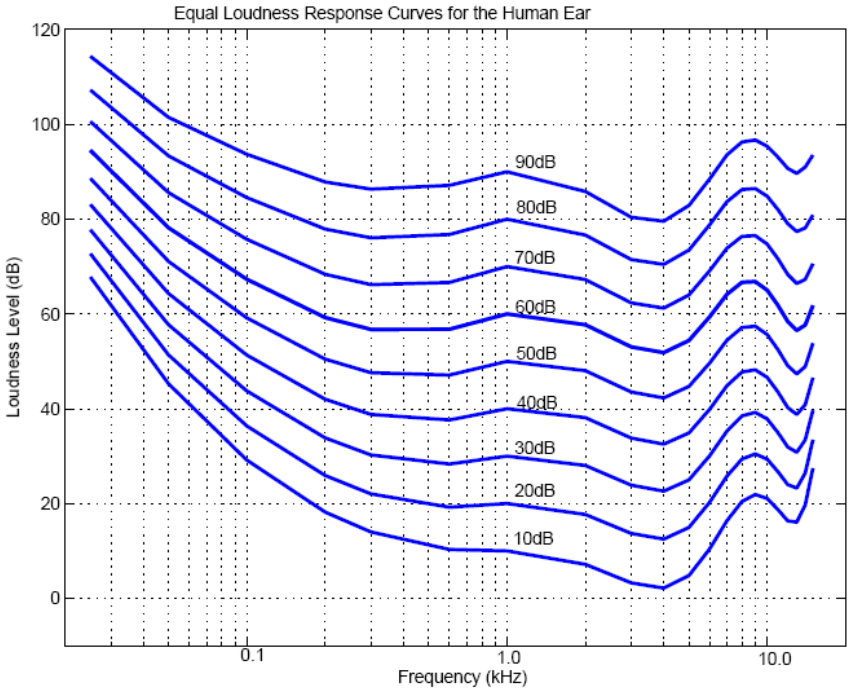
Psychoacoustic Model

- Basically: If you can't hear the sound, don't encode it
  - Frequency range is about 20 Hz to 20 kHz, most sensitive at 2 to 4 KHz.
  - Dynamic range (quietest to loudest) is about 96 dB
  - Normal voice range is about 500 Hz to 2 kHz
    - Low frequencies are vowels and bass
    - High frequencies are consonants
- Threshold of Hearing
  - Experiment: Put a person in a quiet room. Raise level of 1 kHz tone until just barely audible. Vary the frequency and plot



Psychoacoustic Model con'td

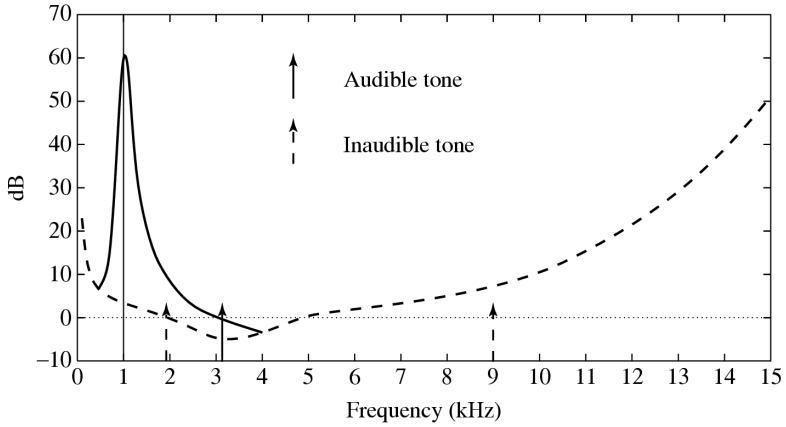
- Fletcher-Munson Curves

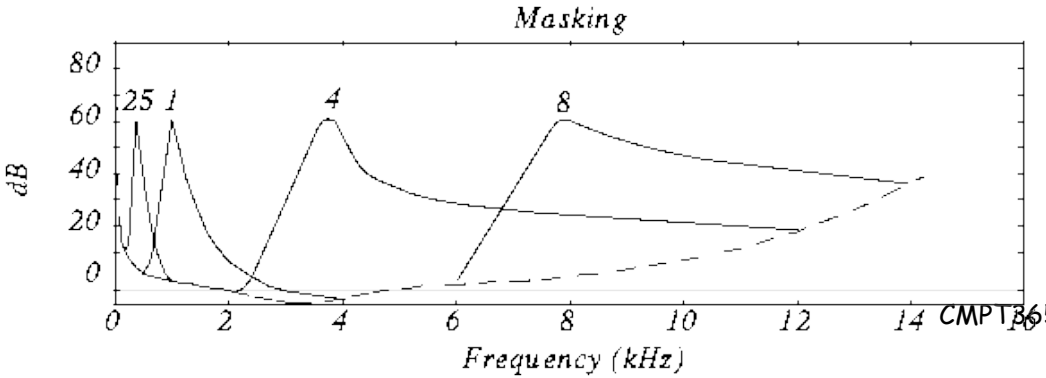


- Equal loudness curves that display the relationship between perceived loudness ("Phons", in dB) for a given stimulus sound volume ("Sound Pressure Level", also in dB), as a function of frequency
- The bottom curve shows what level of pure tone stimulus is required to produce the perception of a 10 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

Psychoacoustic Model con'td

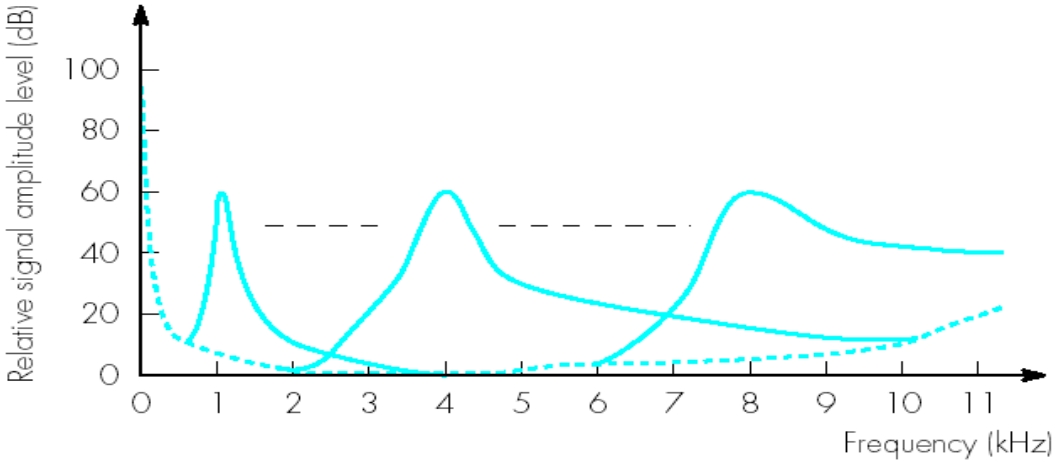
- Frequency masking: Do receptors interfere with each other?
- Experiment:
  - Play 1 kHz tone (*masking tone*) at fixed level (60 dB). Play *test tone* at a different level and raise level until just distinguishable.
  - Vary the frequency of the test tone and plot the threshold when it becomes audible:





Psychoacoustic Model con'td

- Frequency masking: If within a critical band a stronger sound and weaker sound compete, you can't hear the weaker sound. Don't encode it.



Our brains perceive the sounds through 25 distinct ***critical bands***. The bandwidth grows with frequency (above 500Hz).

- At 100Hz, the bandwidth is about 160Hz;
- At 10kHz it is about 2.5kHz in width.

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

Psychoacoustic Model con'td

- **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)

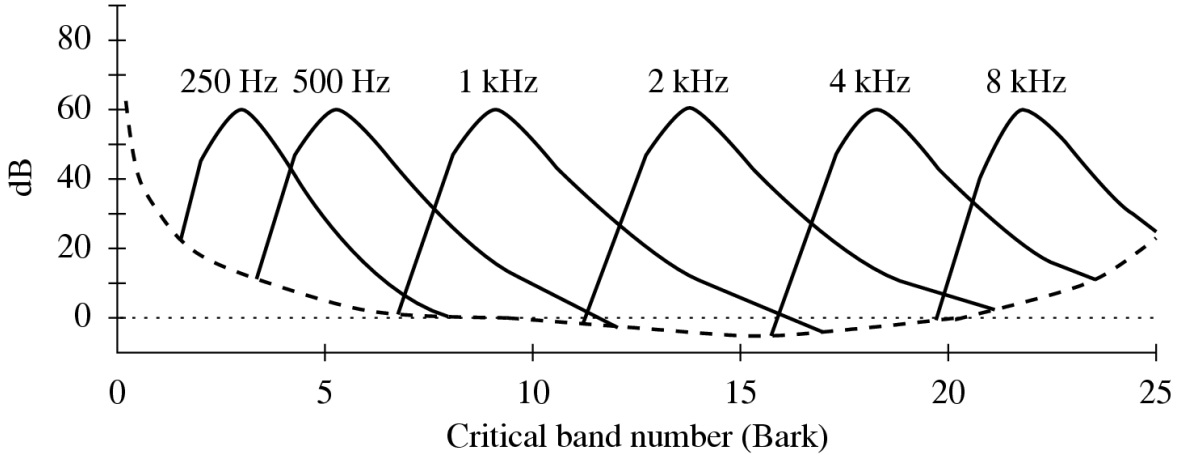


Fig. 14.5: Effect of masking tones, expressed in Bark units

Conversion: Frequency & Critical Band Number

- • Conversion expressed in the Bark unit:

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

- • Another formula used for the Bark scale:

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

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

- The inverse equation:

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

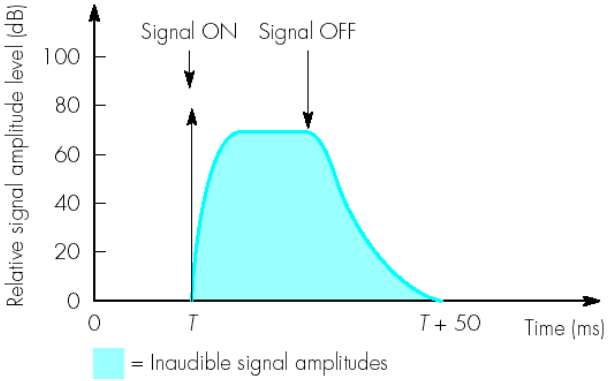
- The critical bandwidth (*df*) for a given center frequency *f* can also be approximated by:

(14.5)

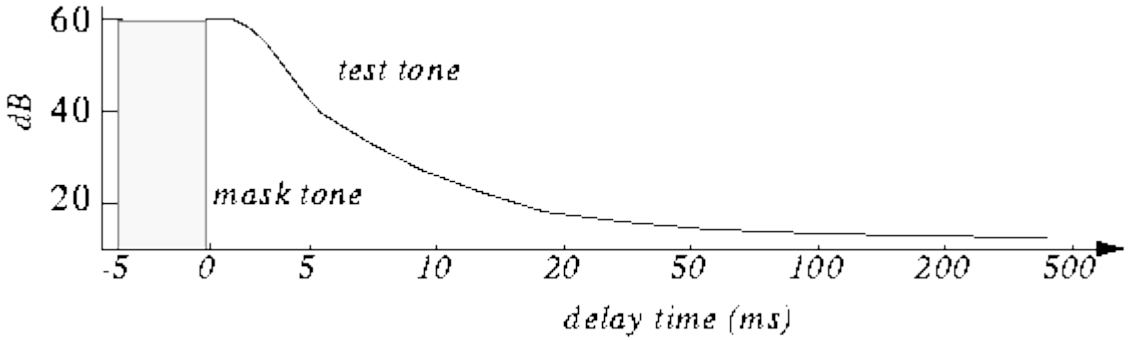
$$df = 25 + 75 \times [1 + 1.4(f^2)]^{0.69}$$

Psychoacoustic Model con'td

- Temporal masking: If we hear a loud sound, it takes a little while until we can hear a soft tone nearby.



- Experiment:
  - Play 1 kHz *masking tone* at 60 dB, plus a *test tone* at 1.1 kHz at 40 dB. Test tone can't be heard (it's masked). Stop masking tone, then stop test tone after a short delay.
  - Adjust delay to the shortest time when test tone can be heard.
  - Repeat with different level of the test tone and plot:



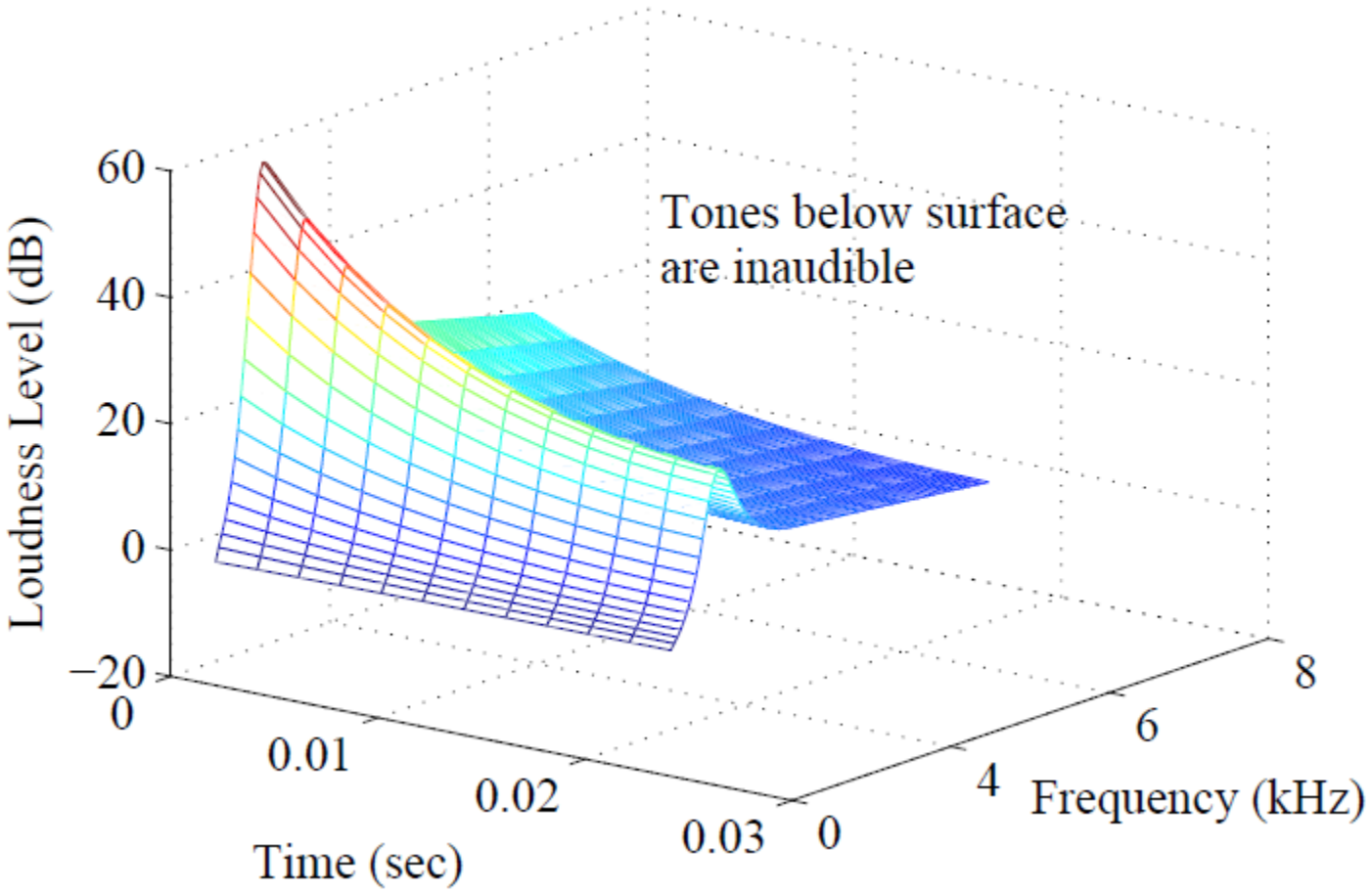


Fig. 14.7: Effect of temporal masking depends on both time and closeness in frequency.

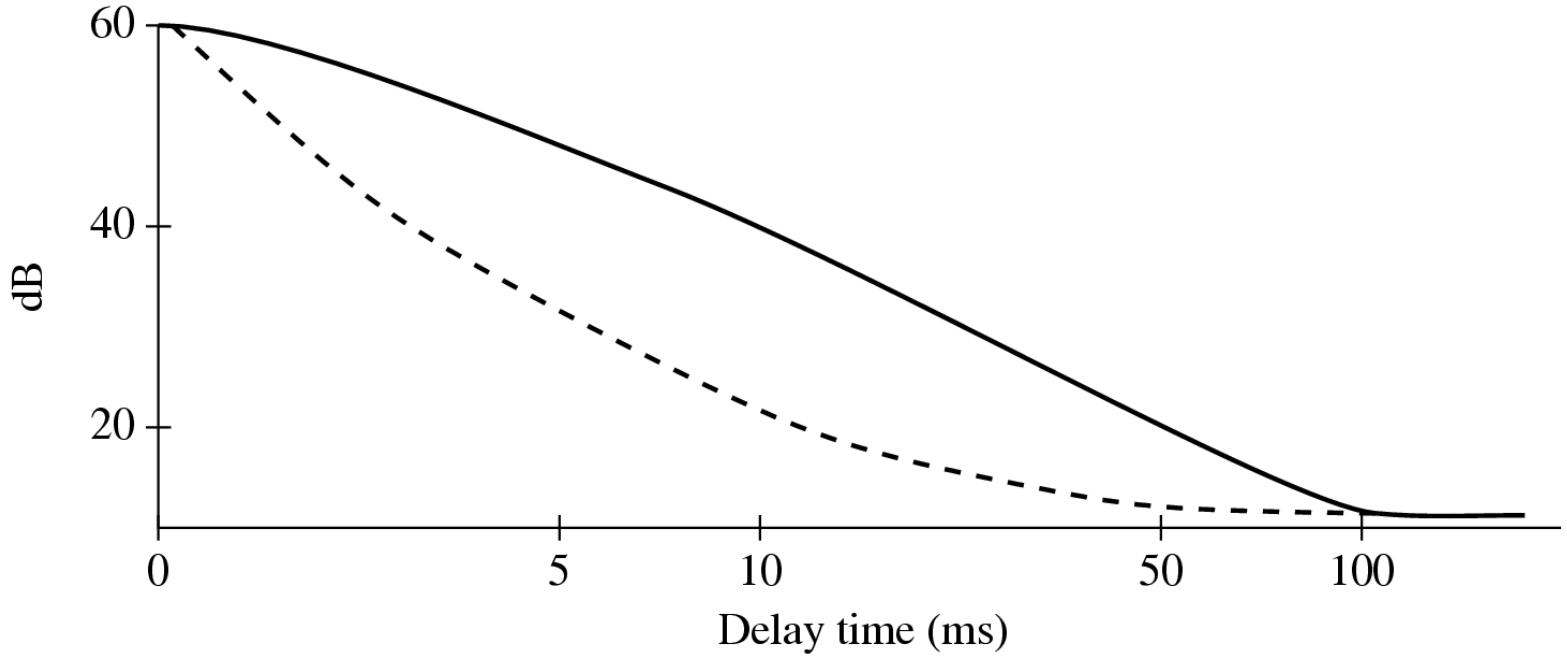
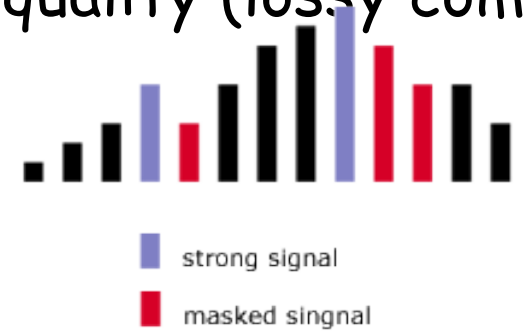


Fig. 14.8: For a masking tone that is played for a longer time, it takes longer before a test tone can be heard. Solid curve: masking tone played for 200 msec; dashed curve: masking tone played for 100 msec.

Perceptual Coding

- Makes use of psychoacoustic knowledge to reduce the amount of information required to achieve the same perceived quality (lossy compression)



- Example:
  - Sony MiniDisc uses Adaptive TRAnsform Coding (ATRAC) to achieve a 5:1 compression ratio (about 141 kbps)
  - MPEG audio (MP3)

http://www.mpeg.org
http://www.minidisc.org/aes_atrac.html

Outline

- Lossless Audio Coding
- Lossy Audio Coding
- MPEG audio (MP3)

MPEG (Moving Picture Expert Group) Audio

- MPEG-1: 1.5 Mbits/sec for audio and video
  - About 1.2 Mbits/sec for video, 0.3 Mbits/sec for audio
  - Cf. Uncompressed CD audio is 44,100 samples/sec * 16 bits/sample * 2 channels > 1.4 Mbits/sec
  - Compression factor ranging from 2.7 to 24.
- With Compression rate 6:1 (16 bits stereo sampled at 48 KHz is reduced to 256 kbits/sec), expert could not distinguish
- Supports sampling frequencies of 32, 44.1 and 48 KHz.
- Supports one or two audio channels in one of the four modes:
  - Monophonic - single audio channel
  - Dual-monophonic - two independent channels, e.g., English and French
  - Stereo - for stereo channels that share bits, but not using Joint-stereo coding
  - Joint-stereo - takes advantage of the correlations between stereo channels

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

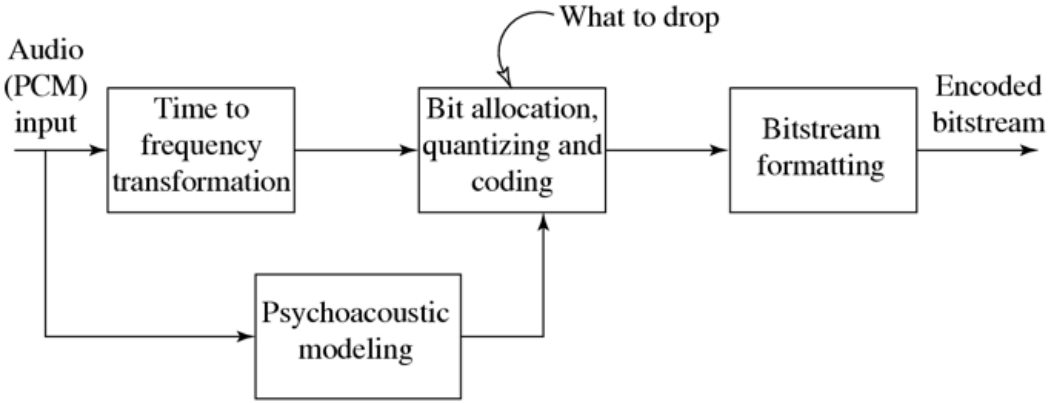
- **MPEG encoder** employs a bank of filters to:
  - Analyze the frequency ("spectral") components of the audio signal by calculating a frequency transform of a window of signal values
  - Decompose the signal into subbands by using a bank of filters (Layer 1 & 2: "quadrature-mirror"; Layer 3: adds a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

Algorithm

- Divide the audio signal (e.g., 48 kHz sound) into 32 frequency subbands --> *subband filtering*.
  - Modified discrete cosine transform (MDCT) -
- Masking for each band caused by nearby band
  - *psychoacoustic model*
  - If the power in a band is below the masking threshold, don't encode it.
  - Otherwise, determine number of bits needed to represent the coefficient such that noise introduced by quantization is below the masking effect
    - One fewer bit introduces about 6 dB of noise).
- Format bitstream



(a) MPEG Audio Encoder

Example

- After analysis, the first levels of 16 of the 32 bands:

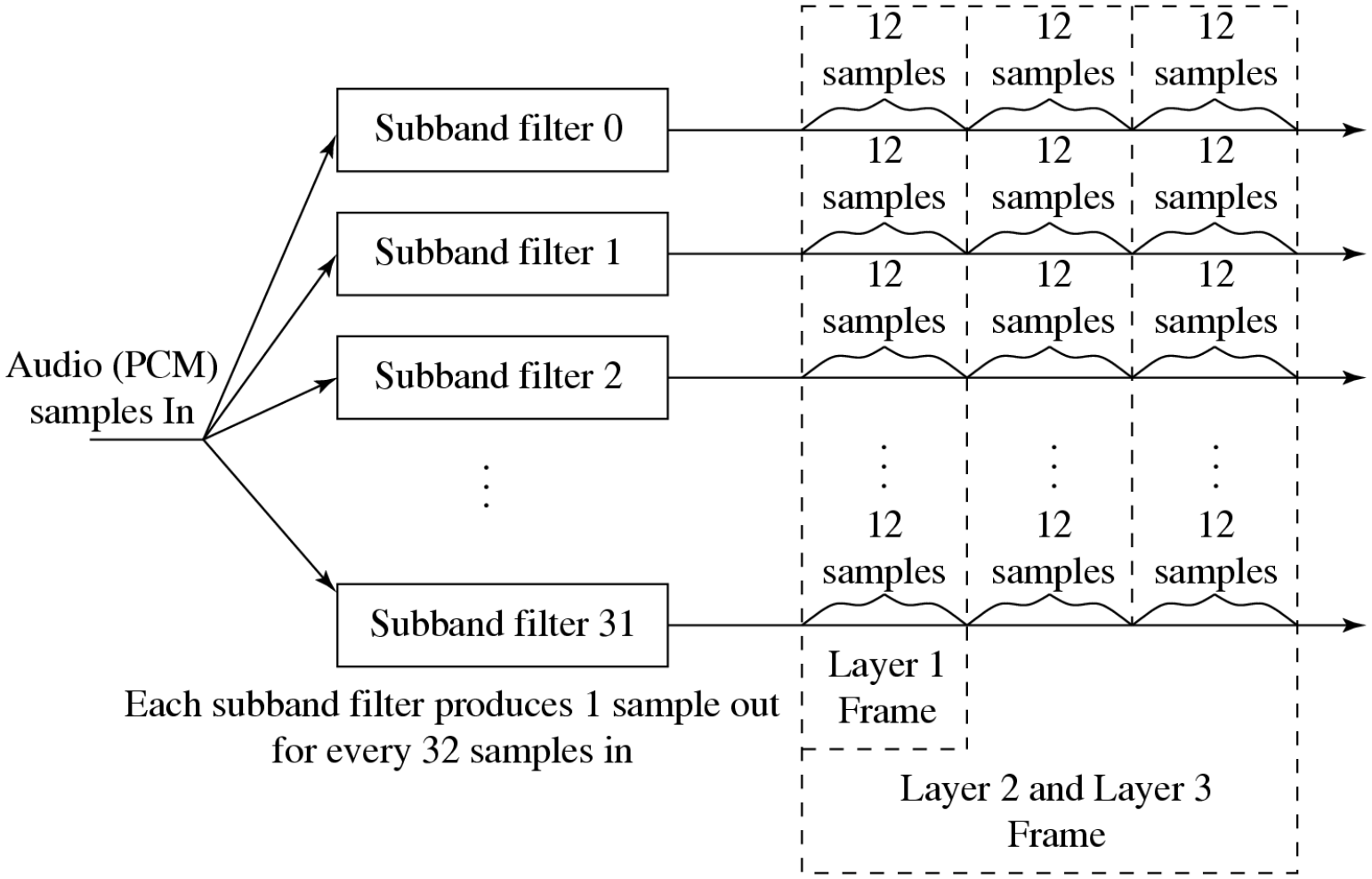
Band	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Level (db)	0	8	12	10	6	2	10	60	35	20	15	2	3	5	3	1

- If the level of the 8th band is 60dB, it gives a masking of 12 dB in the 7th band, 15dB in the 9th.
- Level in 7th band is 10 dB (< 12 dB), so ignore it.
- Level in 9th band is 35 dB (> 15 dB), so send it.

[Only the amount above the masking level needs to be sent, so instead of using 6 bits to encode it, we can use 4 bits -- a saving of 2 bits (12 dB).]

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



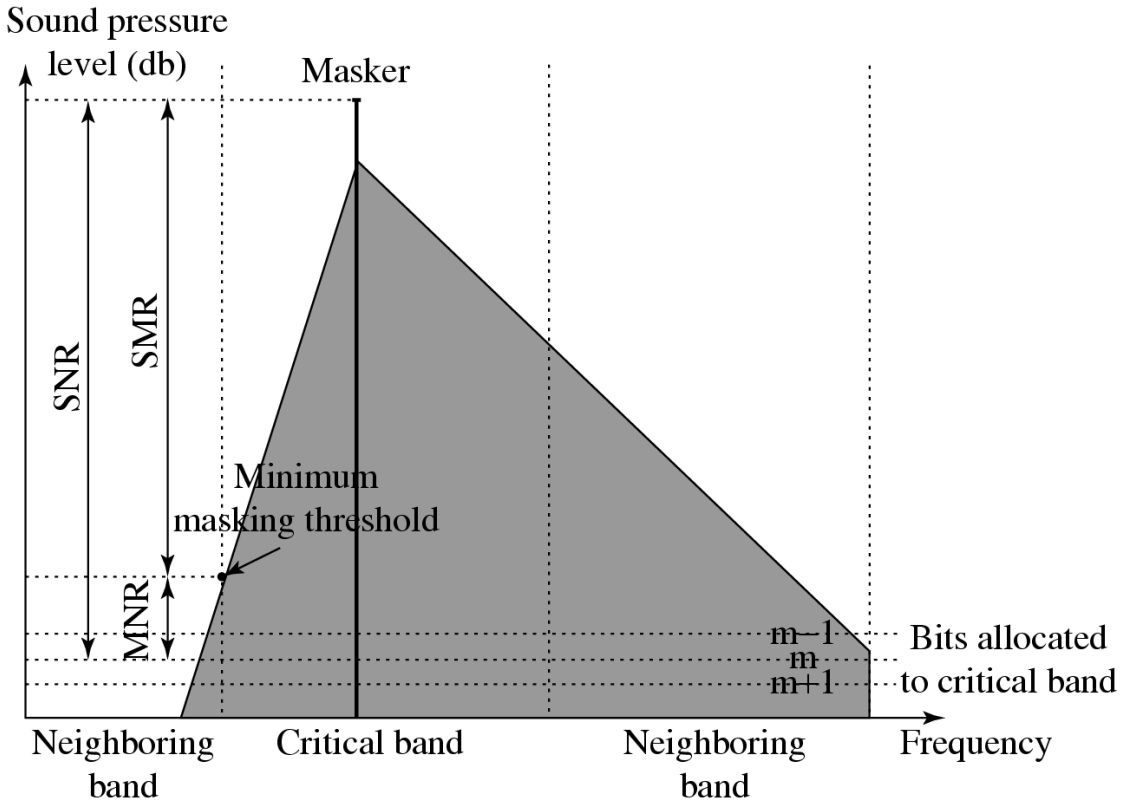
- **Fig. 14.11: MPEG Audio Frame Sizes**

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):

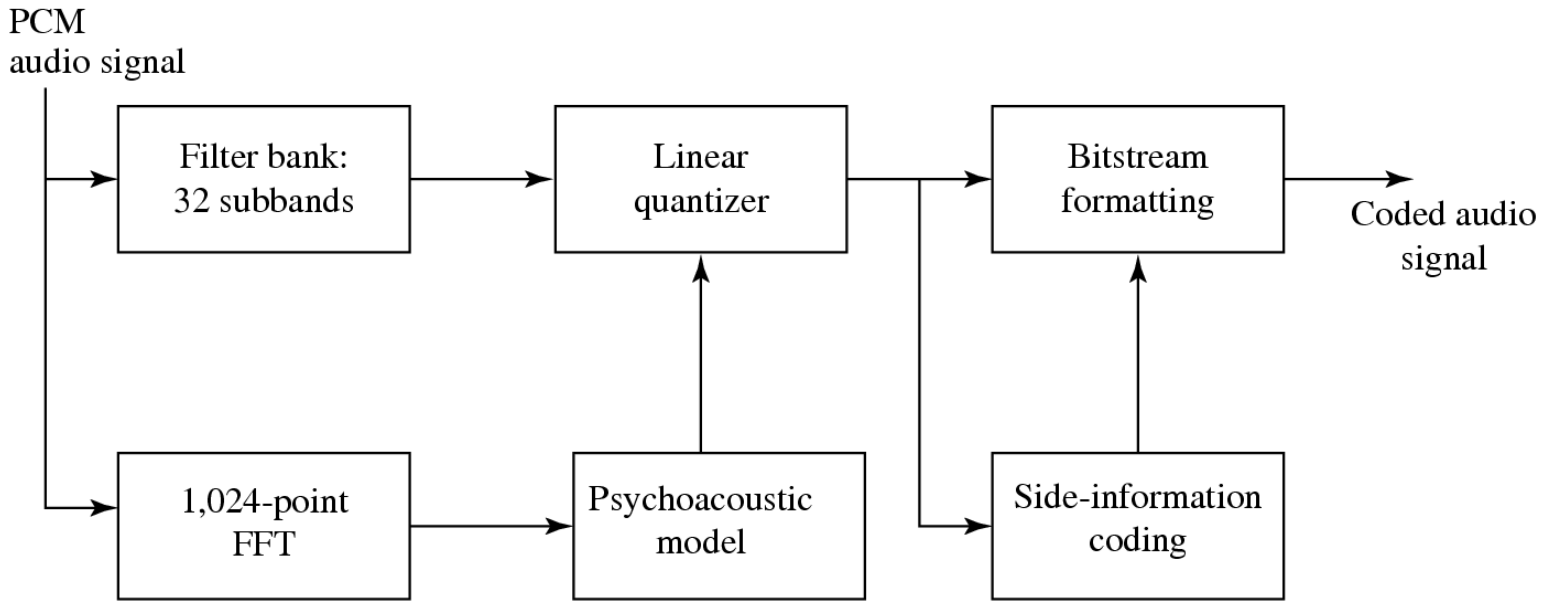
$$\mathrm{MNR}_{\mathrm{dB}} \equiv \mathrm{SNR}_{\mathrm{dB}} - \mathrm{SMR}_{\mathrm{dB}} \qquad (14.6)$$

  - 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

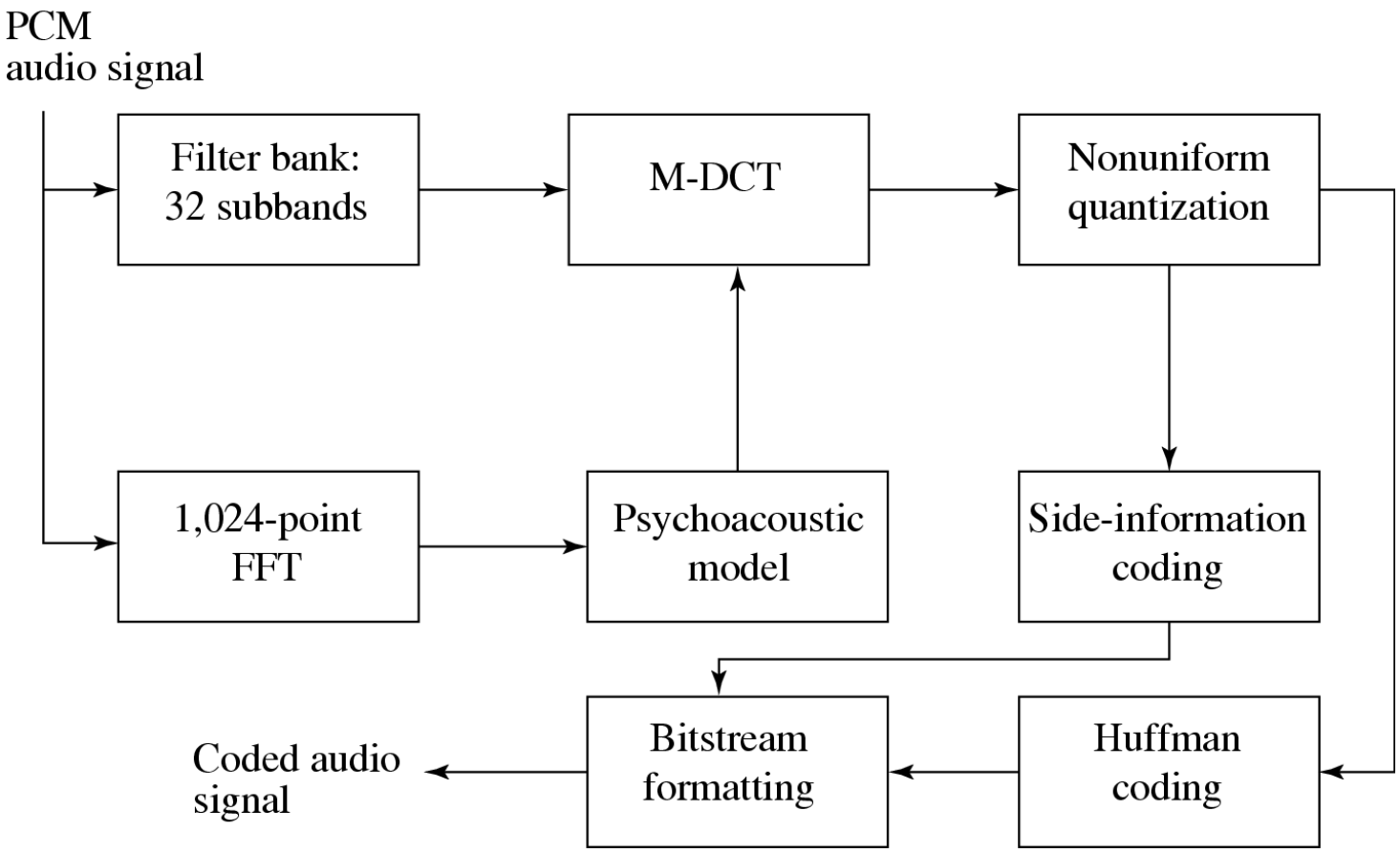
- · **Main difference**:
  - Three groups of 12 samples are encoded in each frame and temporal masking is brought into play, as well as frequency masking
  - Bit allocation is applied to window lengths of 36 samples instead of 12
  - The resolution of the quantizers is increased from 15 bits to 16
- · **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%:

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

  - (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 Audio Layers

- MPEG defines 3 layers for audio. Basic model is same, but codec complexity increases with each layer. .
- Layer 1: DCT type filter with one frame and equal frequency spread per band. Psychoacoustic model only uses frequency masking.
- Layer 2: Three frames in filter (before, current, next, a total of 1152 samples). This models a little bit of the temporal masking.
- Layer 3 (MP3): Better critical band filter is used (non-equal frequencies), psychoacoustic model includes temporal masking effects, takes into account stereo redundancy, and uses Huffman coder.
- **Stereo Redundancy Coding:**
  - Intensity stereo coding -- at upper-frequency subbands, encode summed signals instead of independent signals from left and right channels.
  - Middle/Side (MS) stereo coding -- encode middle (sum of left and right) and side (difference of left and right) channels.

Effectiveness of MPEG Audio

Layer	Target Bit-rate	Ratio	Quality at 64 kb/s	Quality at 128 kb/s	Theoretical Min. Delay
Layer 1	192 kb/s	4:1	---	---	19 ms
Layer 2	128 kb/s	6:1	2.1 to 2.6	4+	35 ms
Layer 3	64 kb/s	12:1	3.6 to 3.8	4+	59 ms

- **Quality factor**: 5 - perfect, 4 - just noticeable, 3 - slightly annoying, 2 - annoying, 1 - very annoying
- **Real delay** is about 3 times of the theoretical delay

Artefacts of compression

- Mp3 encoded recordings rarely sound identical to original uncompressed audio files
- Whole areas of the spectrum are lost in the encoding process
- On small domestic 'hi-fi' or PC speakers, however, mp3 compressed audio can be acceptable

Artefacts of compression

- Sound Test: Difference between WAV and MP3
  - http://www.noiseaddicts.com/2010/04/sound-test-difference-between-wav-vs-mp3/
- Do 320kbps mp3 files really sound better (vs 128kbps) for your ear?
  - www.noiseaddicts.com/2009/03/mp3-sound-quality-test-128-320/

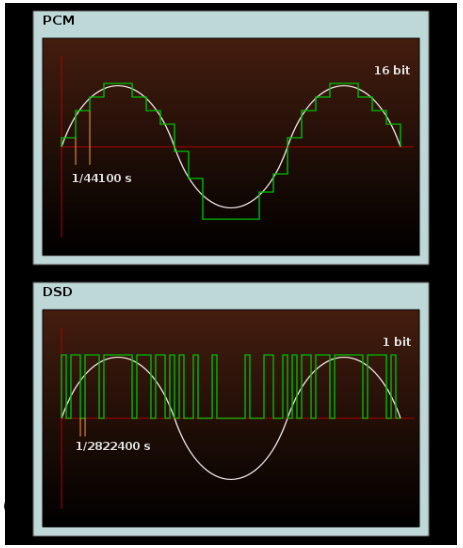
Advanced Audio Coding (AAC)

- An improvement to MP3
  - Part of MPEG-2/4 specifications
  - Standard format for iTunes, DivX, Playstation 3, Android, Blackberry ...
- More sample frequencies (8-96 KHz)
  - 16-48 for MP3
  - Better handling of frequencies above 16 kHz
- Up to 48 channels
  - 2 channel in original MP3 and 5.1 channel in current MP3
- Higher efficiency and accuracy
  - Additional modules/tools to increase compression efficiency

Future of Digital Audio

- Ultra high compression
  - Internet streaming
  - Internet phone
- Ultra high quality
  - CD (compact disc - Redbook)
    - Sony 1976; Philips 1979
  - SACD (super audio CD)
    - Sony/Philips 1999
    - DSD (Direct Stream Digital)
      - 1 bit 2.8224 MHz
      - Equivalent to about PCM 24-bit 176.4 kHz
      - First practical DSD converter implementations by EMM Labs (Calgary, AB, Canada)

	CD	SACD
Format	16 bit PCM	1 bit DSD
Sampling frequency	44.1 kHz	2.8224 MHz
Dynamic range	96 dB	120 dB
Frequency range	20 Hz – 22.05 kHz	20 Hz – 50 kHz
Disc capacity	700 MB	7.95 GB
Stereo	Yes	Yes



Future of Digital Audio cont'd

- Ultra high quality
  - CD (compact disc - Redbook)
    - Sony 1976; Philips 1979
  - SACD (super audio CD)
    - Sony/Philips 1999
  - DVD-A (DVD Audio)
    - Toshiba 2000
    - 8.5GB disc
    - A war with SACD (HDDVD/BluRay later)
    - Neither very successful in consumer market

	16-, 20- or 24-bit (DVD-A)					
	44.1 kHz	48 kHz	88.2 kHz	96 kHz	176.4 kHz	192 kHz
Mono (1.0)	Yes	Yes	Yes	Yes	Yes	Yes
Stereo (2.0)	Yes	Yes	Yes	Yes	Yes	Yes
nStereo (2.1)	Yes	Yes	Yes	Yes	No	No
Stereo + mono surround (3.0 or 3.1)	Yes	Yes	Yes	Yes	No	No
Quad (4.0 or 4.1)	Yes	Yes	Yes	Yes	No	No
3-stereo (3.0 or 3.1)	Yes	Yes	Yes	Yes	No	No
3-stereo + mono surround (4.0 or 4.1)	Yes	Yes	Yes	Yes	No	No
Full surround (5.0 or 5.1)	Yes	Yes	Yes	Yes	No	No

Future of Digital Audio cont'd

	SACD	DVD-A	CD
Sampling Rate	2.8224 MHz (64 x 44.1 kHz)	Up to 192 kHz	44.1 kHz
Amount of Bits	1 bit	16 to 24 bits	16 bits
Dynamic Range (Maximum)	120 dB	104 to 108 dB	96 dB
Copy Protection	Yes	Yes	No
Multi Channel Capability	Yes	Yes	Not good support

- Next ?
  - Disc-less
    - USB flashmemory
    - Lady Gaga "Born This Way" £40 (4x online downloading price)
      - 2GB USB-drive: 1GB of content - 17 tracks, six remixes, music videos, photo gallery and special feature for Little Monsters membership card holders
  - Digital download
    - MP3/AAC – APE/FLAC 16 bit 44.1 KHz - 24 bit 96/192 KHz

      High resolution audio samples:

      http://www.2l.no/hires/index.html

      http://www.linnrecords.com/linn-downloads-testfiles.aspx