VORBIS AUDIO CODEC

By

Bharan, Prasad, Kulin and Sanjay

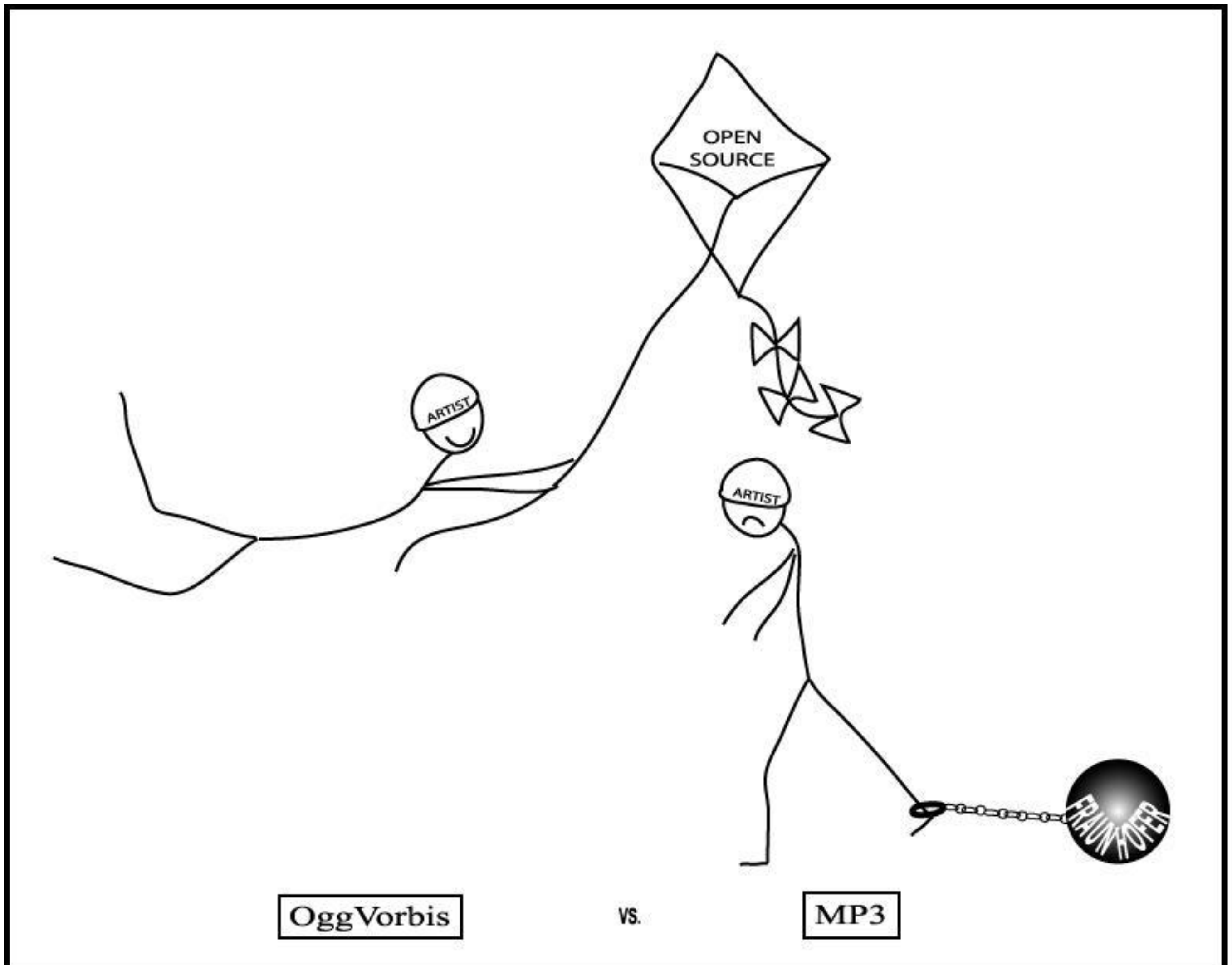
Bird's Eye View

- Currently a growing audio format (internet streaming)
- Open-Source, Non-Proprietary, Patent-Free, and Royalty-Free
- Lossy audio codec – Lossy (psychoacoustic)?, Lossless (entropy)?
- Good all-round performance (>48 kbps - a leading codec at 128 kbps)
  - Same competitive class as MPEG-4 (AAC), and higher performance than MP3, WMA.
- Well written specifications and documentation (from the decoder's point of view)
- High potential for further tuning (Vorbis II)
- Supported by most portable Digital Audio Players (DAPs)

Background

- Result of making MP3 proprietary by Fraunhofer Institute, Germany
- First of a family of codec by xiph.org
- Ogg= general purpose container stream format
- Vorbis= psychoacoustic audio codec
- Ogg comes from a Netrek term ‘ogging’
- Vorbis – Discworld character- High priest – Small Gods

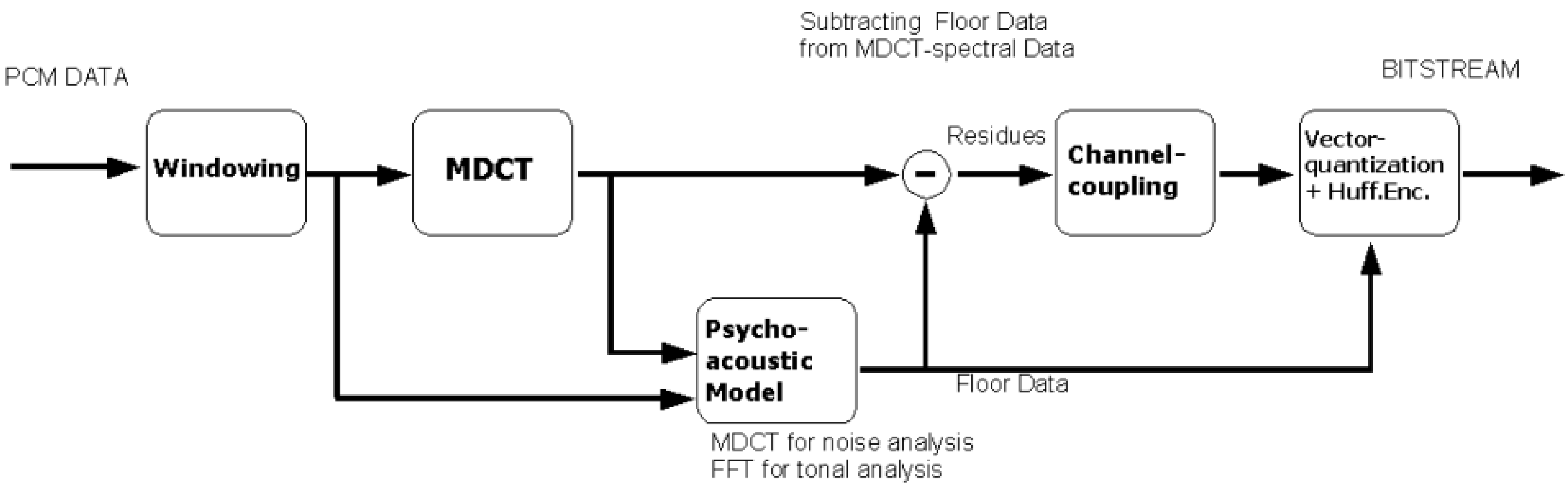
Ogg or MP3?



Speech Vs. Audio CODEC

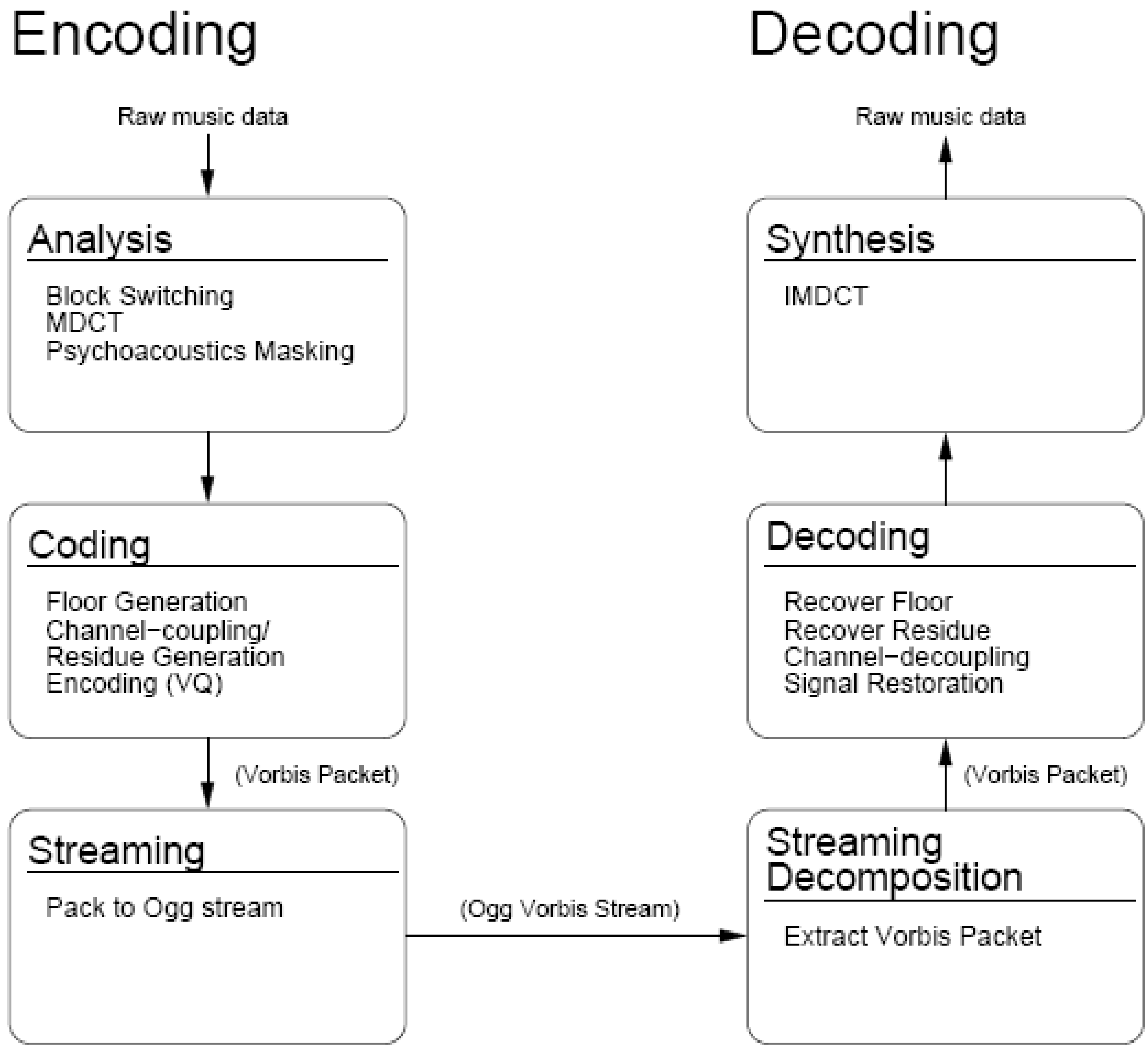
- **Speech Codec**
  - Designed to deal with characteristics of voice.
  - Speech has maximum frequency of 7 kHz. It looks for speech patterns and tries to compress data further.
  - Only intelligibility of speech is important for codecs and hence a lot of statistical info. related to emotions, etc. can be dropped. They has comparatively lower bit rate.
  - Ex. u-law PCM
- **Audio Codec**
  - Audio codec are developed for music (Audio signals).
  - Audio signals have comparatively larger frequency range of 20 Hz to 20 kHz.
  - They have comparatively higher bit rate.
  - Ex. Ogg vorbis

Vorbis Audio Codec



Channel Coupling

- Eliminates inter-channel redundancies
- Eliminates stereo image info labelled unimportant by the psychoacoustic model
- Two types
  - Channel interleaving
  - Square polar mapping
- Vorbis supports multi channels (>2, ex. Vorbis 5.1 5-channel surround)


Encoding
Decoding
Raw music data
Analysis
Block Switching
MDCT
Psychoacoustics Masking
Coding
Floor Generation
Channel−coupling/
Residue Generation
Encoding (VQ)
(Vorbis Packet)
Streaming
Pack to Ogg stream
(Ogg Vorbis Stream)
Raw music data
Synthesis
IMDCT
Decoding
Recover Floor
Recover Residue
Channel−decoupling
Signal Restoration
(Vorbis Packet)
Streaming
Decomposition
Extract Vorbis Packet

MDCT(MLT)

- Most expensive step as far as Vorbis decoding is concerned
- Maps an K real nos. to an array of K/2 real nos.
- Used in most second generation audio codecs including MP3 and Ogg Vorbis
- Basically DCT with 50% overlap window

MDCT Equations

$$(\overrightarrow{\mathcal{F}_M}\, x)[m] = \sum_{k=0}^{K-1} x[k] \cos\left(\frac{2\pi}{K}(k+d)(m+\frac{1}{2})\right)$$

$$(\overleftarrow{\mathcal{F}_M}\, X)[j] = \frac{4}{K} \sum_{m=0}^{K/2-1} X[m] \cos\left(\frac{2\pi}{K}(j+d)(m+\frac{1}{2})\right)$$

MDCT followed by IMDCT - ?

$$(\overleftarrow{\mathcal{F}_M}\, \overrightarrow{\mathcal{F}_M}\, x)[j] = \begin{cases} x[j] - x[\frac{1}{2}K - j - 1], & j < K/2 \\ x[j] + x[\frac{3}{2}K - j - 1], & K/2 \leq j \end{cases}$$

Prove it…

MDCT Equations

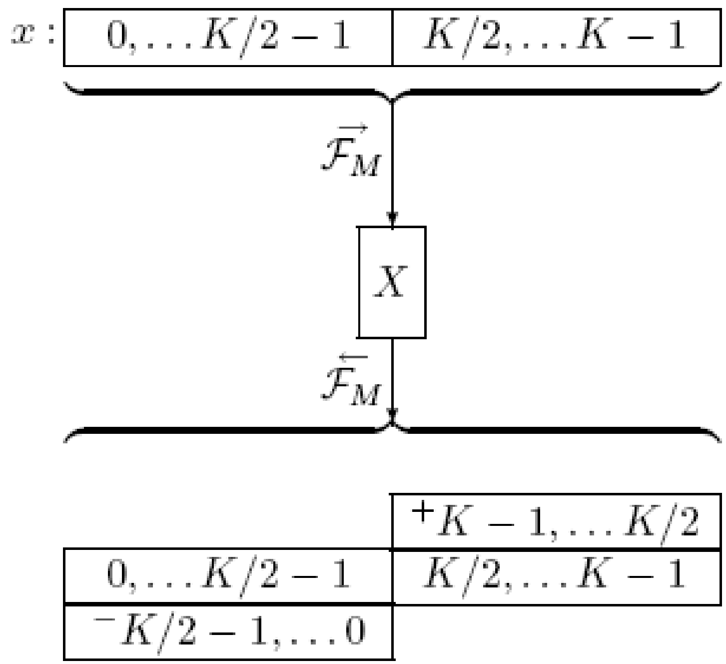
$$\sum_{n=0}^{N-1} \cos(nx) = \begin{cases} \frac{1}{2} + \frac{\sin((N-\frac{1}{2})x)}{2\sin(x/2)} & \textit{if } x \bmod 2\pi \neq 0, \\ N & \textit{if } x \bmod 2\pi = 0 \end{cases}$$

$$\sum_{n=0}^{N-1} \cos((n+\frac{1}{2})x) = \begin{cases} \frac{\sin(Nx)}{2\sin(x/2)} & \textit{if } x \bmod 2\pi \neq 0, \\ N & \textit{if } x/2\pi \textit{ is even} \\ -N & \textit{if } x/2\pi \textit{ is odd} \end{cases}$$

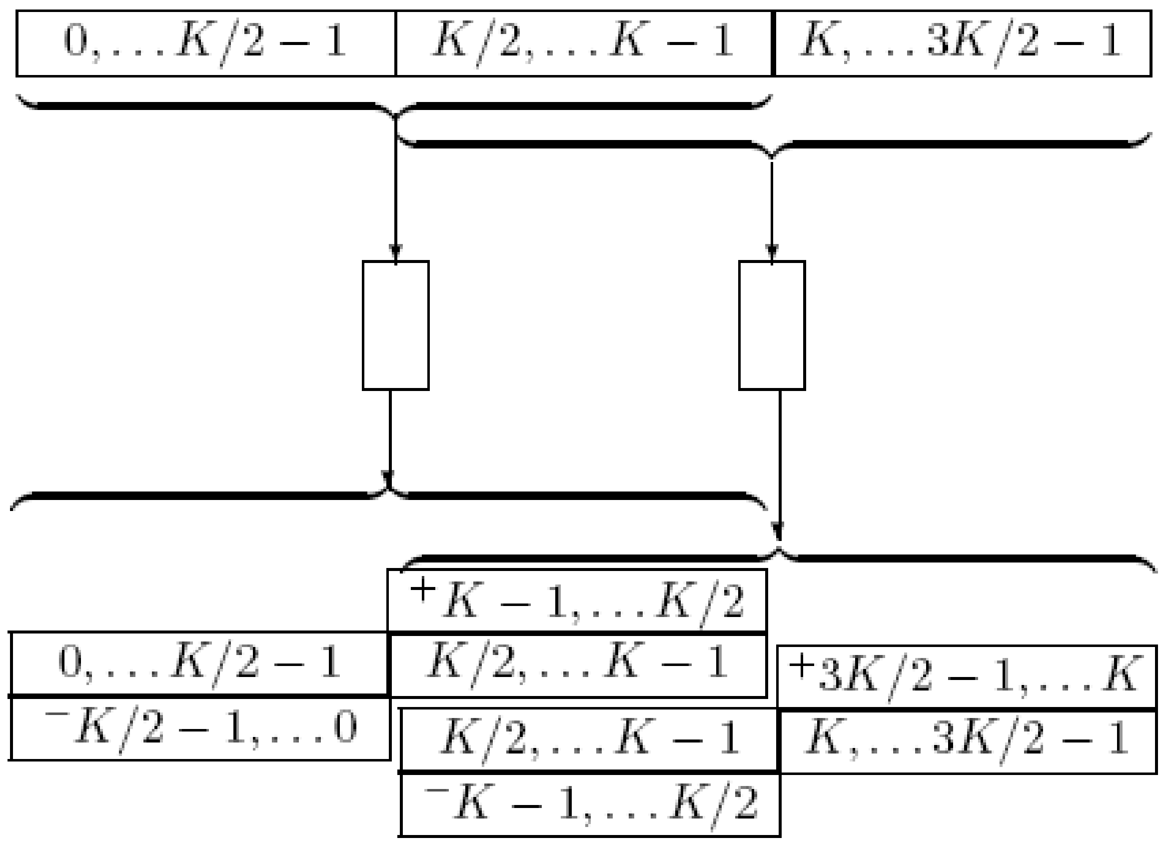
First prove: assume j>=0 and k<K

$$\sum_{m=0}^{K/2-1} \cos\left(\frac{2\pi}{K}(k+d)(m+\frac{1}{2})\right) \cos\left(\frac{2\pi}{K}(j+d)(m+\frac{1}{2})\right)$$
$$= \frac{K}{4}\left([\![k=j]\!] - [\![k=K-2d-j]\!] + [\![k=2K-2d-j]\!]\right)$$

Result of MDCT followed by IMDCT



The beauty of MDCT- TDAC



$$(\overleftarrow{\mathcal{F}_M}\ \overrightarrow{\mathcal{F}_M}\, x_0)[j+K/2] + (\overleftarrow{\mathcal{F}_M}\ \overrightarrow{\mathcal{F}_M}\, x_1)[j] = 2x[j+K/2]$$

So what’s the Big deal?

- Could have used FFT on non overlapping window?? Issues? – ‘ Block Artifacts’

(a) Original image

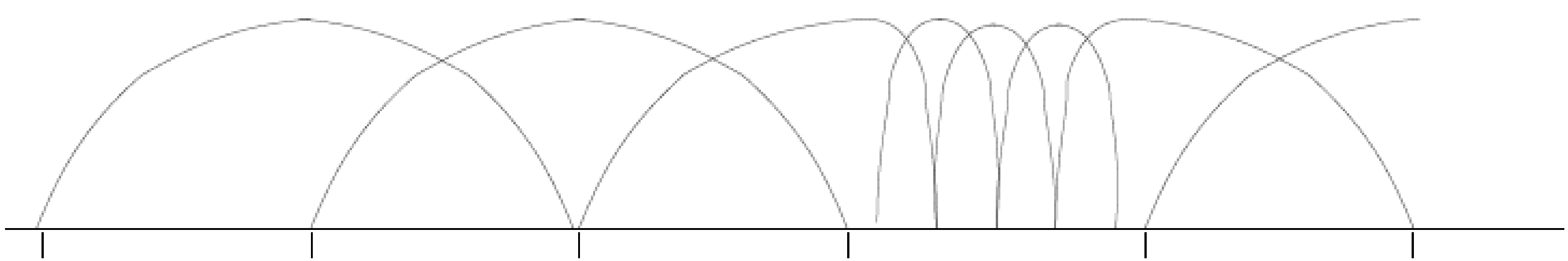
(b) 20x compression using the DCT

(c) 20x compression using the DFT (FFT)

- Multiply by window before and after the transform – known as ‘block switching’
- Long window- 2048 samples, short window- 256 samples

Windowing results

- To cancel alias:
  - Shape of windows in succeeding blocks must fit to each other only in the overlapping part
  - Longer- shorter split
  - Symmetry in each half of window



Windowing results

$$f_0[j+K/2](\overleftarrow{\mathcal{F}_M}\,\overrightarrow{\mathcal{F}_M}\,x_0)[j+K/2] + f_1[j](\overleftarrow{\mathcal{F}_M}\,\overrightarrow{\mathcal{F}_M}\,x_1)[j]$$

$$\begin{aligned}
&= f_0[j+K/2]x_0[j+K/2] \\
&\quad + f_0[j+K/2]x_0[K-j-1] \\
&\quad + f_1[j]x_1[j] - f_1[j]x_1[K/2-j-1] \\
&= f_0[j+K/2]h_0[j+K/2]x[j+K/2] \\
&\quad + f_0[j+K/2]h_0[K-j-1]x[K-j-1] \\
&\quad + f_1[j]h_1[j]x[j+K/2] \\
&\quad - f_1[j]h_1[K/2-j-1]x[K-j-1]
\end{aligned}$$

$$f_0[j+K/2]h_0[j+K/2] + f_1[j]h_1[j] = 1$$

$$f_0[j+K/2]h_0[K-j-1] - f_1[j]h_1[K/2-j-1] = 0$$

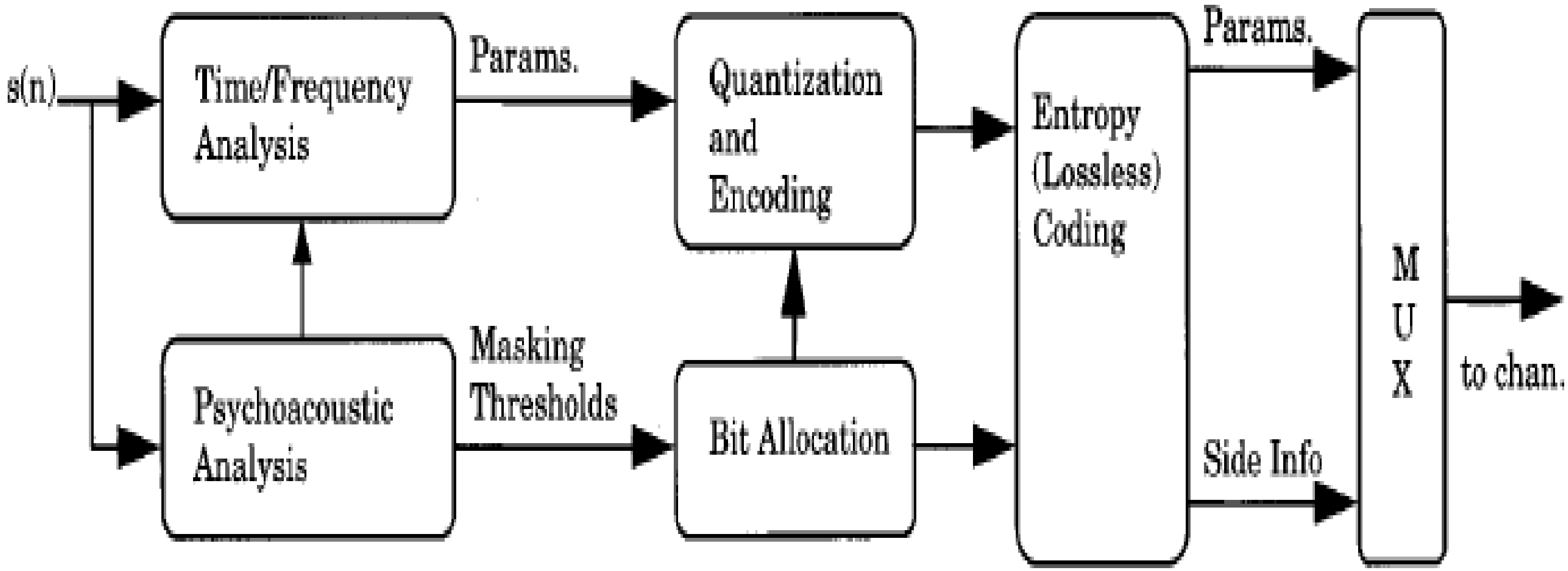
Windowing results

- The window used in vorbis is

$$f[j] = h[j] = \sin\left(\frac{\pi}{2}\sin^2\left(\pi\,\frac{j+\frac{1}{2}}{K}\right)\right)$$

- Better stop-band attenuation at the expense of lesser pass-band selectivity
- You can verify that it satisfies the previous set of eqns.

GENERIC PSYCHOACOUSTIC CODER

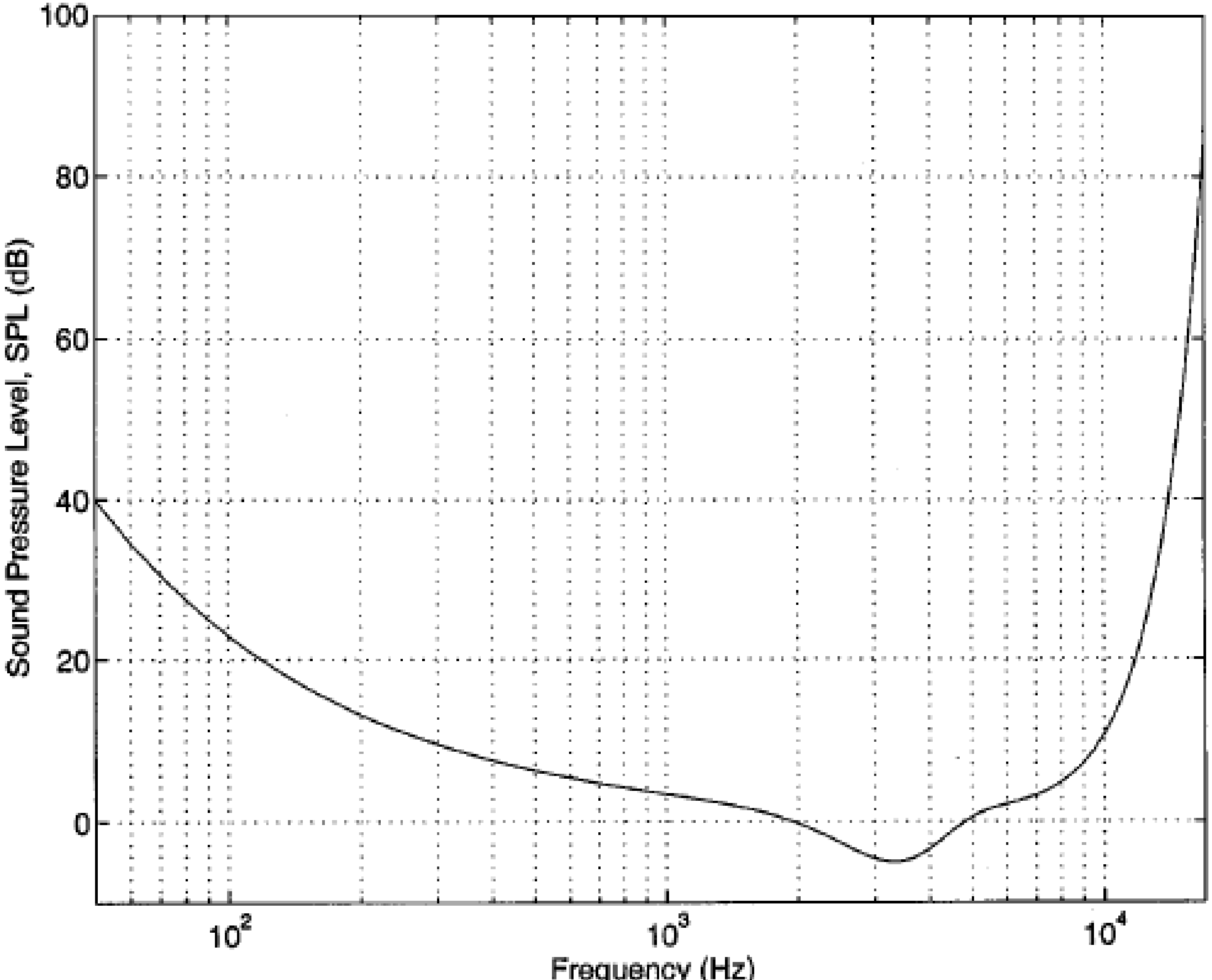


.. Generic perceptual audio encoder.

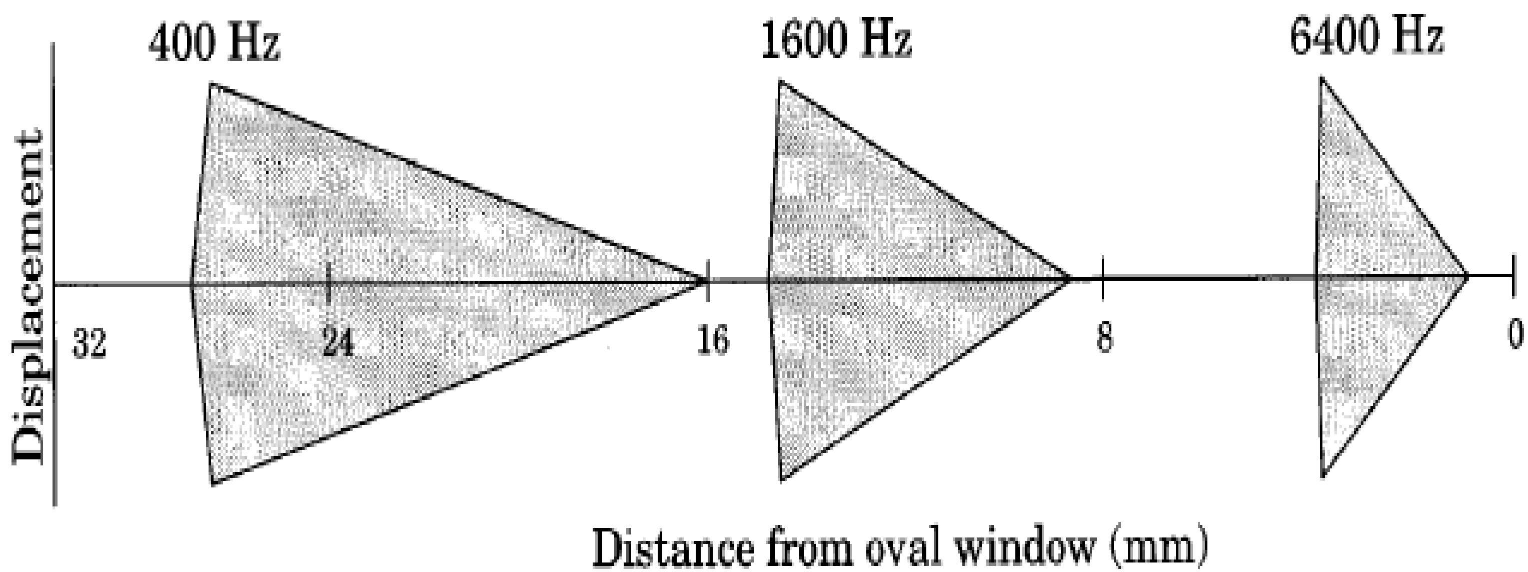
PSYCHOACOUSTIC PRINCIPLES

- Removing the “irrelevant” signal information
- Psychoacoustic metrics are
  - Absolute threshold of hearing
  - Critical bands
    - Neural receptors in Cochlea(inner ear) containing basilar membrane
    - Bank of highly overlapped BPF
    - Non-uniform bandwidth increases with frequency

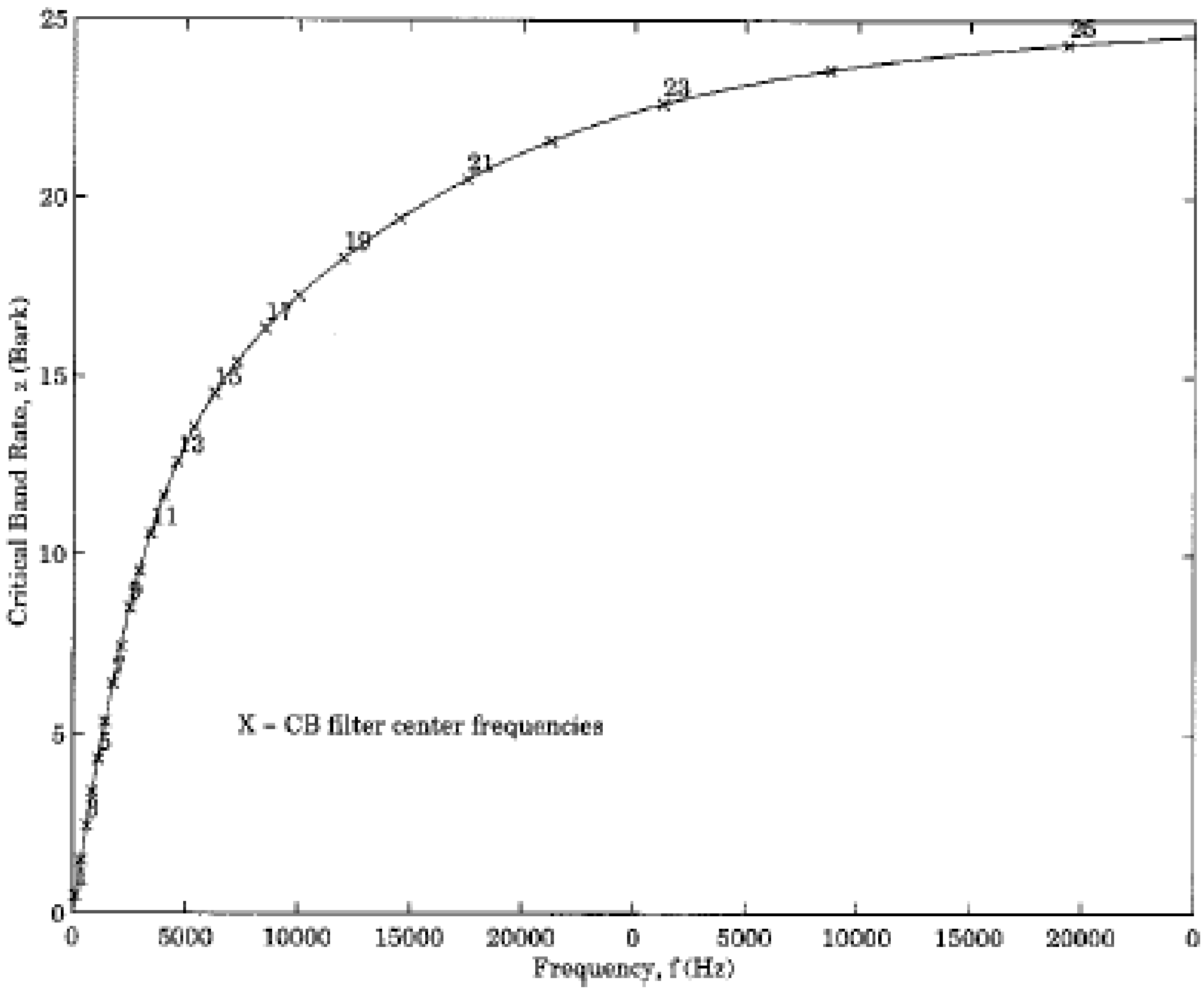
Absolute threshold of hearing in quiet



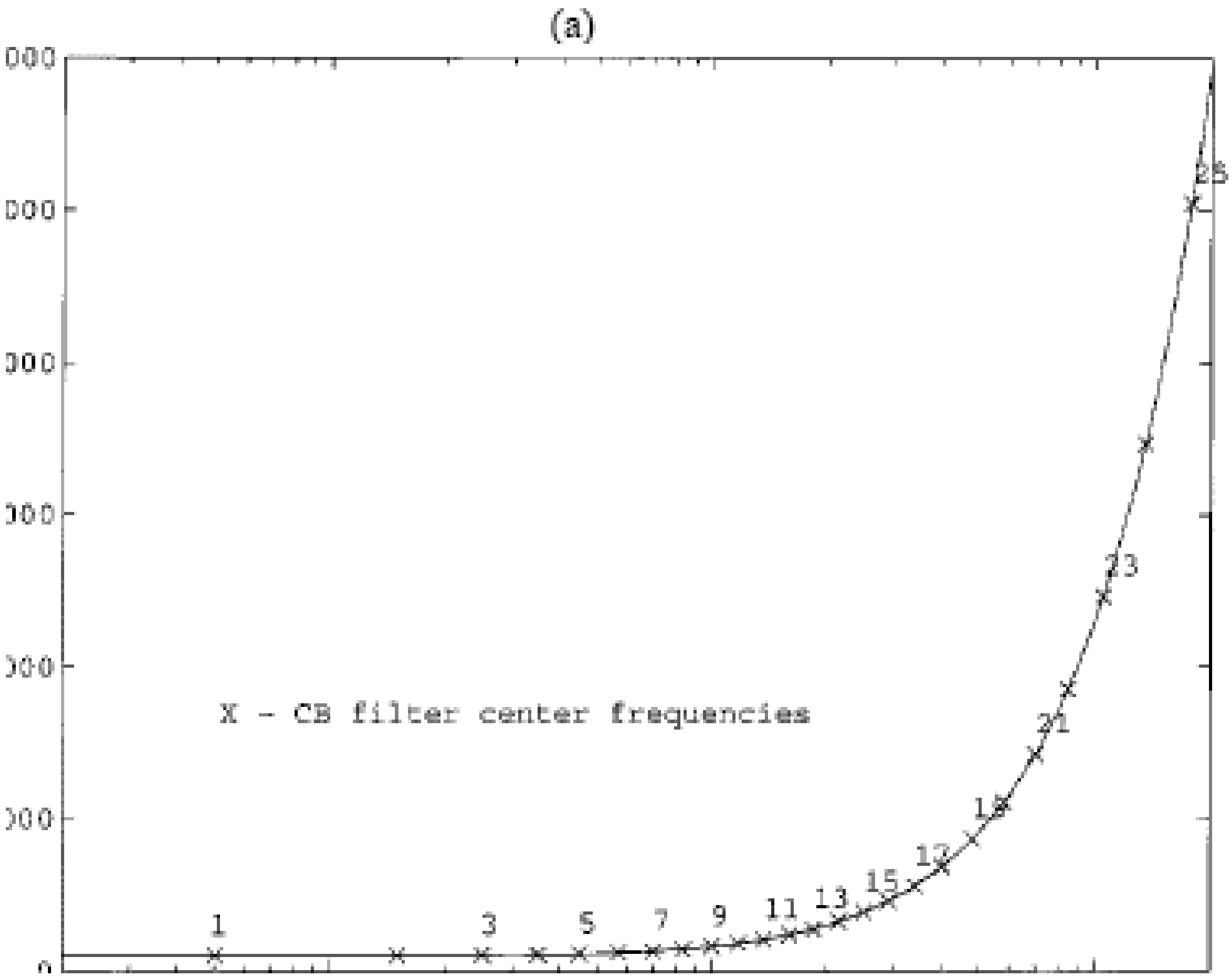
Critical bands-frequency to place transformation



Critical band-rate as a function of frequency

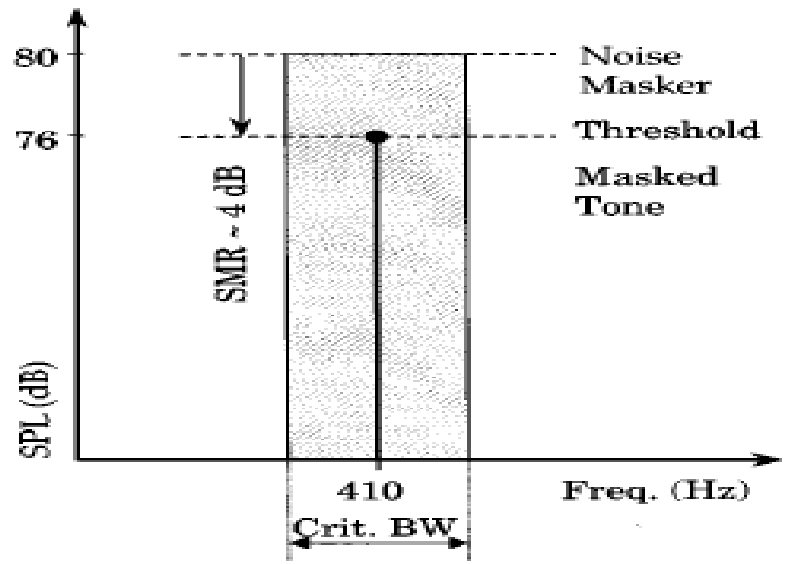


Critical bandwidth as a funtion of center-frequency

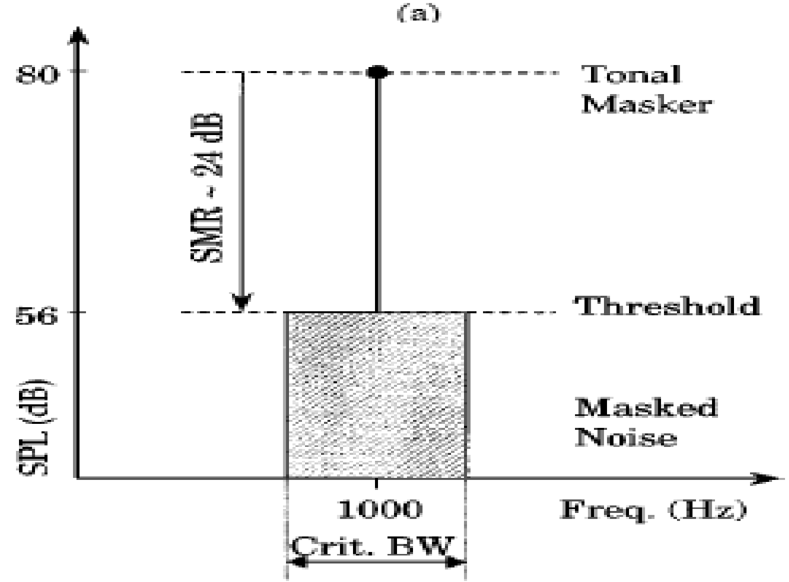


PSYCHOACOUSTIC PRINCIPLES Contd.

- Simultaneous masking, asymmetric masking and spread of masking
  - Noise masking tone
  - Tone masking noise
  - Noise masking noise

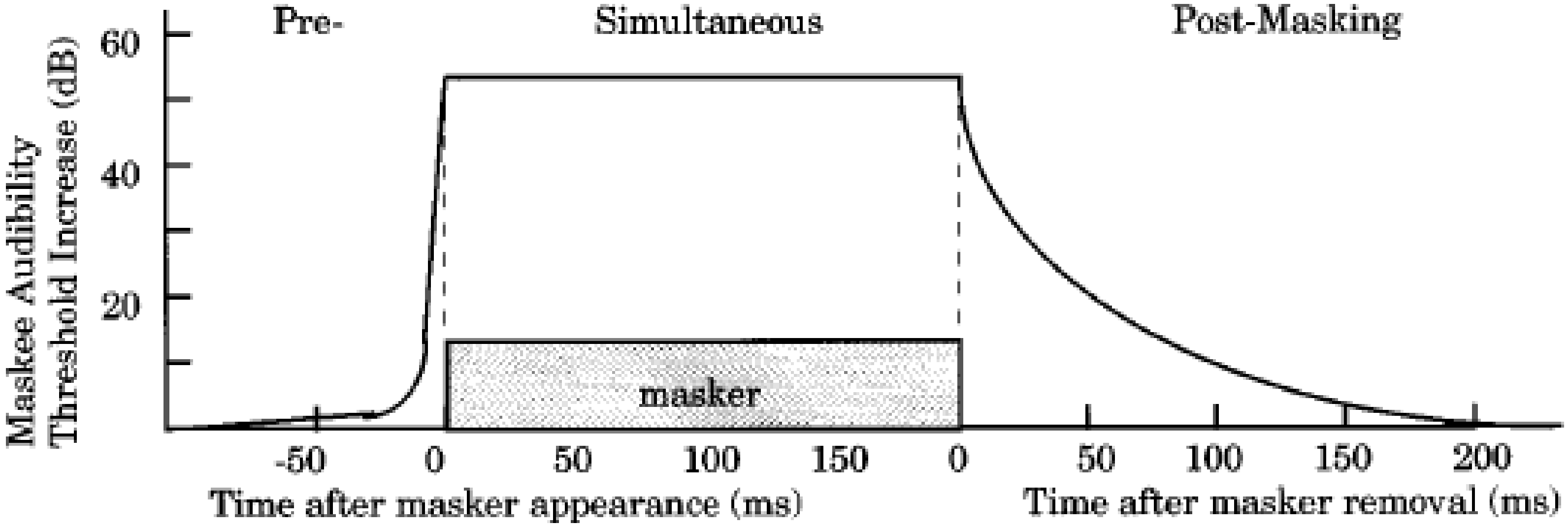

80
76
Noise
Masker
Threshold
Masked
Tone
SMR ~ 4 dB
SPL (dB)
410
Freq. (Hz)
Crit. BW

(a)


80
56
Tonal
Masker
Threshold
Masked
Noise
SMR ~ 24 dB
SPL (dB)
1000
Freq. (Hz)
Crit. BW

PSYCHOACOUSTIC METRIC

- Non-simultaneous masking



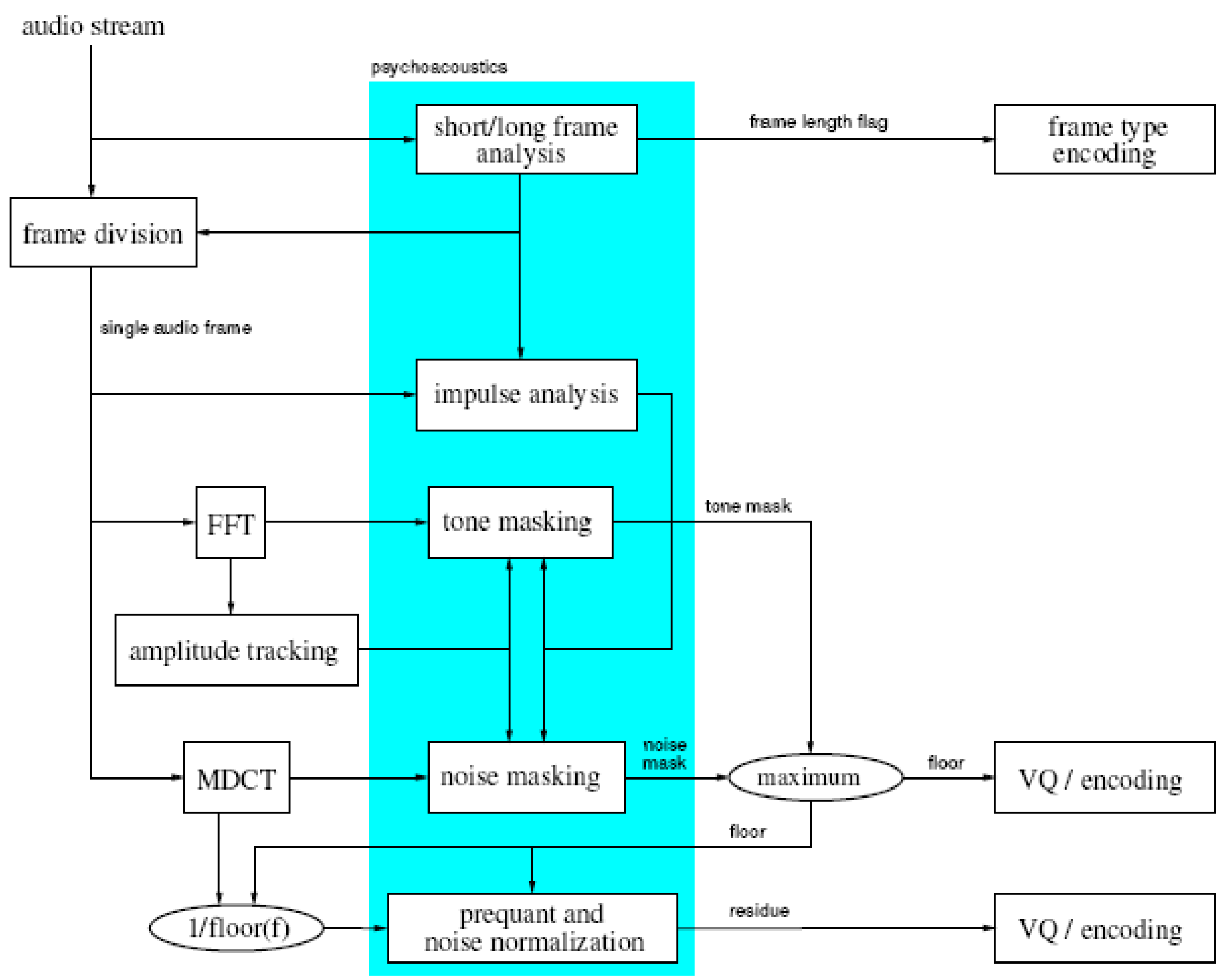
PSYCHOACOUSTIC METRIC

- Perceptual entropy
  - Signal is windowed and transformed to frequency domain
  - Masking threshold using perceptual rules
  - Using PE histogram the no. of bits required to quantize the spectrum

VORBIS MODEL

- Encode side psycho-acoustic heuristics
- Each audio-spectrum floor curve is generated
  - Tone masking
    - Single spectral curve per frame which represents the threshold of perception per spectral line
  - Noise masking
    - Envelope of noise energy in the spectrum
    - Adds hard-wired noise-offset producing noise-mask curve

Psycho acoustic model in VORBIS



VORBIS MODEL

- Noise normalization
  - To preserve the noise-masking curve through quantization process

Impulse analysis

  - Refers to localized temporal events.
  - Improves the non-sinusoidal, non-random-noise content

VORBIS MODEL

- Floor curve= max(tone mask, noise mask) after direct superposition
- Spectral residues= MDCT spectrum- floor curve
- Floor and residue curves are then vector quantized
- Vorbis audio codec uses Spectral flatness measure curve
- This curve is made using Geometric median and envelope followers by smoothing the log spectrum using sliding window of 1 bark
- Distances tell about tonality and noisiness