ECE160 / CMPS182
Multimedia

Lecture 14: Spring 2007
MPEG Audio Compression

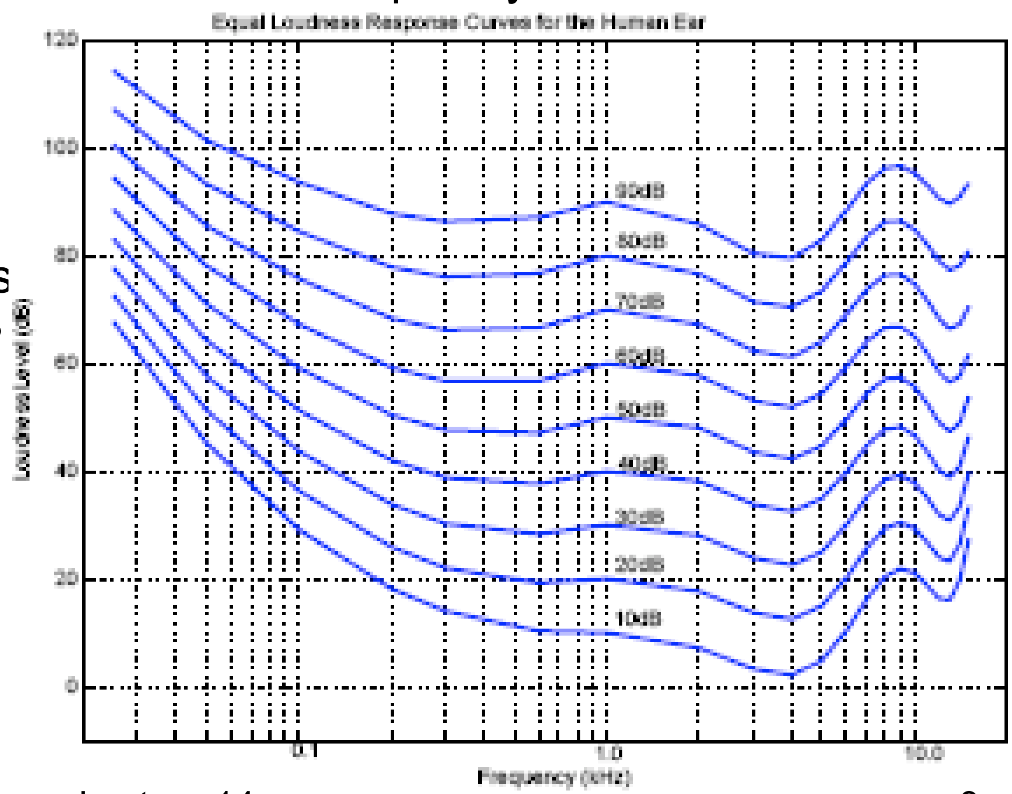
Psychoacoustics

- 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

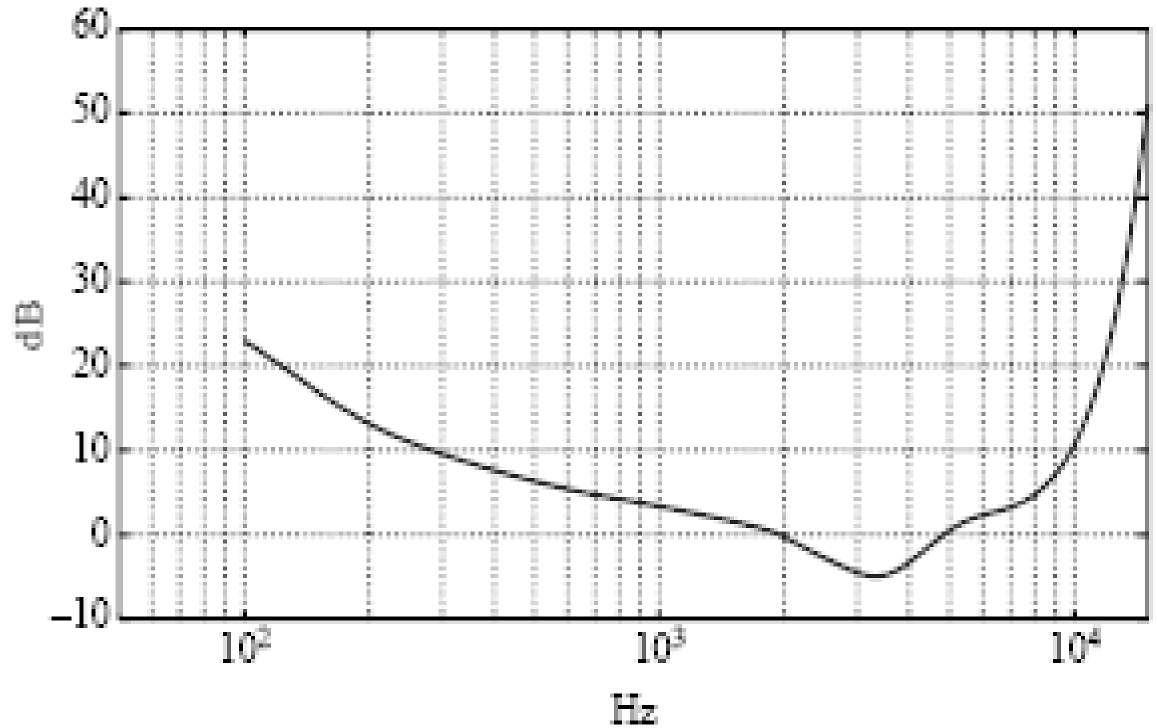
Fletcher-Munson Curves

- 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
- 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



Threshold of Hearing

- Threshold of human hearing, for pure tones: 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:



Frequency Masking

- Lossy audio data compression methods, such as MPEG/Audio encoding, do not encode 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

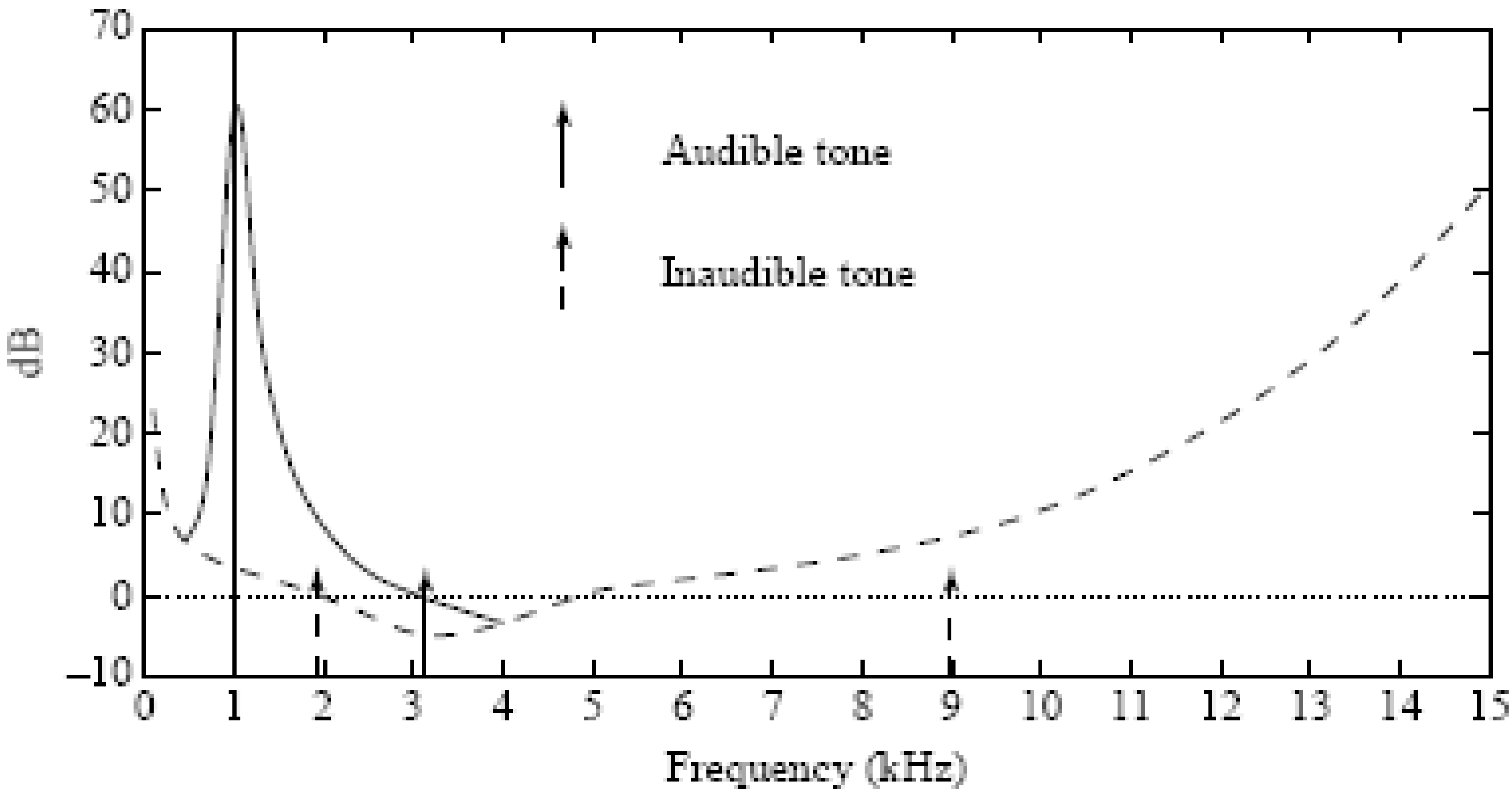
  3. The greater the power in the masking tone, the wider is its influence - the broader the range of frequencies it can mask.

  4. As a consequence, if two tones are widely separated in frequency then little masking occurs

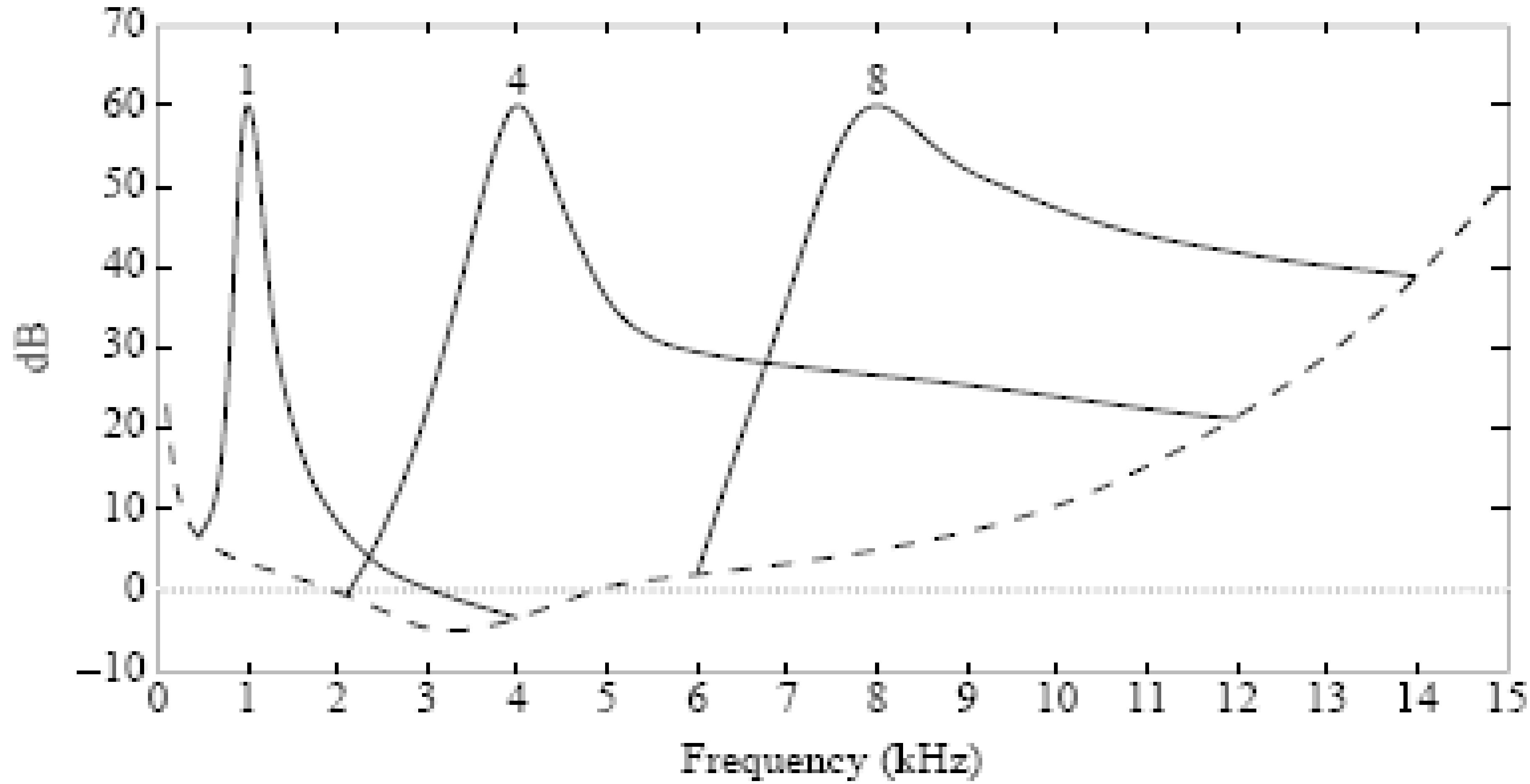
Frequency Masking Curves

- 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 plots the audible level for a single masking tone (1 kHz) and a single sound level
- The plot changes if other masking frequencies or sound levels are used.

Frequency Masking Curve



Frequency Masking Curve



Critical Bands

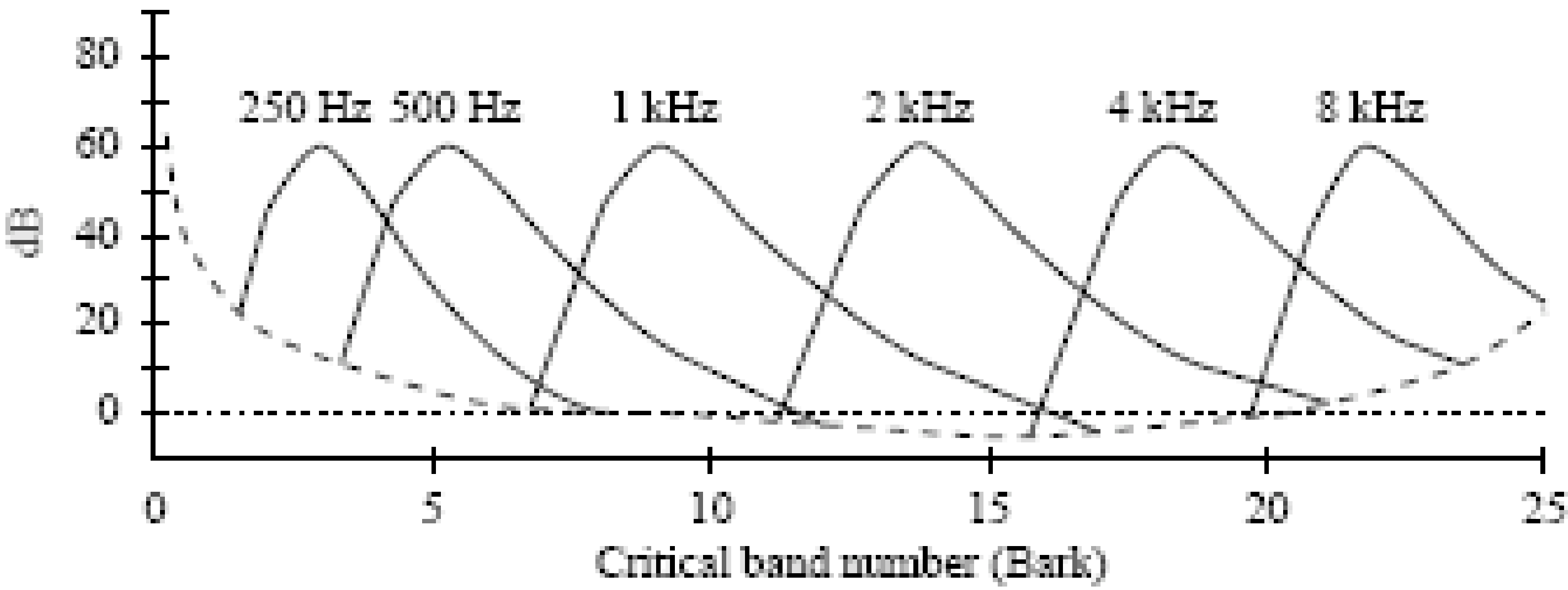
- **Critical bandwidth** represents the ear's resolving power for simultaneous tones or partials
  - At the low-frequency end, a critical band is less than 100 Hz wide, while for high frequencies the width can be greater than 4 kHz
- Experiments indicate that the critical bandwidth:
  - for masking frequencies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

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

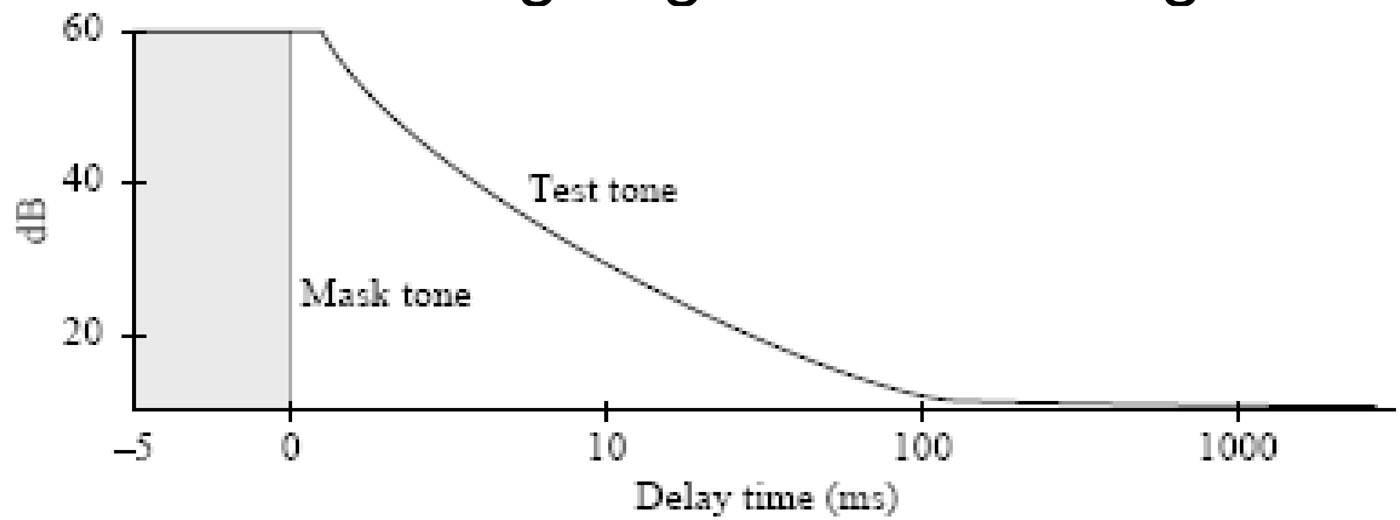
Bark Unit

- **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

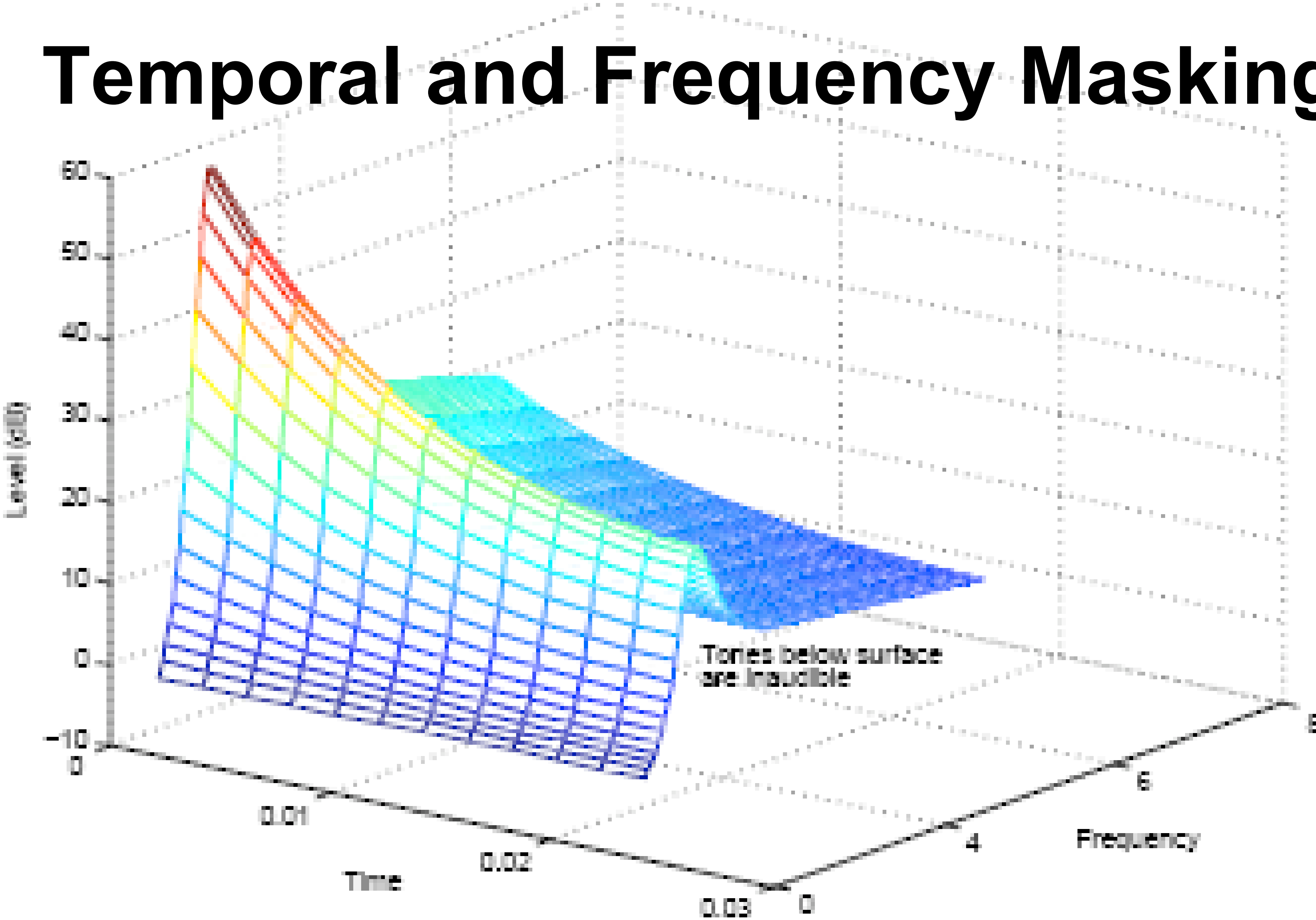


Temporal Masking

- **Phenomenon**: any loud tone will cause the hearing receptors in the inner ear to become *saturated* and require time to recover
- The louder is the test tone, the shorter it takes for our hearing to get over hearing the masking.

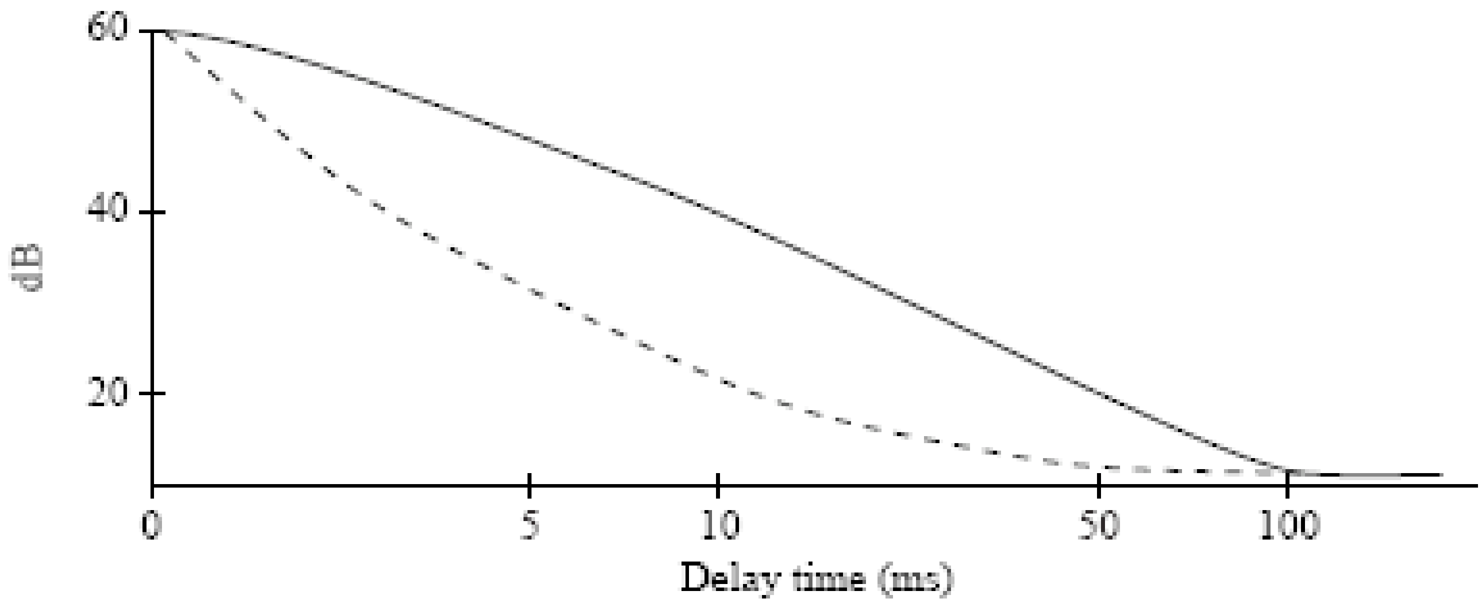


Temporal and Frequency Masking



Temporal and Frequency Masking

- 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.



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

- 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
- Human auditory system is not accurate within the width of a critical band (perceived loudness and audibility of a frequency)

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

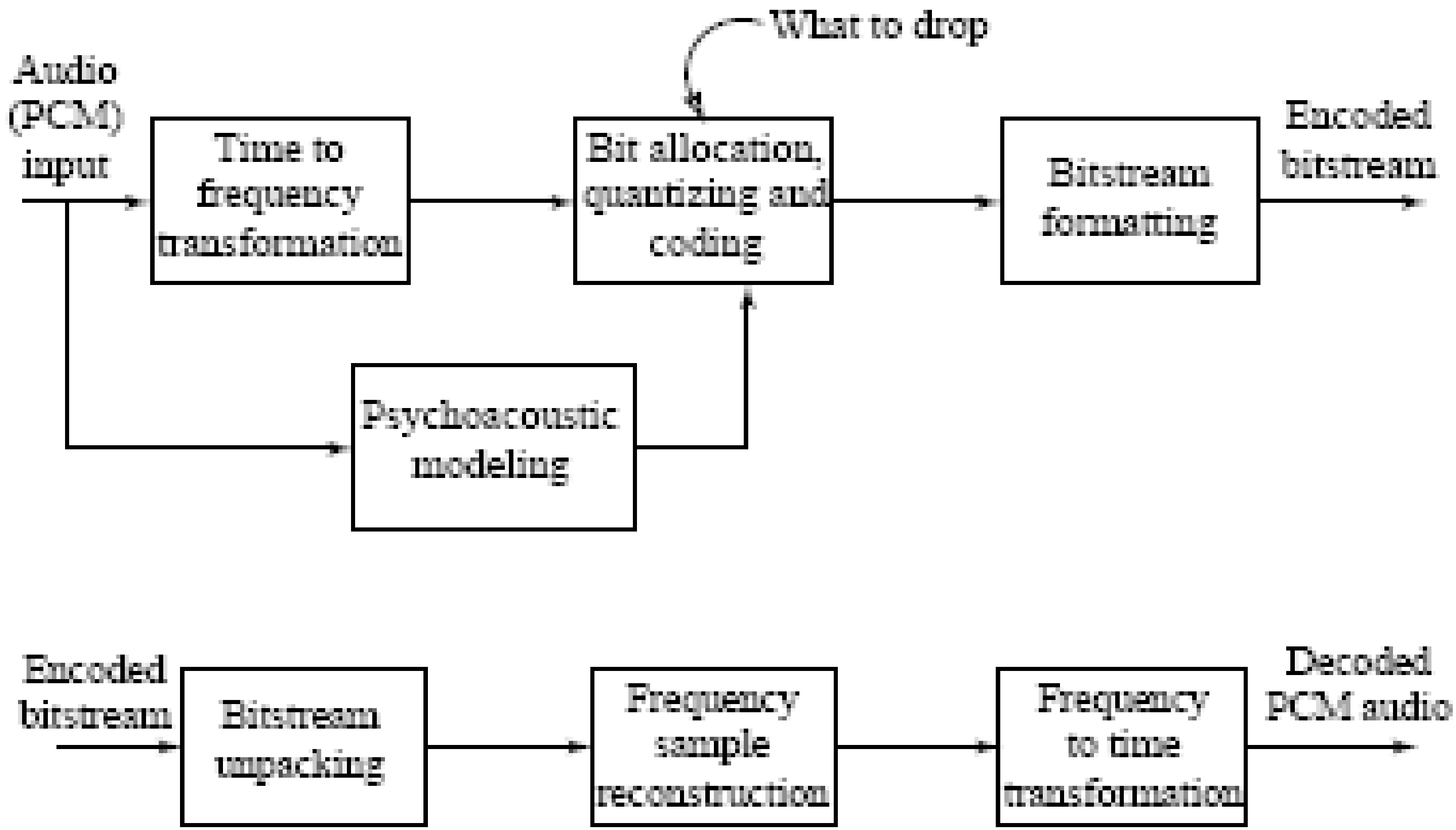
  (Layer 1 & 2: “quadrature-mirror";

  Layer 3: adds a DCT; psychoacoustic model: Fourier transform)

MPEG Audio Strategy

- **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
  - For simplicity, adopts a *uniform width* for all frequency analysis filters, using 32 overlapping subbands

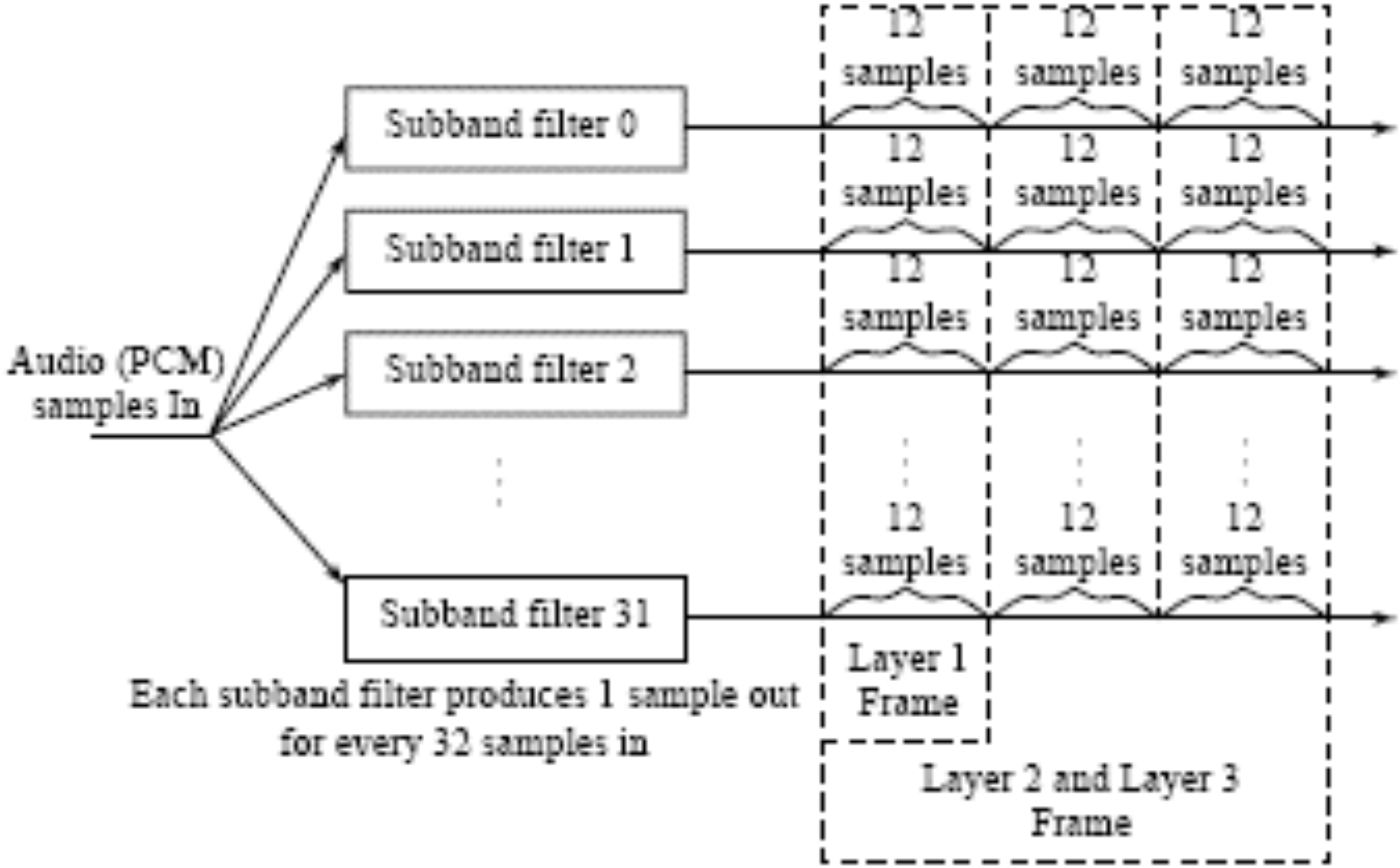
MPEG
Audio Compression Algorithm



MPEG
Audio Compression Algorithm

- 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., 12x32) samples
- A Layer 2 or Layer 3, frame actually accumulates more than 12 samples for each subband: a frame includes 1,152 samples

MPEG Audio Compression Algorithm



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

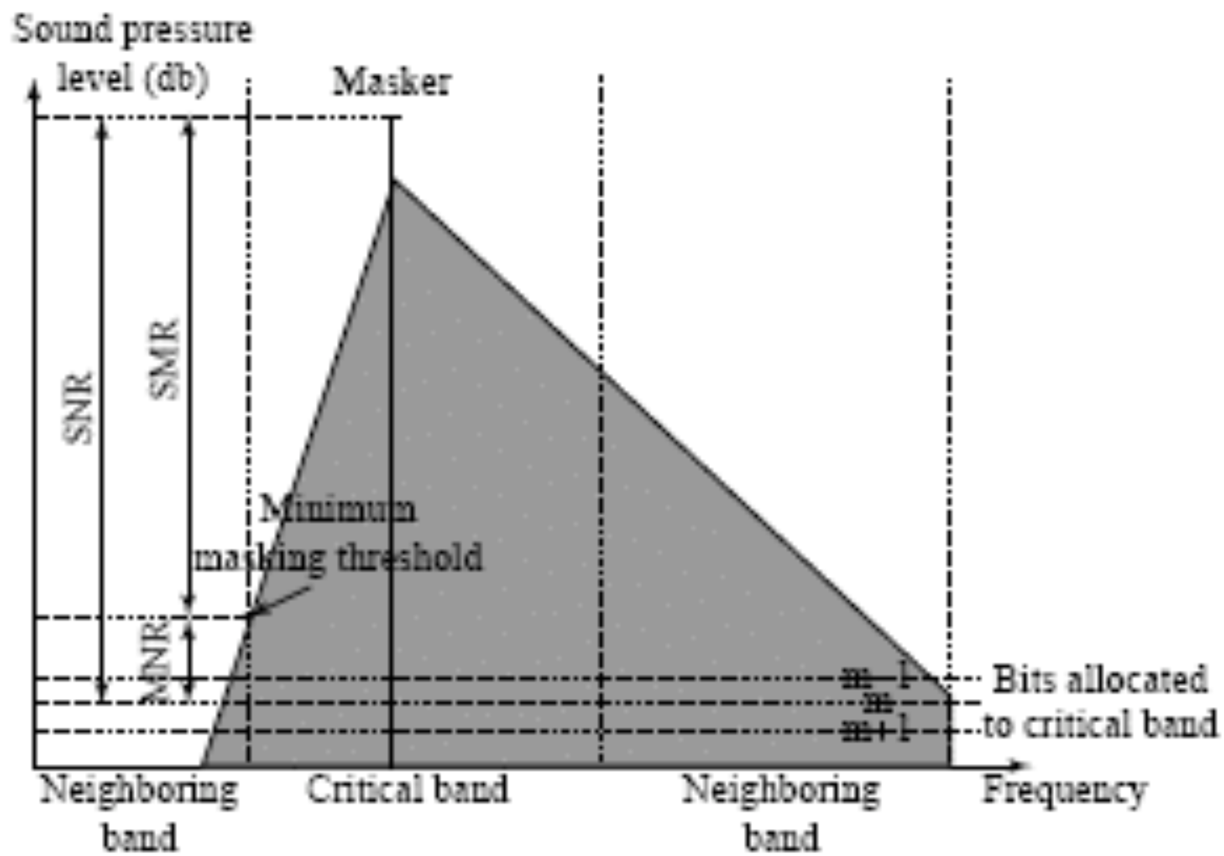
$$MNR_{dB} = SNR_{dB} - SMR_{dB}$$

- 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

Bit Allocation Algorithm

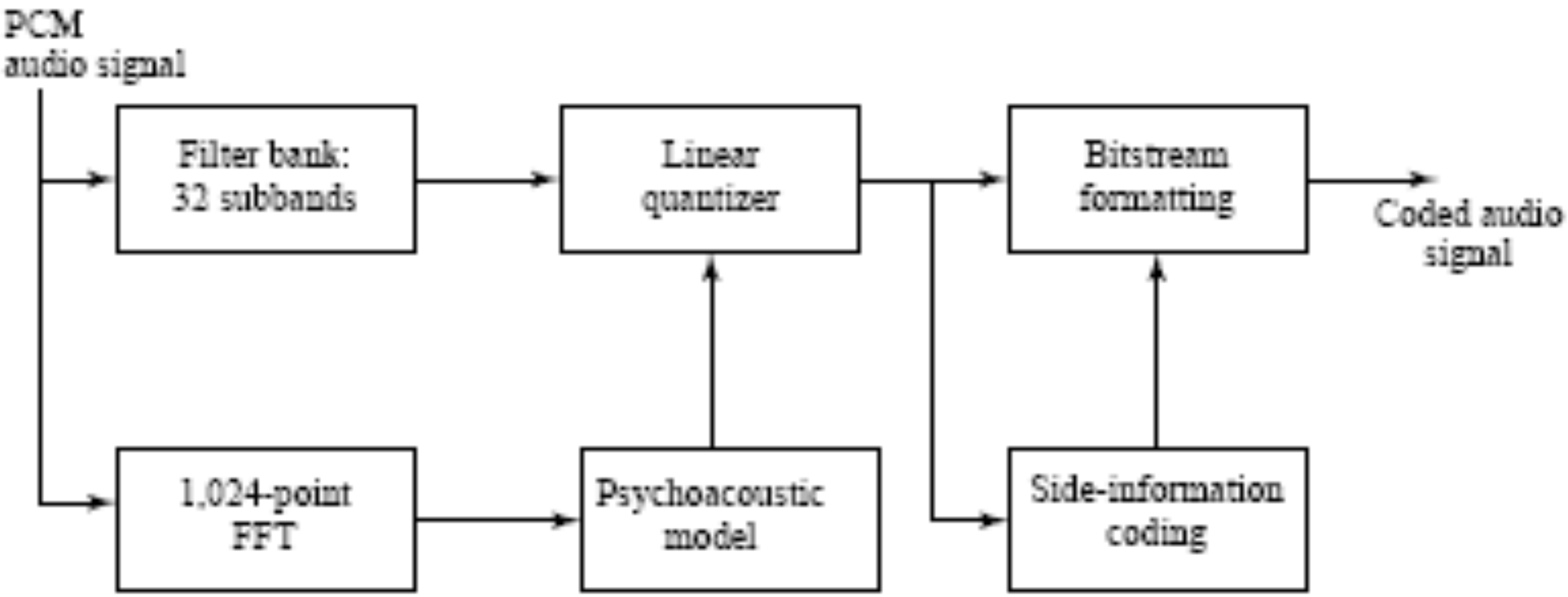
A qualitative view of SNR

SMR and MNR are shown, with one dominant masker and *m* bits allocated to a particular critical band.



MPEG Layers 1 and 2

- Mask calculations are performed in parallel with subband filtering



Layer 2 of MPEG 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

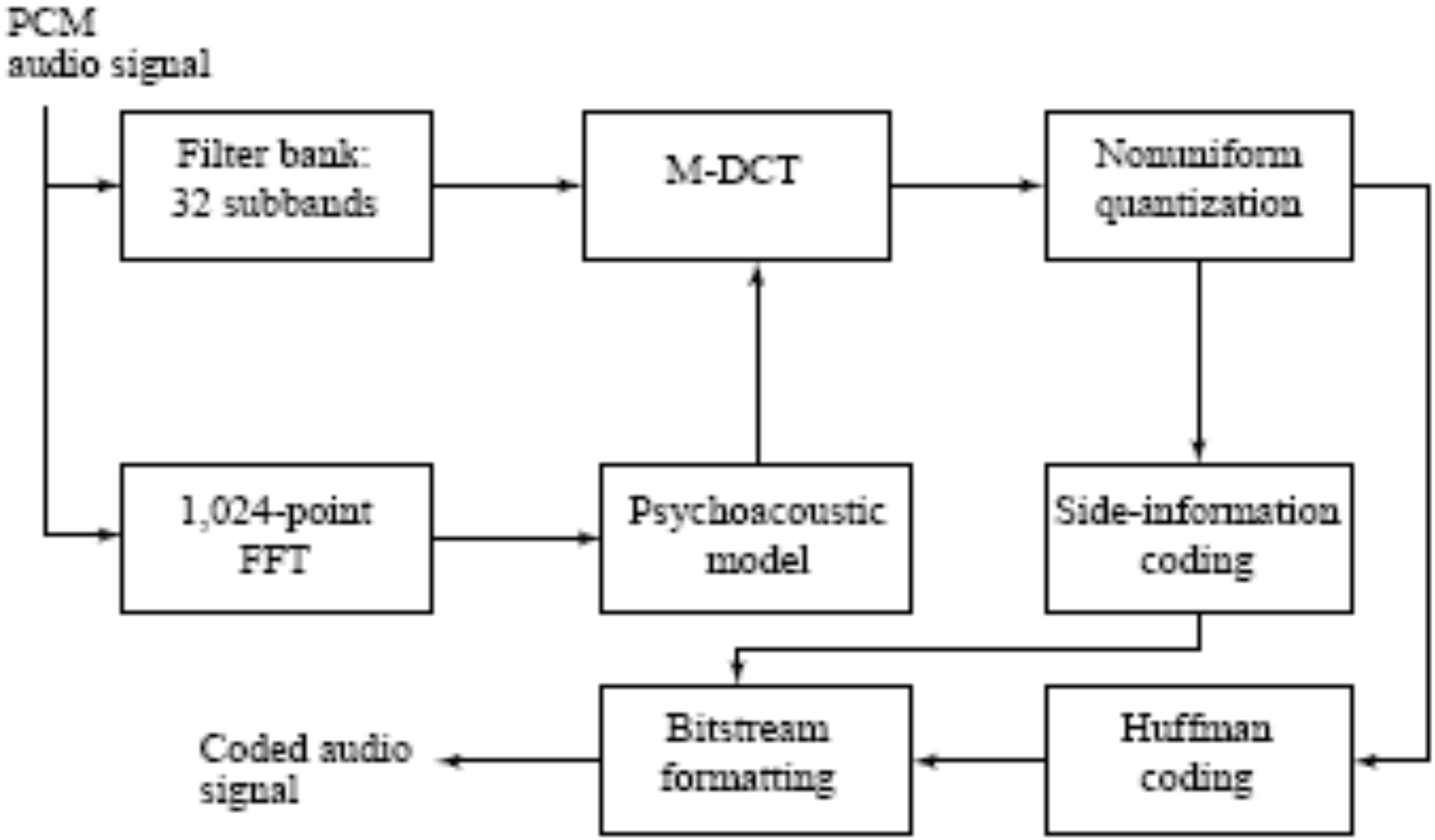
Layer 3 of MPEG Audio

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%:

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

MPEG Layer 3 Coding



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)

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

MPEG-2 AAC

- Also capable of delivering high-quality stereo sound at bit-rates below 128 kbps
- 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

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

Other Commercial Audio Codecs

Codec	Bit-rate kbps/channel	Complexity	Main Application
Dolby AC-2	128-192	low (en-/decoder)	p-to-p, cable
Dolby AC-3	32-640	low (decoder)	HDTV, cable, DVD
Sony ATRAC	140	low (en-/decoder)	minidisc

MPEG-7 and MPEG-21

MPEG-7: A means of standardizing meta-data for audiovisual multimedia sequences - meant to represent information about multimedia information

– In terms of audio: facilitate the representation and search for sound content. Example application supported by MPEG-7: *automatic speech recognition* (**ASR**).

MPEG-21: Ongoing effort, aimed at driving a standardization effort for a Multimedia Framework from a consumer's perspective, particularly *interoperability*

– In terms of audio: support of this goal, using audio.

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