Audio Codecs

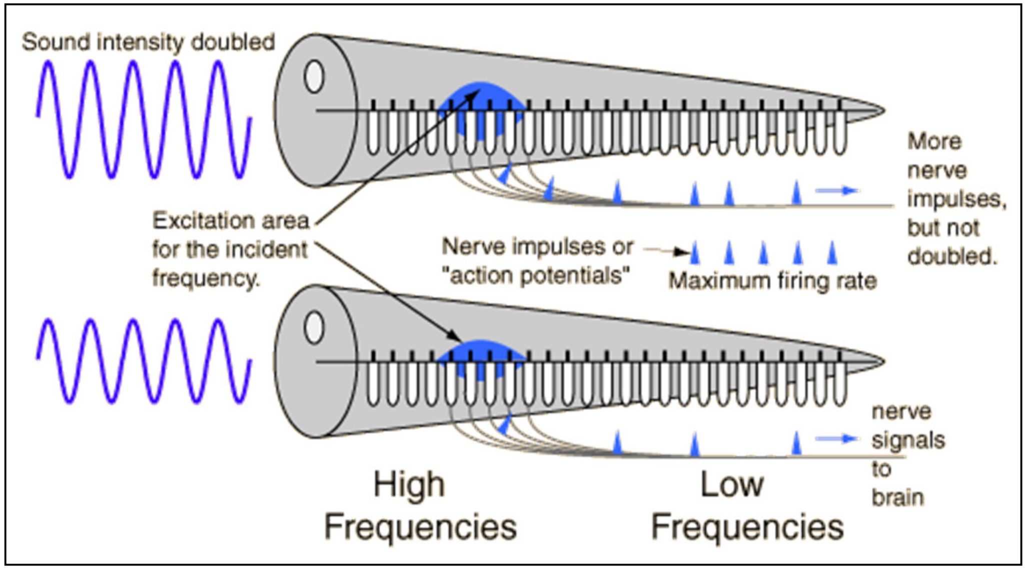
National Chiao Tung University

Chun-Jen Tsai

12/25/2014

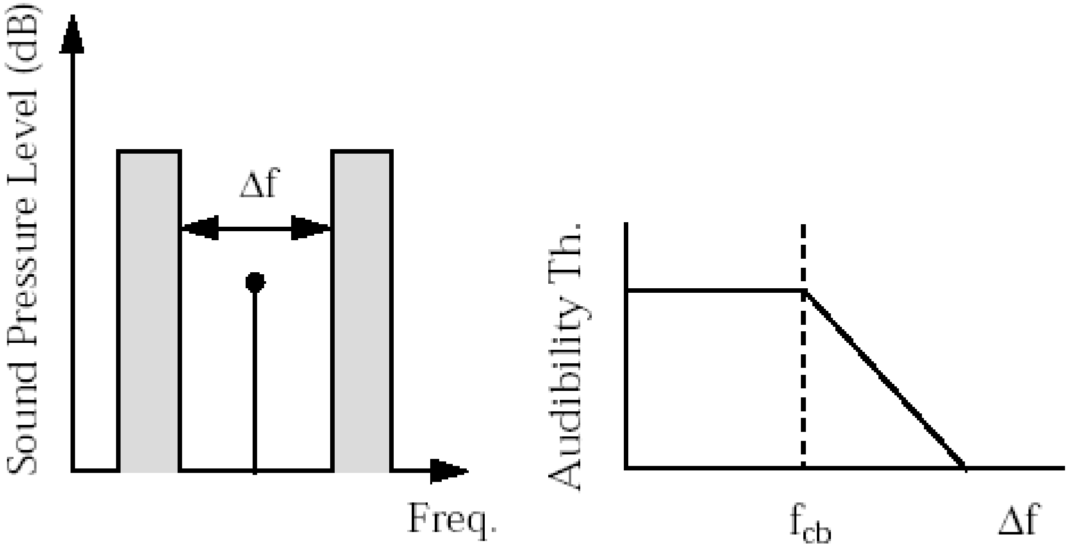
Perceptual Coding

- Human auditory system model:
  - A bandpass filterbank with 25 overlapping critical bands (CB) covering 20~20k Hz
  - For a given frequency, the critical band is the smallest neighborhood of frequencies around it which excites the same nerve cells


Sound intensity doubled
More nerve impulses, but not doubled.
Excitation area for the incident frequency.
Nerve impulses or "action potentials"
Maximum firing rate
nerve signals to brain
High Frequencies
Low Frequencies

CB Bandwidth Measurement

- The bandwidth of a CB can be measured by taking a sine tone barely masked by a band of white noise around it; when the noise band is narrowed until the point where the sine tone becomes audible, its width at that point is the critical bandwidth



Critical Band Frequencies

- The width of one critical band is commonly referred to as “one bark”

Band #	Center Freq.	Range
1	50	~ 100
2	150	100 ~ 200
3	250	200 ~ 300
4	350	300 ~ 400
5	450	400 ~ 510
6	570	510 ~ 630
20	5800	5300 ~ 6400
21	7000	6400 ~ 7700
22	8500	7700 ~ 9500
23	10500	9500 ~ 12000
24	13500	12000 ~ 15500
25	19500	15500 ~

Sound Pressure Level (SPL)

- ❑ Sound pressure level is a measure of the magnitude of sound:

$$\text{SPL} = 10 \log (\rho/\rho_{\text{ref}})^2$$

  - ■ ρ is the given sound pressure (in N/m^2)
  - ■ ρ_{ref} is the reference sound (just-audible in a quiet room)

- ❑ Typical SPL:

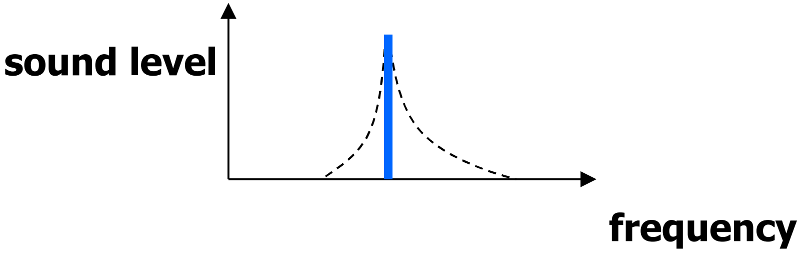
Sound Source	SPL, dB
Gunshot at close range	140
Loud rock group	120
Shouting at close distance	100
Normal conversation	70
Quiet conversation	50
Soft whisper	30
Ref. Level	0

Masking Effect and Audio Coding

- ❑ "Signal masking" is a key to audio compression
  - ■ Masker: dominating strong signal
  - ■ Maskee: low-level "hard-to-hear" signal
- ❑ Masking effects:
  - ■ In frequency domain → simultaneous masking
  - ■ In temporal domain → temporal masking
- ❑ There are four types of masking:
  - ■ tone-mask-noise
  - ■ noise-mask-tone
  - ■ noise-mask-noise → too complicated to use!
  - ■ tone-mask-tone → too complicated to use!

Simultaneous Masking

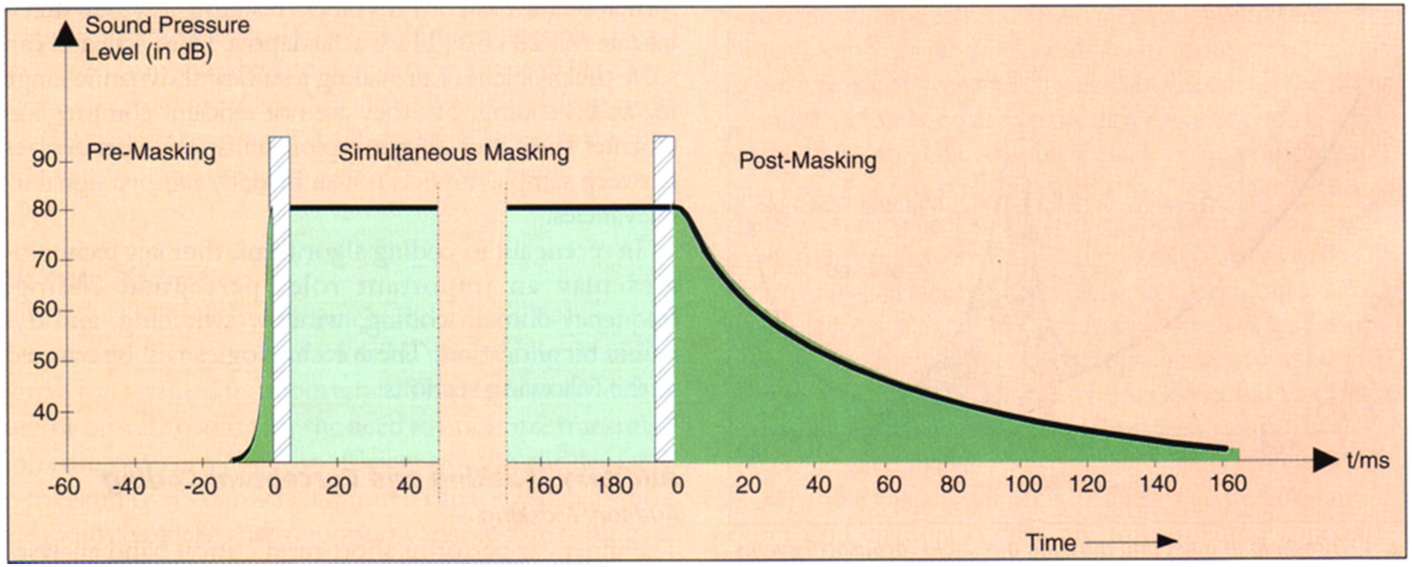
- A strong signal creates a masking envelop around it's frequency neighborhood



- The masked signal can be:
  - Low-level audio signals
  - Quantization noise signals
  - Aliasing distortion
  - Transmission errors

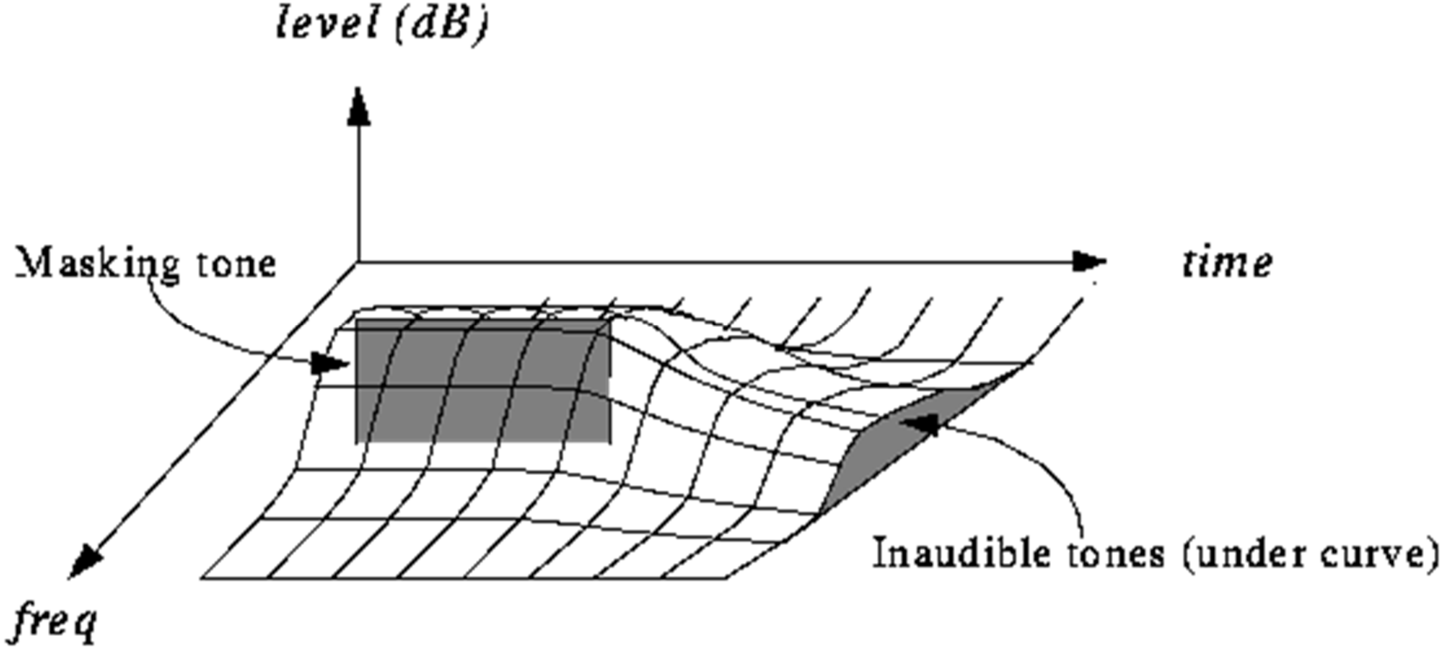
Temporal Masking

- ❑ An audio signal also mask signals before and after its existence[†]



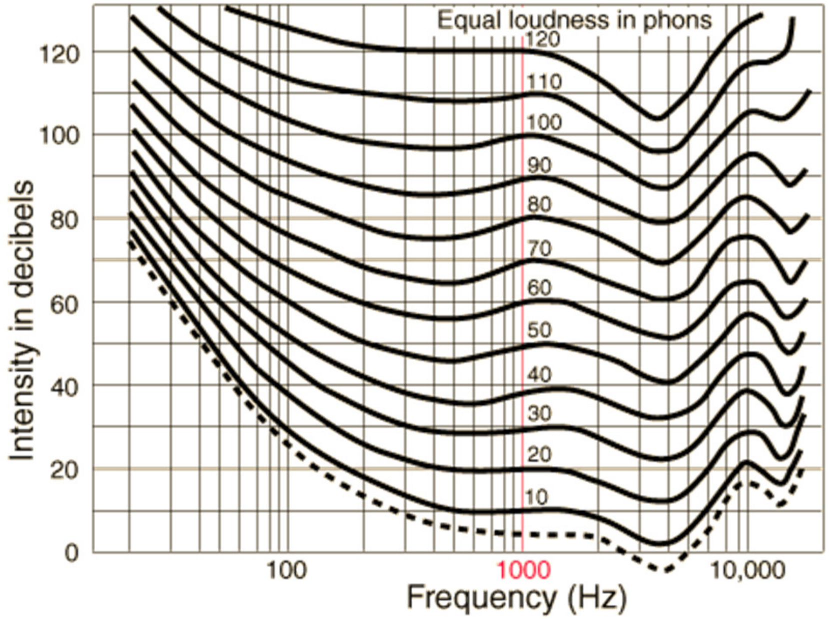
† Peter Noll, "MPEG Digital Audio Coding," *IEEE Signal Processing Magazine*, Vol. 14, No. 5, Sep. 1997

Total Effect of Masking



Threshold in Quiet

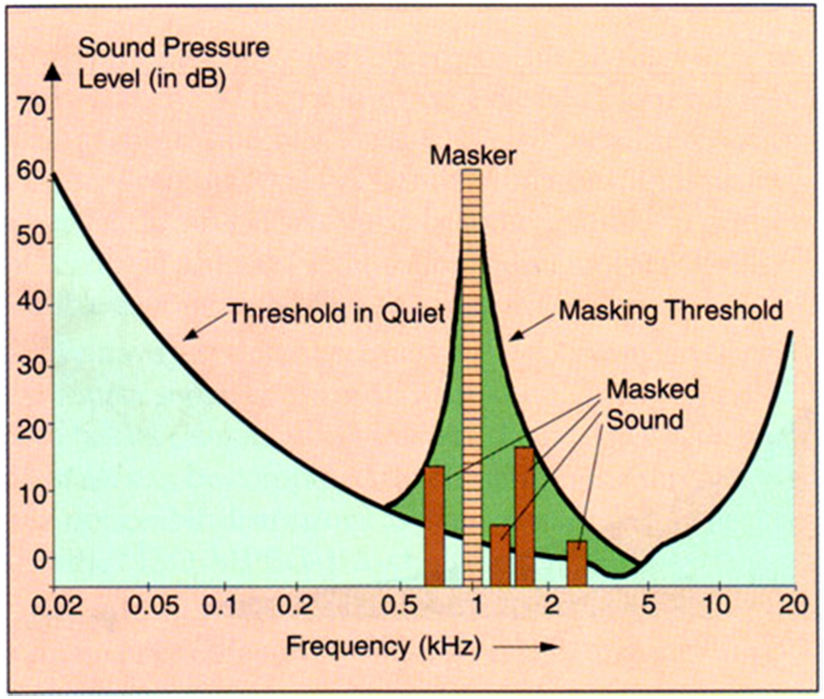
- “Threshold in quiet” is the threshold of “just audible sound” across all frequencies



⇨ The lower bound of the equal loudness curves is the threshold in quiet

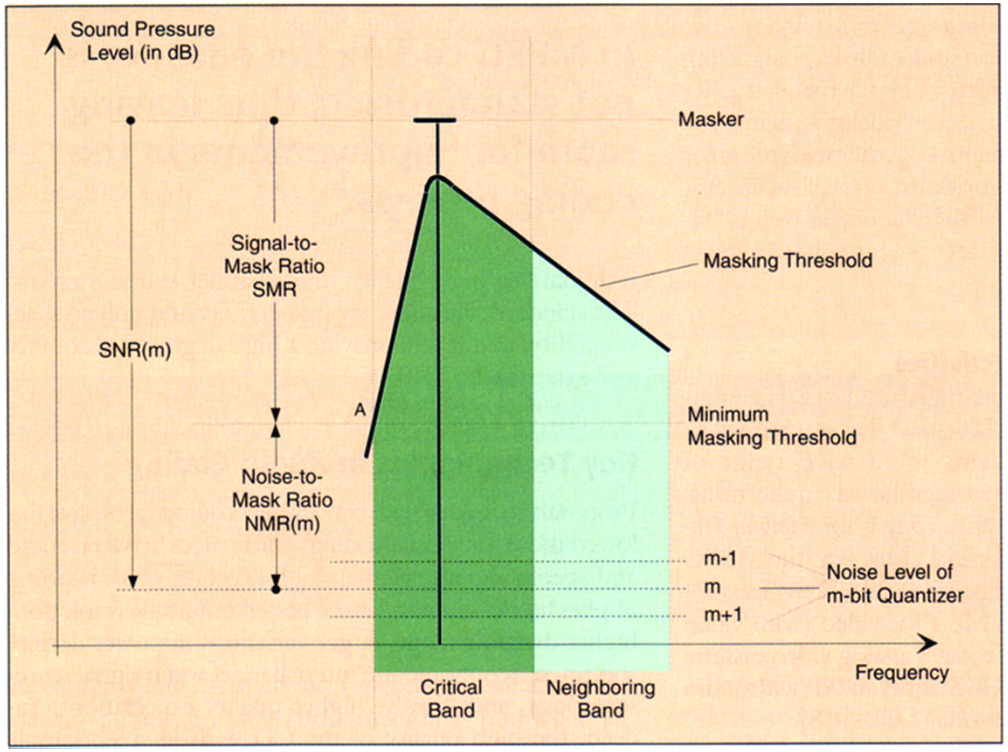
Inaudible Thresholds

- ❑ An audio signal must have SPL higher than inaudible threshold, or it's not audible
- ❑ The threshold is also known as threshold of "just noticeable distortion" (JND)
- ❑ These thresholds are time-varying



†Peter Noll, "MPEG Digital Audio Coding," *IEEE Signal Processing Magazine*, Vol. 14, No. 5, Sep. 1997

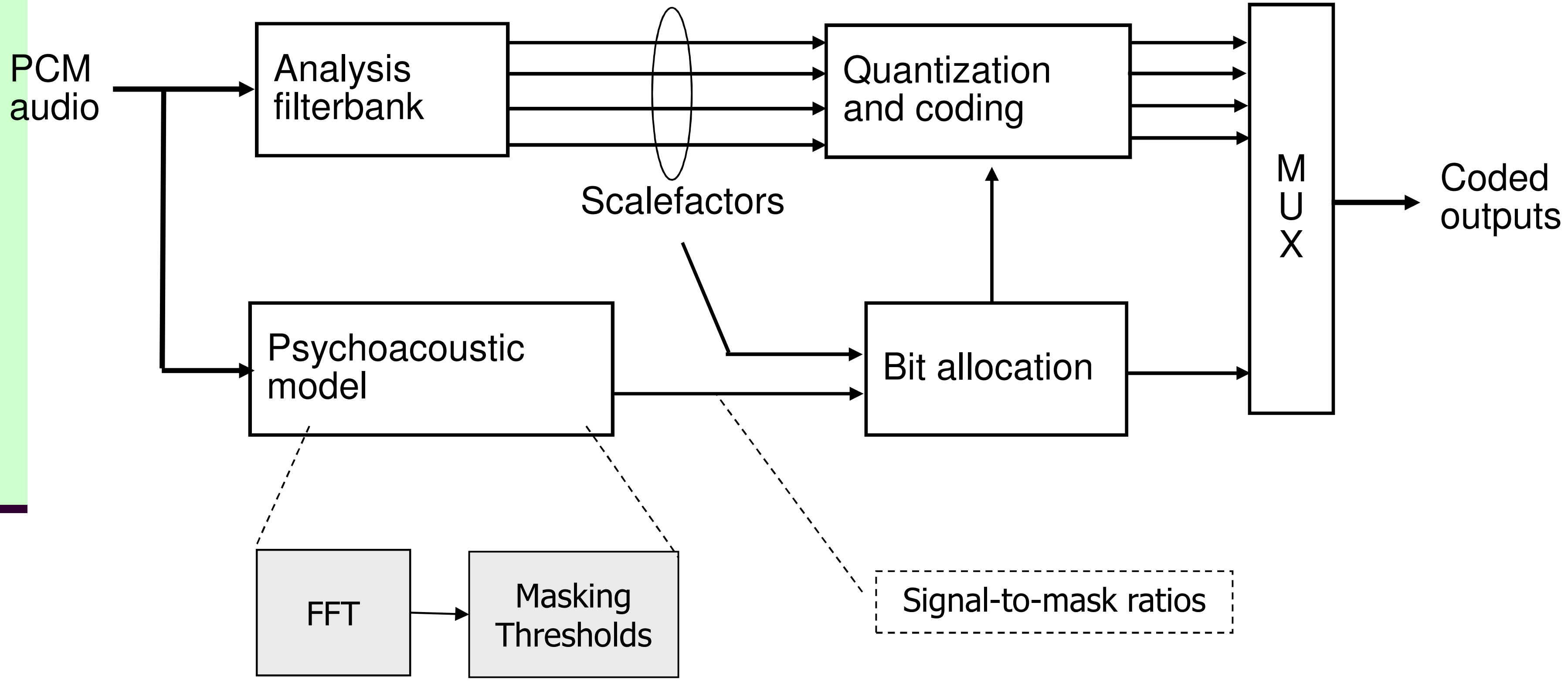
Signal-to-Mask Ratio (SMR)


Sound Pressure Level (in dB)
Masker
Signal-to-Mask Ratio SMR
SNR(m)
Masking Threshold
A
Minimum Masking Threshold
Noise-to-Mask Ratio NMR(m)
m-1
m
m+1
Noise Level of m-bit Quantizer
Frequency
Critical Band
Neighboring Band

Note: NMR ≤ 0. When NMR = 0, the distortion is just noticeable distortion

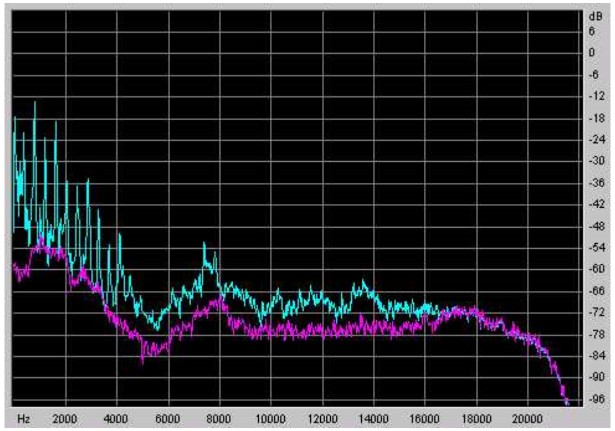
† Peter Noll, “MPEG Digital Audio Coding,” *IEEE Signal Processing Magazine*, Vol. 14, No. 5, Sep. 1997

General Audio Coding Structure



Perceptually Transparent Coding

- ❑ If the signal is coded with a complete masking of distortion, the coded signal is subjectively indistinguishable from the source signal
- ❑ JND coding is not desirable because:
  - ■ End-user processing amplifies noises
  - ■ Transcoding may take places during transmission
- ❑ Example: Blue – coded signal, Purple – noises

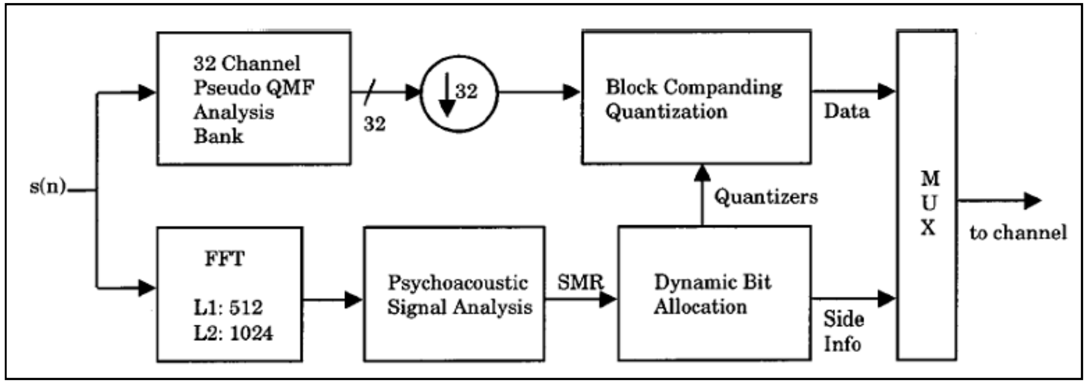


MPEG Audio Standards

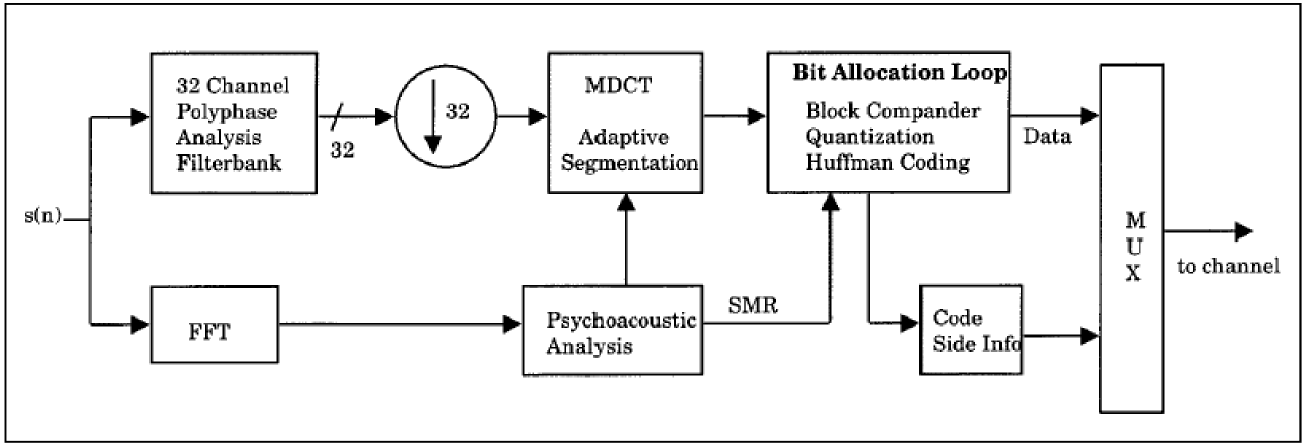
- MPEG-1 Layer 1, 2 (1992)
  - 32, 44.1, 48k sampling rates; 32~448 kbps; 1~2 channels
- MPEG-1 Layer 3 (1993)
  - 32, 44.1, 48k sampling rates; 32~320 kbps; 1~2 channels
- MPEG-2 Layer 1,2,3 (1994)
  - Extra 16, 22.05, 24k sampling rate; 1~5.1 channels
- MPEG-2 AAC (1997), MPEG-4 AAC (1999), Enhanced AAC+ (2004)
  - 8~64 kbps/ch, 1~96 channels

MPEG Audio Encoder Models

- ❑ Layers 1 & 2 Encoder

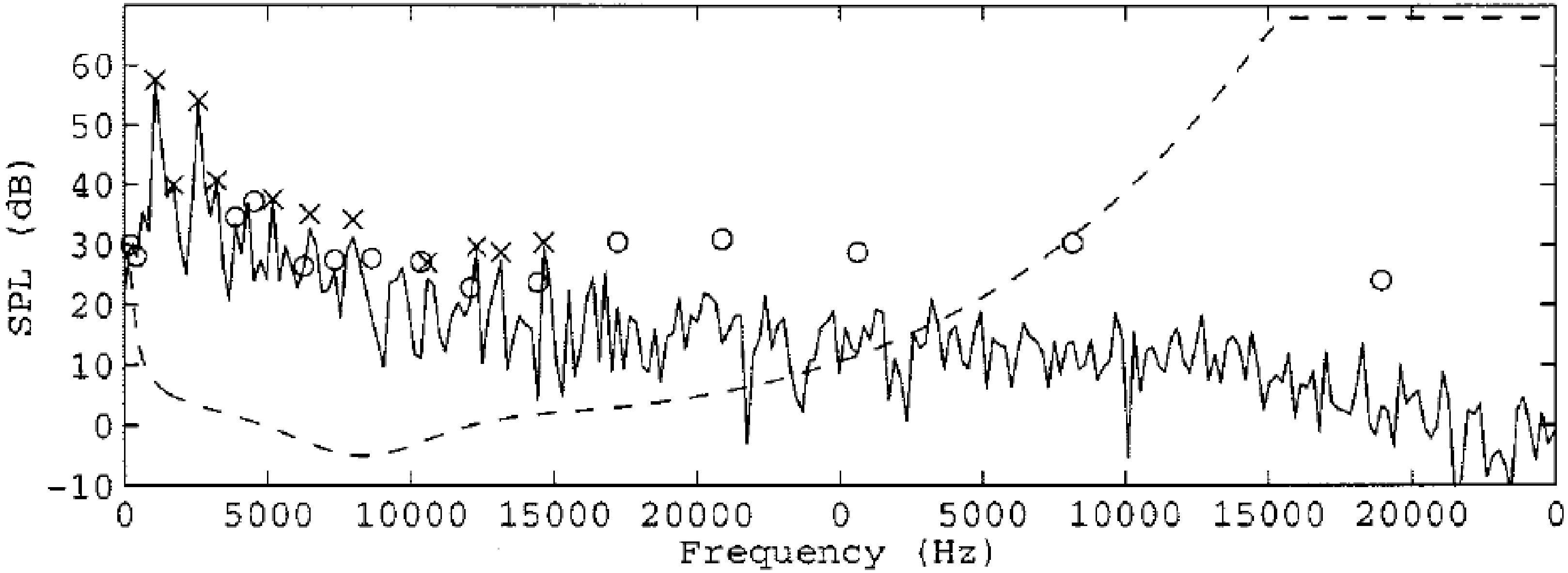

32 Channel Pseudo QMF Analysis Bank
32
32
Block Companding Quantization
Data
s(n)
Quantizers
M U X
to channel
FFT
L1: 512
L2: 1024
Psychoacoustic Signal Analysis
SMR
Dynamic Bit Allocation
Side Info

- ❑ Layer 3 Encoder



MPEG-1 Perceptual Model Example

- ❑ For each audio frame, perform 512/1024-point FFT analysis and construct a threshold mask as follows:



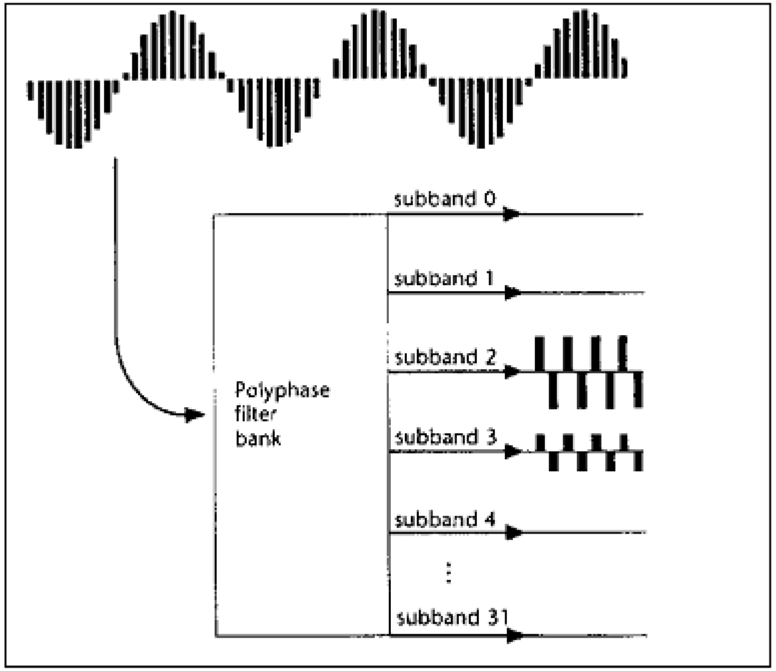
MPEG-1 Filter Banks

❑ 32-band poly-phase filter bank

- Critically sampled: 32 input samples → 32 outputs
- Low frequency resolution with overlapping bands

❑ Example output

- Input: 1500 Hz sine wave sampled at 32kHz
- Output: subband 2 & 3 have significant outputs

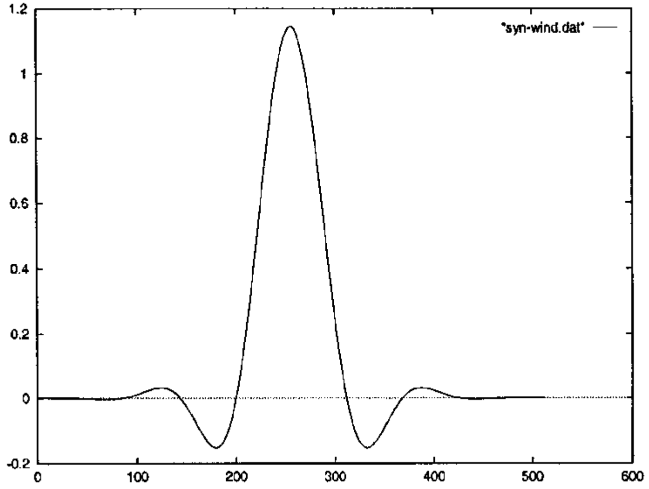


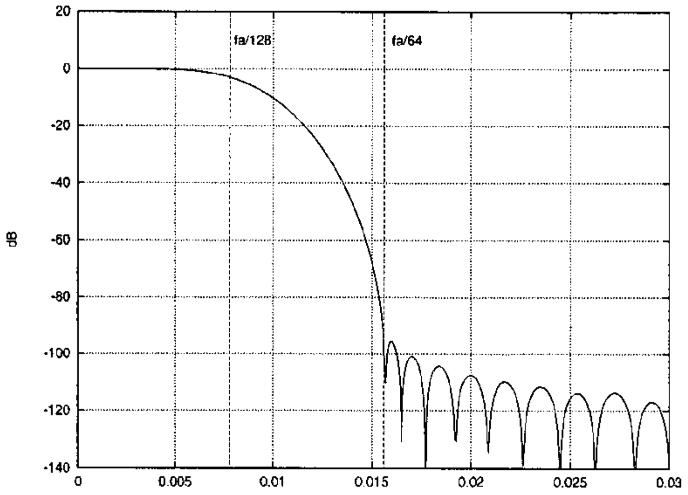
Lowpass Prototype Filter

❑ Left: impulse response; right: frequency response

(time domain)

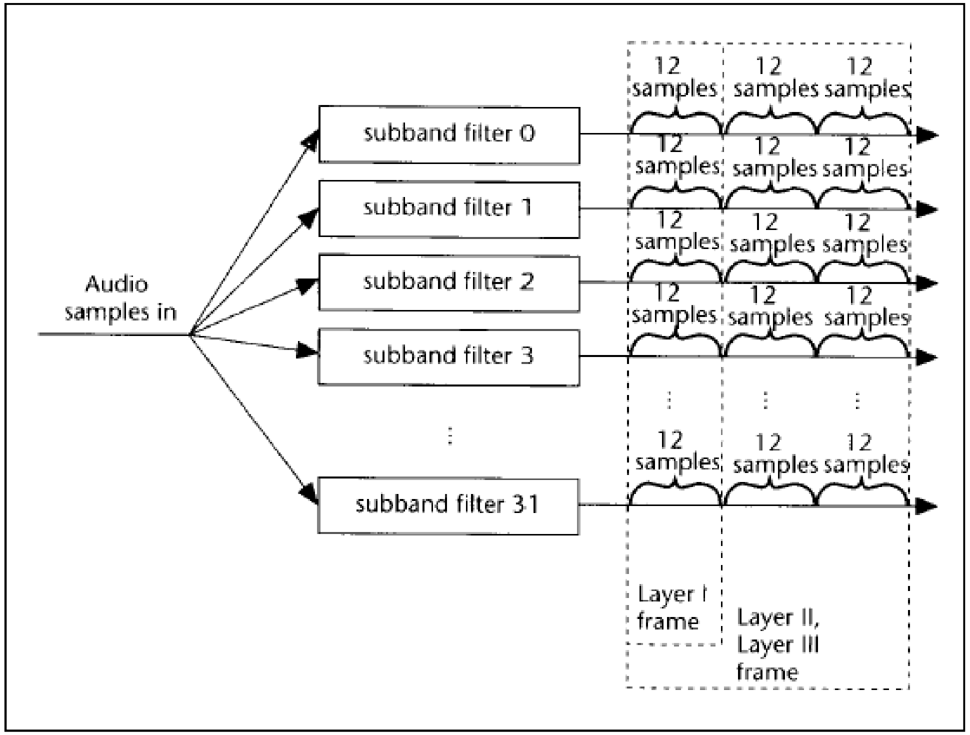
(frequency domain)





MPEG Audio Frame Sizes

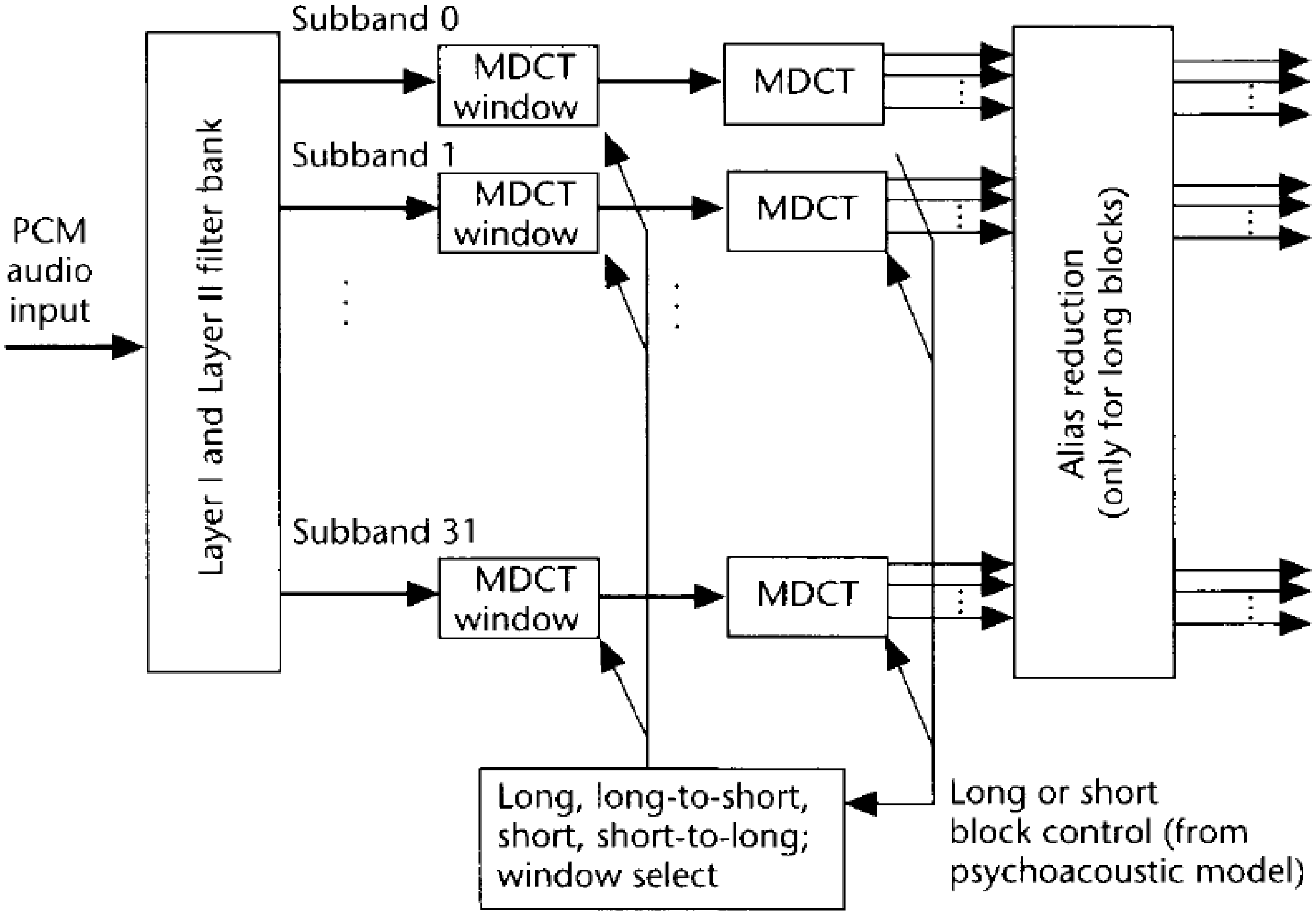
- ❑ Layer 1: 384 (= 32×12) samples/frame
- ❑ Layer 2&3: 1152 (= 32×3×12) samples/frame



MPEG Layer 3 MDCT (1/2)

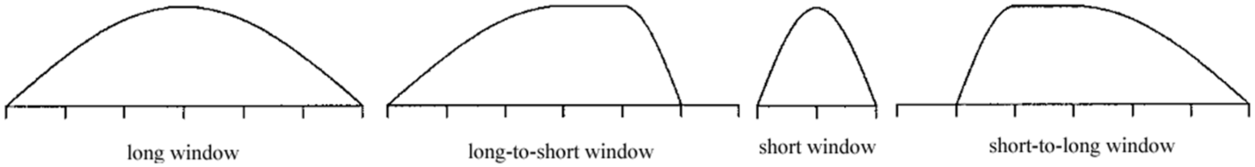
- MPEG Layer 3 inserted a cascaded transform module, MDCT, between the filter bank and the quantizer to further increase the coding efficiency
- Three subband block length: for each subband of each frame, block length can be long (18 sample), short (6 samples), or mixed

MPEG Layer 3 MDCT (2/2)

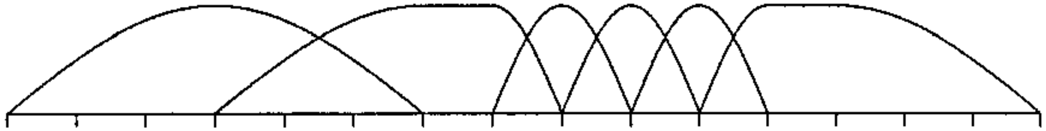

Subband 0
Subband 1
Subband 31
PCM audio input
Layer I and Layer II filter bank
MDCT window
MDCT window
MDCT window
MDCT
MDCT
MDCT
Alias reduction (only for long blocks)
Long, long-to-short, short, short-to-long; window select
Long or short block control (from psychoacoustic model)

MDCT Window Switching

- ❑ Four window types: long (36-point), short (12-point), long-to-short (36-point), and short-to-long (36-point)

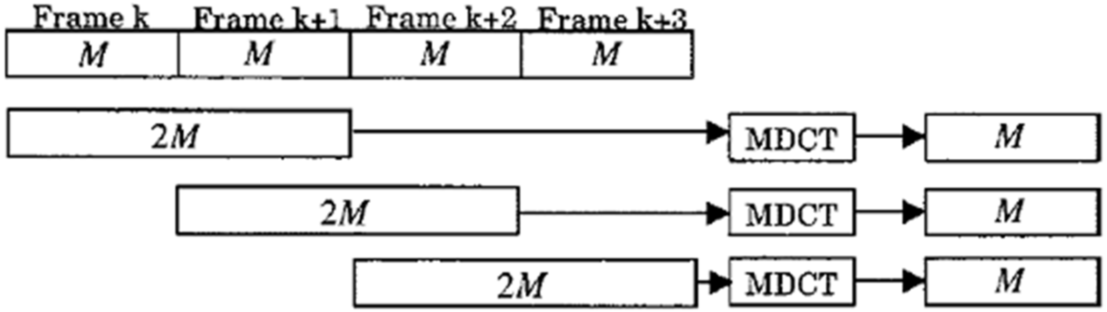


- ❑ Window-switching in a subband:

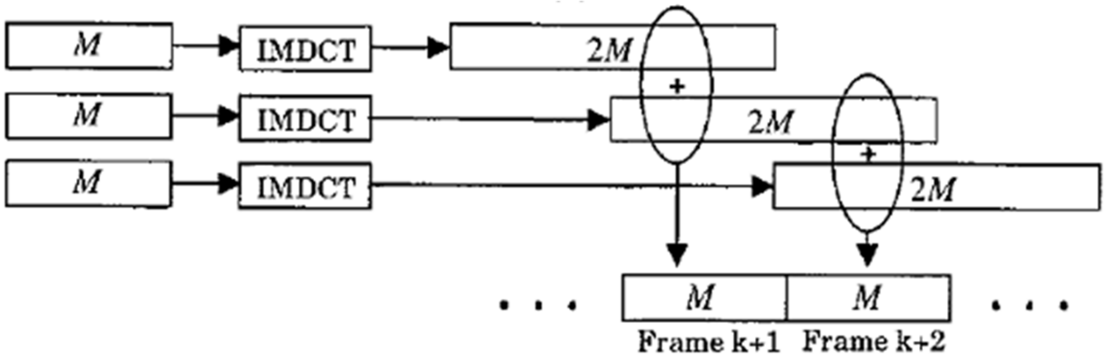


MDCT Operations

- Analysis filter: $2M$ overlapped inputs, M output components

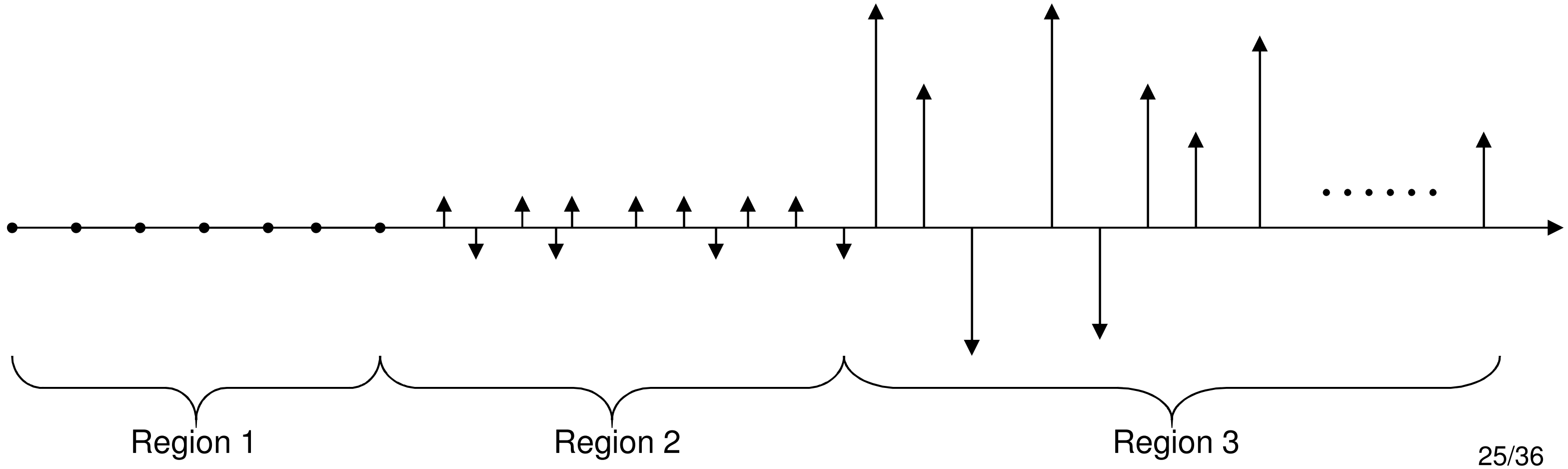


- Synthesis filter: M input spectral components, $2M$ lapped outputs



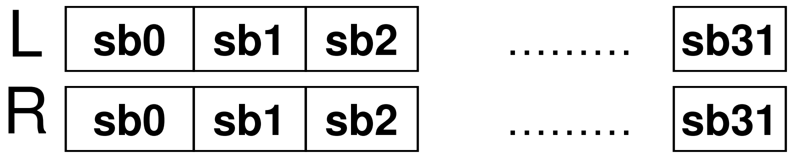
Spectral Region Partition

- ❑ 576 (32×18) MDCT coefficients are ordered (by freq, then by window) first
- ❑ Partition the ordered coefficients into three regions
  - Region 1: leading zeros
  - Region 2: runs of only ±1 and 0 (called "count 1" region)
  - Region 3: remaining coefficients (called "big values" region)



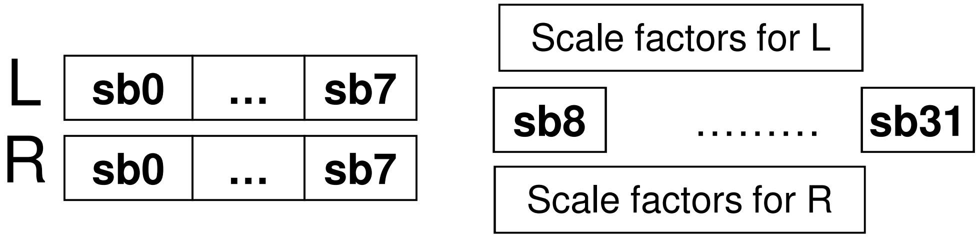
Stereo Coding

- ❑ Stereo signals:



L	sb0	sb1	sb2	………	sb31
R	sb0	sb1	sb2	………	sb31

- ❑ Intensity stereo coding:



L	sb0	…	sb7
R	sb0	…	sb7

Scale factors for L

sb8 ……… sb31

Scale factors for R

- ❑ Middle/side (MS) coding (Layer 3 only): convert signal to sum/difference instead of L/R

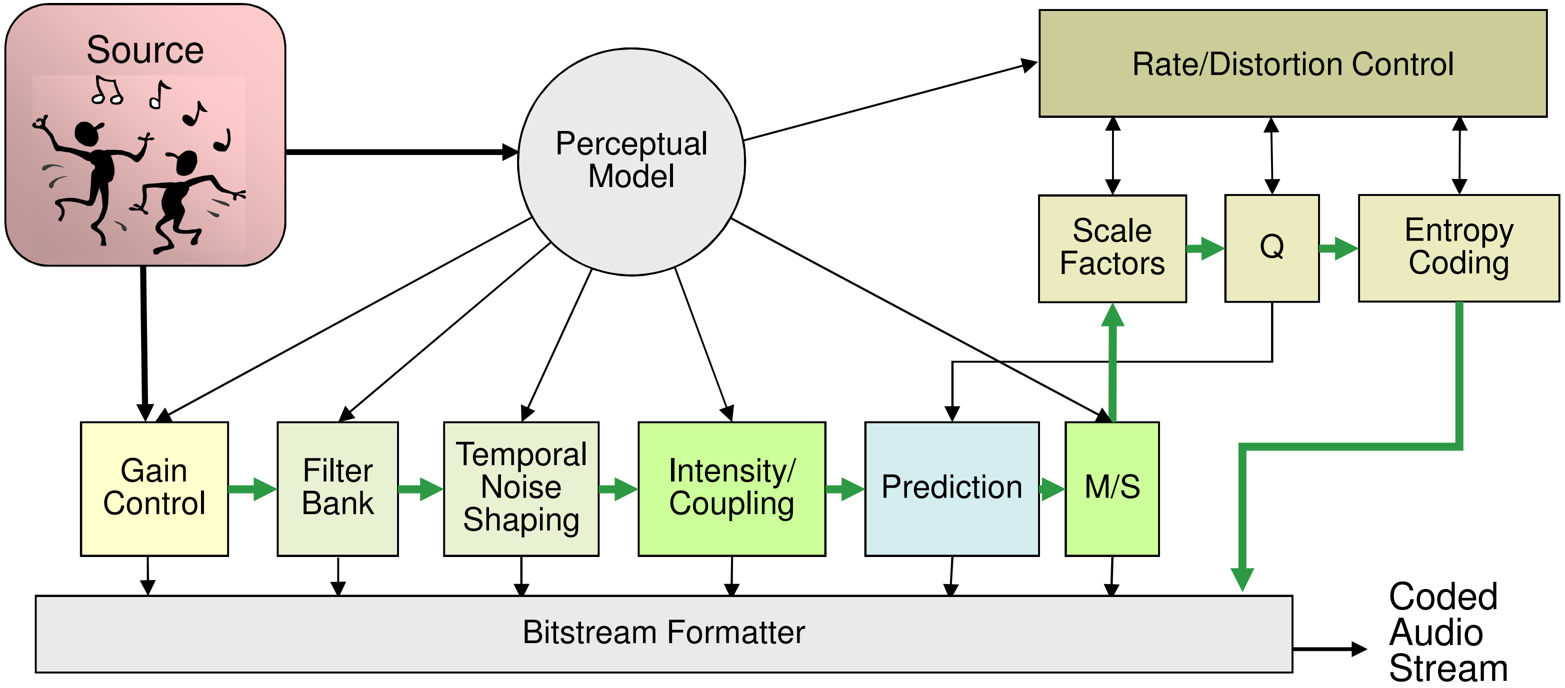
Note: Scale factor is the amplifier value for the frequency spectrum in a critical band.

Advanced Audio Coding (AAC)

- Main difference from MP3 are as follows
  - Fine resolution MDCT (2048 window, 1024 output lines)
  - Adaptive windowing (long, short, …)
  - Temporal Noise Shaping
  - Prediction
  - Better perceptual model
  - Additional MPEG-4 tools: PNS, LPT, Twin VQ, …

AAC Audio Codec

- ❑ AAC achieves better efficiency by re-design the codec without considering compatibility



AAC Filter Bank

- Modulated, (50%) overlapped filter bank – MDCT
- Adaptive block switching: 256 and 2048
  - Long widow – good freq. resolution, higher coding gain for "stationary" signals
  - Short widow – good time resolution, higher quality control on "pitchy" signals
- Adaptive window shape: Inter-band leakage – separation between (nearby) freq. bands
  - Sine widow – narrow main-lobe, PR, DC-component is contained in one the (1st) coeffficients
  - Kaiser-Bessel Derived (KBD) widow – optimization of transition BW and rejection, PR

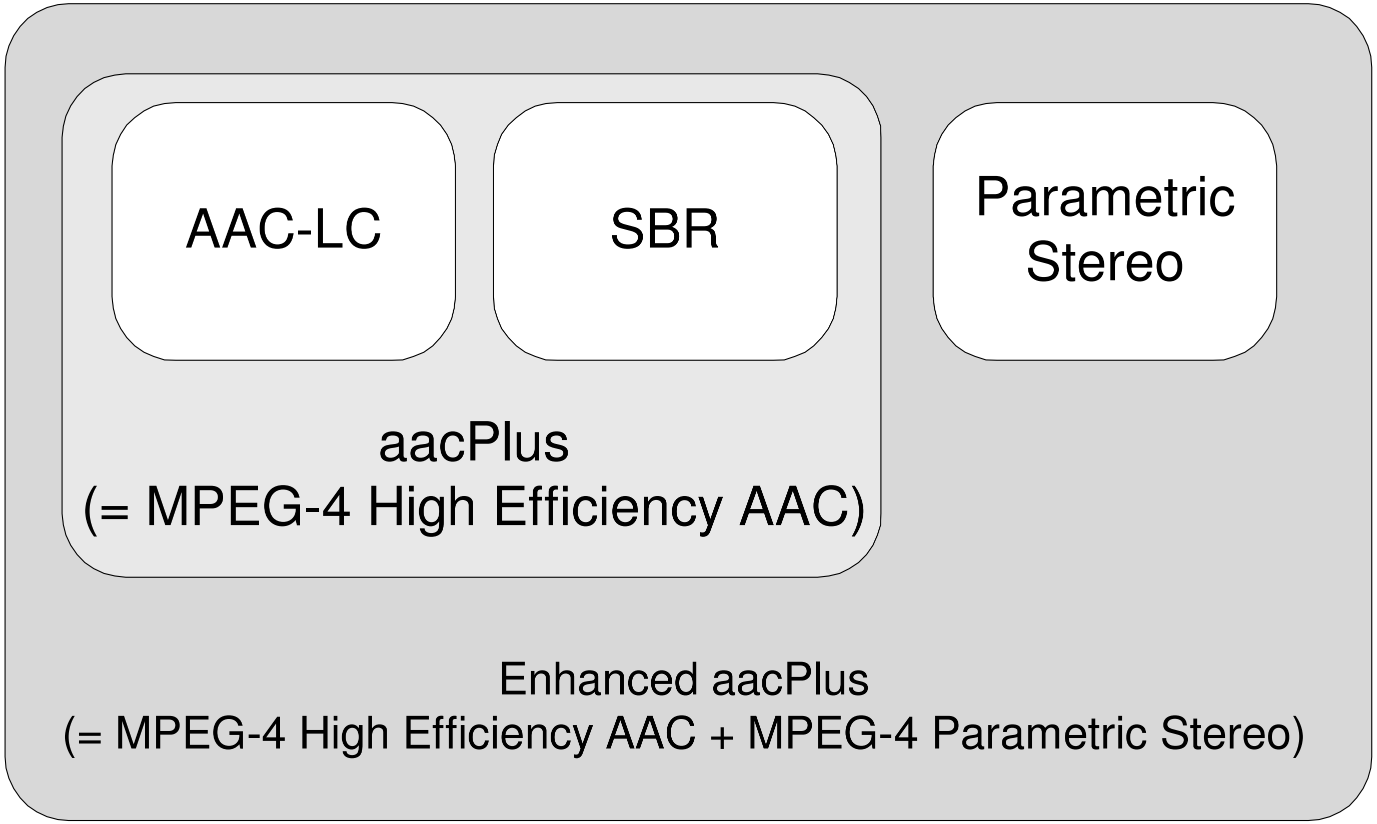
MPEG-4 AAC Tools (1/2)

- ❑ Perceptual Noise Substitution
  - ■ Parametric coding of noise-like signal components has been used widely e.g. in speech coding, why not audio?
- ❑ Long-term Prediction
  - ■ Tone-like signals require much higher coding precision than noise-like signals (e.g. 20 dB vs. 6 dB), but tonal signal components are predictable

MPEG-4 AAC Tools (2/2)

- ❑ SBR Bandwidth extension of audio signals
  - Recover high-frequency (e.g. >5kHz) part of signal content from low-frequency part
  - Use small helper information to approximate original signal t/f distribution “Lower Bitrates”
- ❑ Parametric coding of wide-band signals
  - Complements existing MPEG-4 parametric audio coder towards higher qualities & rates

MPEG-4 AAC Profiles



AAC-LC

SBR

Parametric Stereo

aacPlus
(= MPEG-4 High Efficiency AAC)

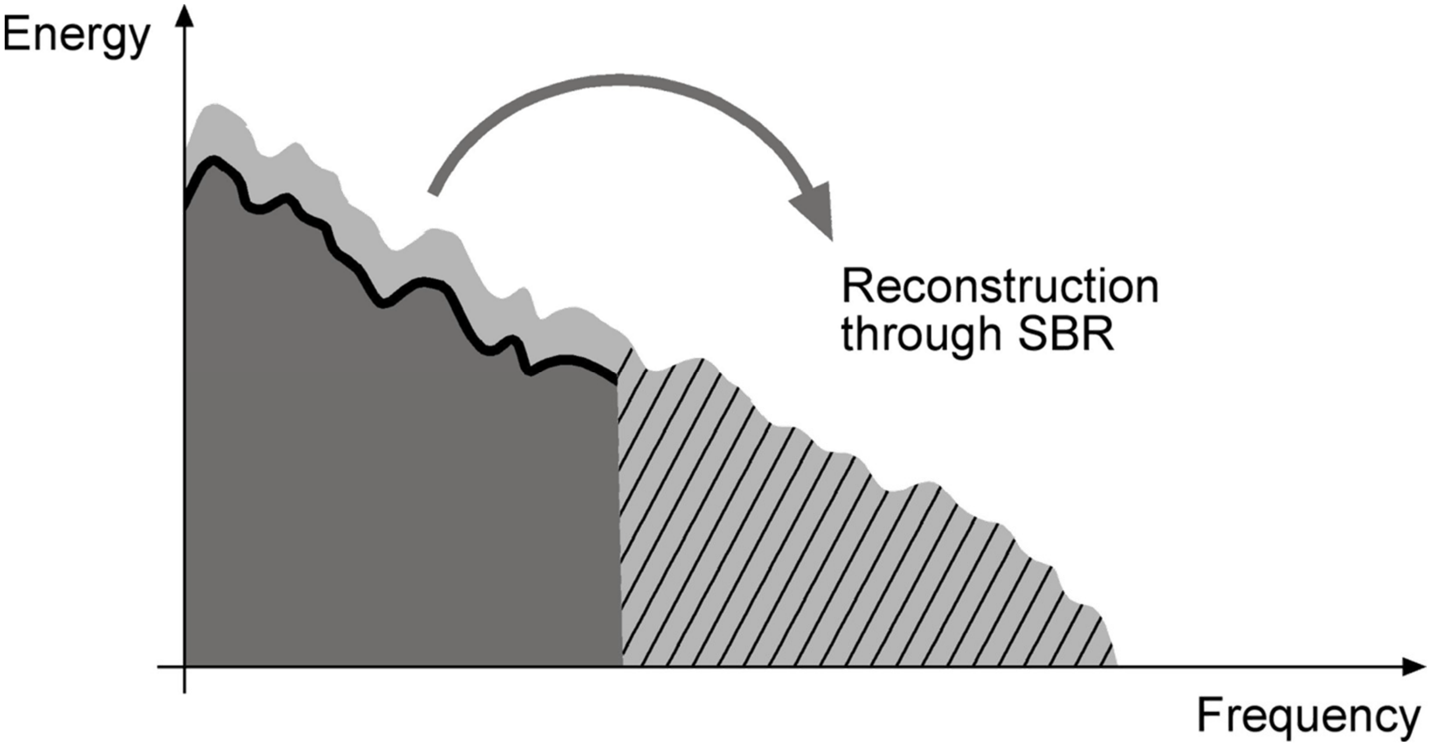
Enhanced aacPlus
(= MPEG-4 High Efficiency AAC + MPEG-4 Parametric Stereo)

The Idea behind SBR Bandwidth Ext.

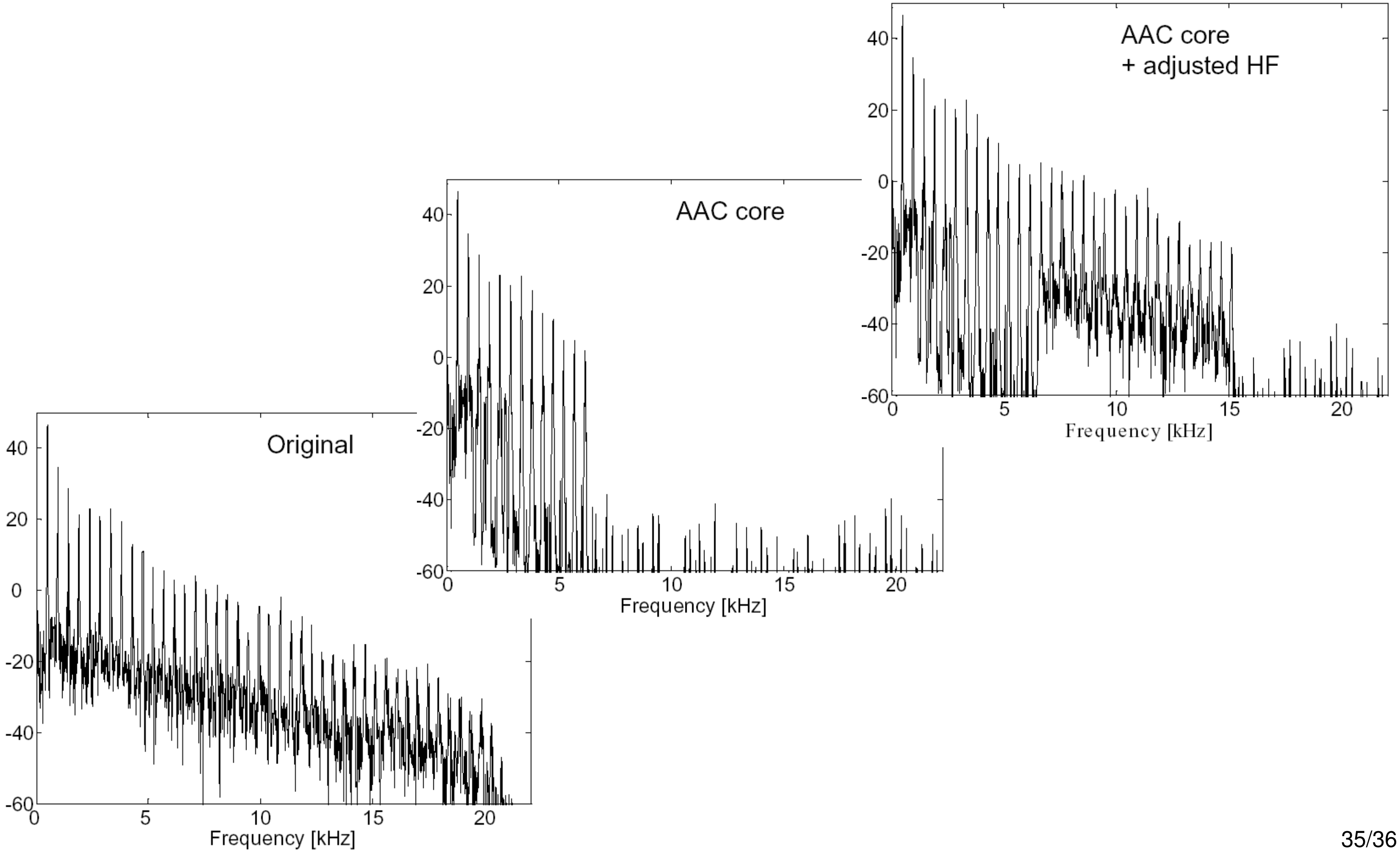
- ❑ There are high correlation between the lower and the higher frequency spectrum in audio signals
  - ■ Use traditional waveform for low frequencies
  - ■ High frequency data can be reconstruct from low frequencies so that there is no need to transmit them as spectral data!
- ❑ Even if correlation is low: Reconstructed data will be nicely related to lower frequencies
- ❑ Only few additional helper information is needed → Large gain in coding efficiency!

Spectral Band Replication (SBR)

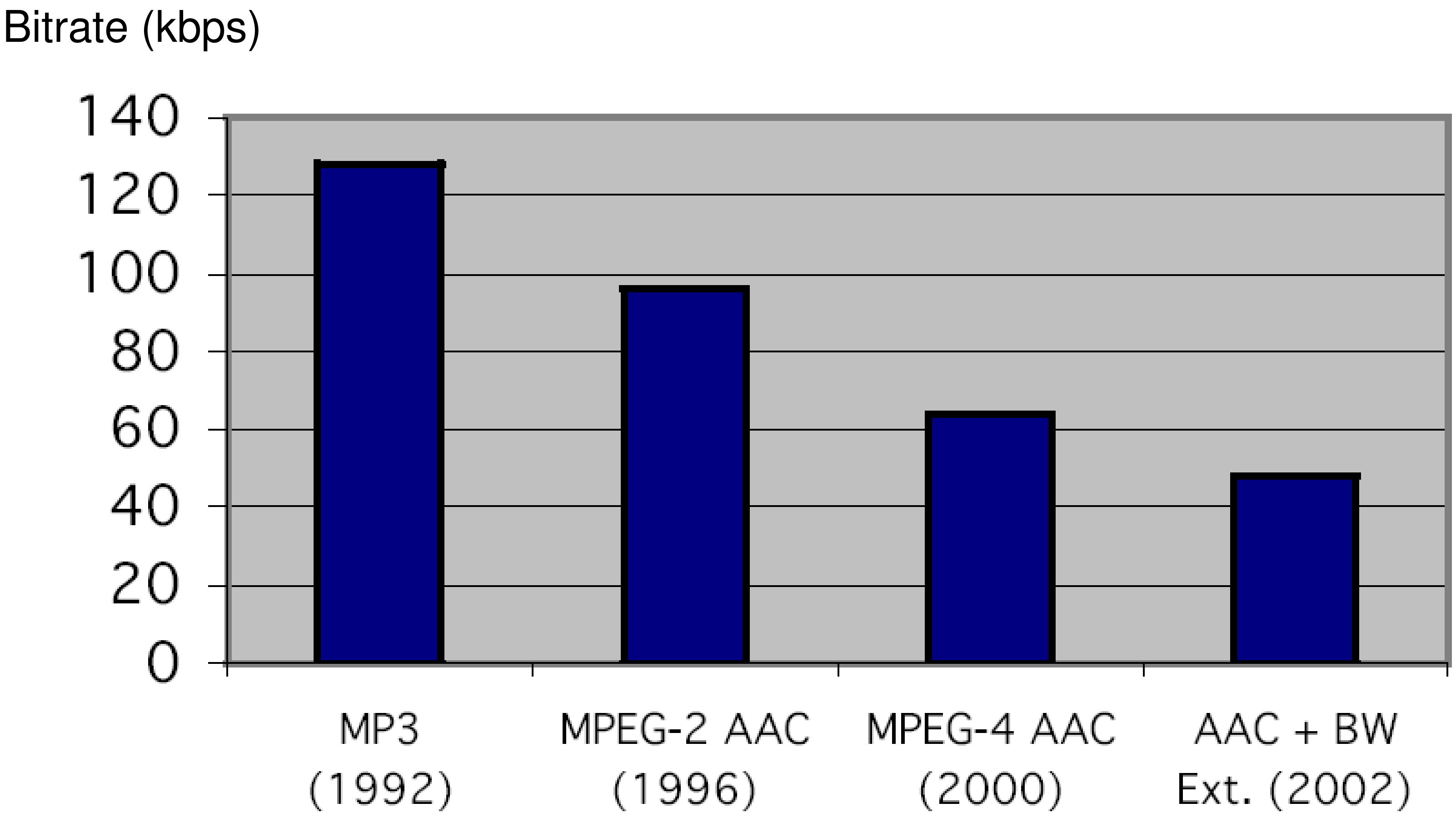
- The high frequencies are reconstructed and adjusted



SBR Example


Original
40
20
0
-20
-40
-60
0
5
10
15
20
Frequency [kHz]
AAC core
40
20
0
-20
-40
-60
0
5
10
15
20
Frequency [kHz]
AAC core
+ adjusted HF
40
20
0
-20
-40
-60
0
5
10
15
20
Frequency [kHz]

Progress in Coding Efficiency


Bitrate (kbps)
MP3 (1992)
MPEG-2 AAC (1996)
MPEG-4 AAC (2000)
AAC + BW Ext. (2002)