MPEG Audio Coding

Nimrod Peleg
Update: March 2004

Introduction

- *High quality low bit-rate audio coding*
- MPEG-1: Mono & Stereo, sampling rates of 32KHz, 44.1KHz and 48KHz.
- MPEG-2: Backward compatible coding of 5+1 multi-channel sound, more sampling rates: 16KHz, 22.05KHz and 24KHz.

Some Facts

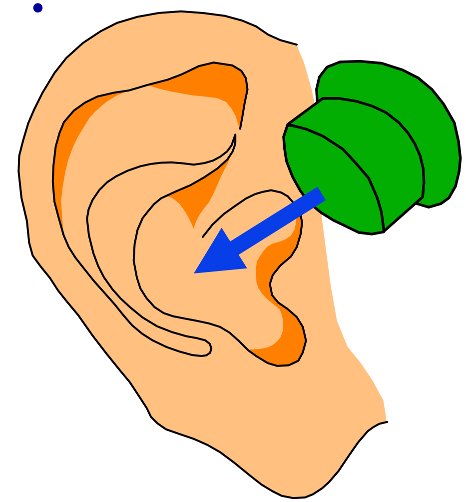
- MPEG-1: 1.5 Mbits/sec for audio and video:
  ~1.1Mbps for video, 0.3-0.4Mbps for audio
- Uncompressed CD audio is 44,100 samples/sec * 16 bits/sample * 2 channels > 1.4Mbps
- Typical Compression factors: from 2.7 to 24
- With Compression rate 6:1 (16 bits stereo sampled at 48 KHz is reduced to 256 kbps) and optimal listening conditions, expert listeners could not distinguish between coded and original audio clips.

Some Facts (Cont'd)

- MPEG-1 audio 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 (similar to stereo)
  - Stereo: for stereo channels that share bits, but not using joint-stereo coding
  - Joint-stereo: takes advantage of the correlation between stereo channels

Basic Idea: PsychoAcoustics

- How much noise can be introduced to the signal without being audible ?



PsychoAcoustic Model

Masking in the frequency domain

Reminder: Cochlear filter mechanism

- The bandwidth of filters (in the 'filter bank') varies strongly from low to high frequencies
- Center frequencies are call 'critical bands': mapping frequency onto a linear distance measure along the basiliar membrane.
- Filters bandwidth variation: 40:1
- Filters time response variation: 1:40
- Simultaneous control of time/frequency artifacts at 40:1 resolution range is difficult !

Barks

- Assuming that each critical band corresponds to a fixed distance along the basiliar membrane, we define a unit of length z(f) to be one critical band, and call it "Bark" (after Barkhausen).
- The approximation of z(f) is done using:

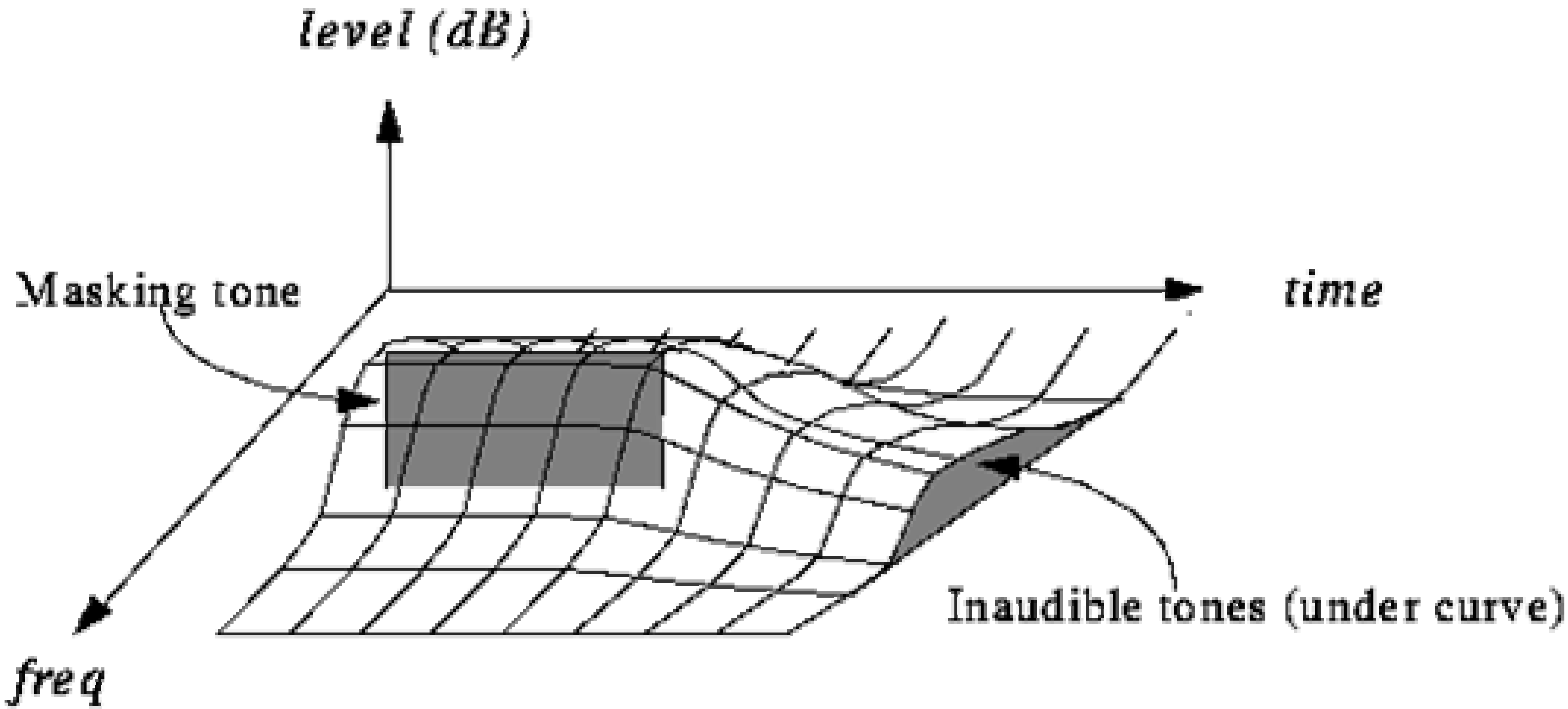
$$z/\text{Bark} = 13\ \text{Arctan}(0.76/1\text{KHz}) + 3.5\ \text{Arctan}[(f/7.5\text{kHz})^2)$$

- Bark width vary from ~100Hz in low freq. and 4kHz at ~15kHz.

The Barks table

Bark#	f_{low}(Hz)	f_{high}(Hz)	f_{center}(Hz)	BandWidth
0	0	100	50	100
1	100	200	150	100
2	200	300	250	100
•				
10	1270	1480	1370	210
11	1480	1720	1600	240
•				
22	9500	12000	10500	2500
23	12000	15500	13500	3500
24	15500			

Reminder: Auditory mechanism (HAS)



PsychoAcoustics Model

- Frequency is divided into “Barks”: bands of non-uniform width (narrower in lower freq.) according to the ear’s “resolution”
- Masking is influenced by two major parameters:
  - Tonal component (Masks the near frequency)
  - Noise component (Masks near and lower noises)
- Given an audio signal, the model creates a masking function

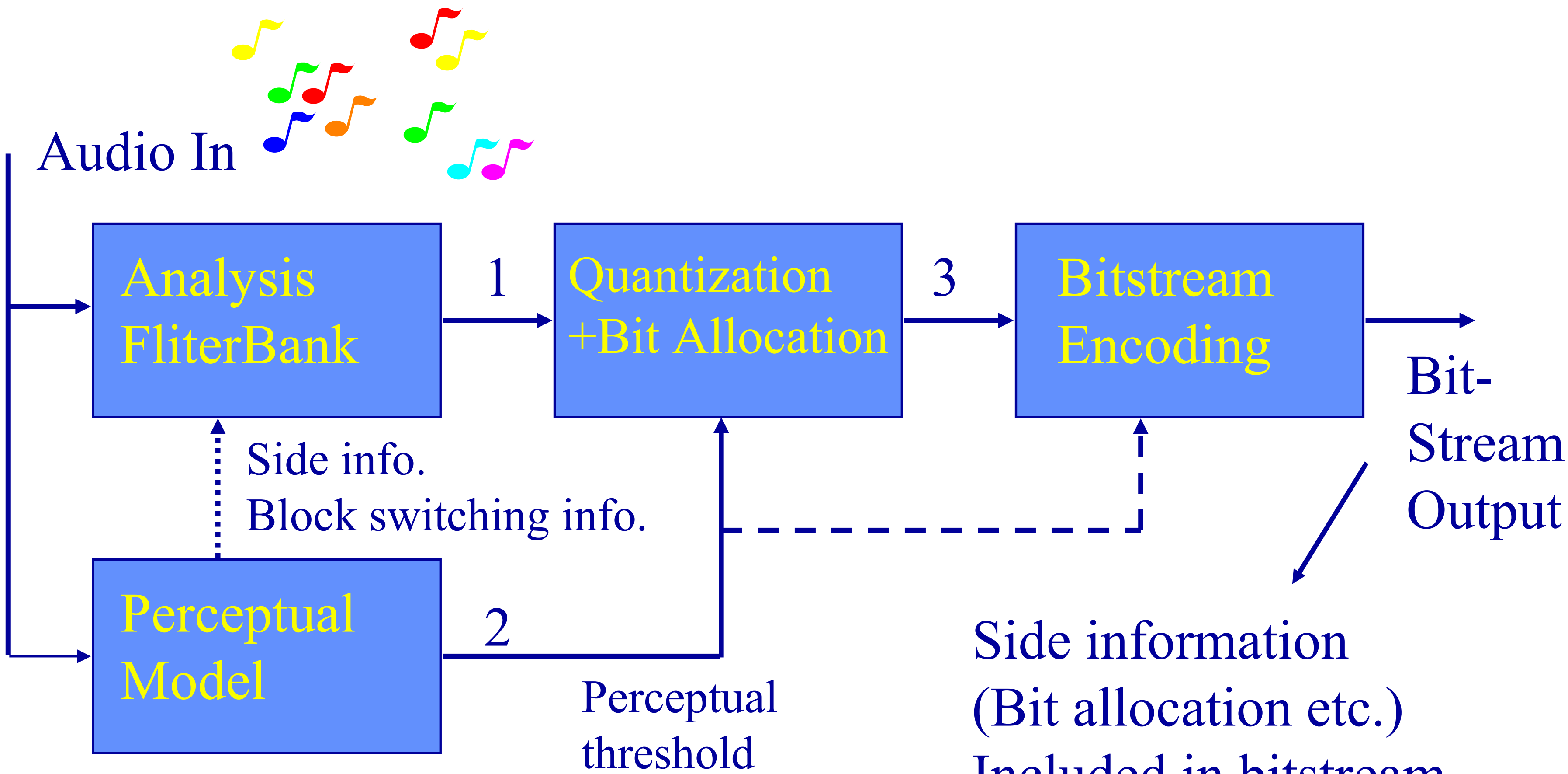
Measures

- SMR: Signal to Mask Ratio
- SNR: Signal to Noise Ratio
- MNR = SNR - SMR

Is the ratio between the mask energy and the quantization noise

- Positive MNR: The noise injected in the quantization process is higher than masking level, and will be heard after reconstruction

Basic Coding Structure



Basic Coding Structure (Cont'd)

1. Input signal is decomposed into sub-sampled spectral components (time/freq. domain)

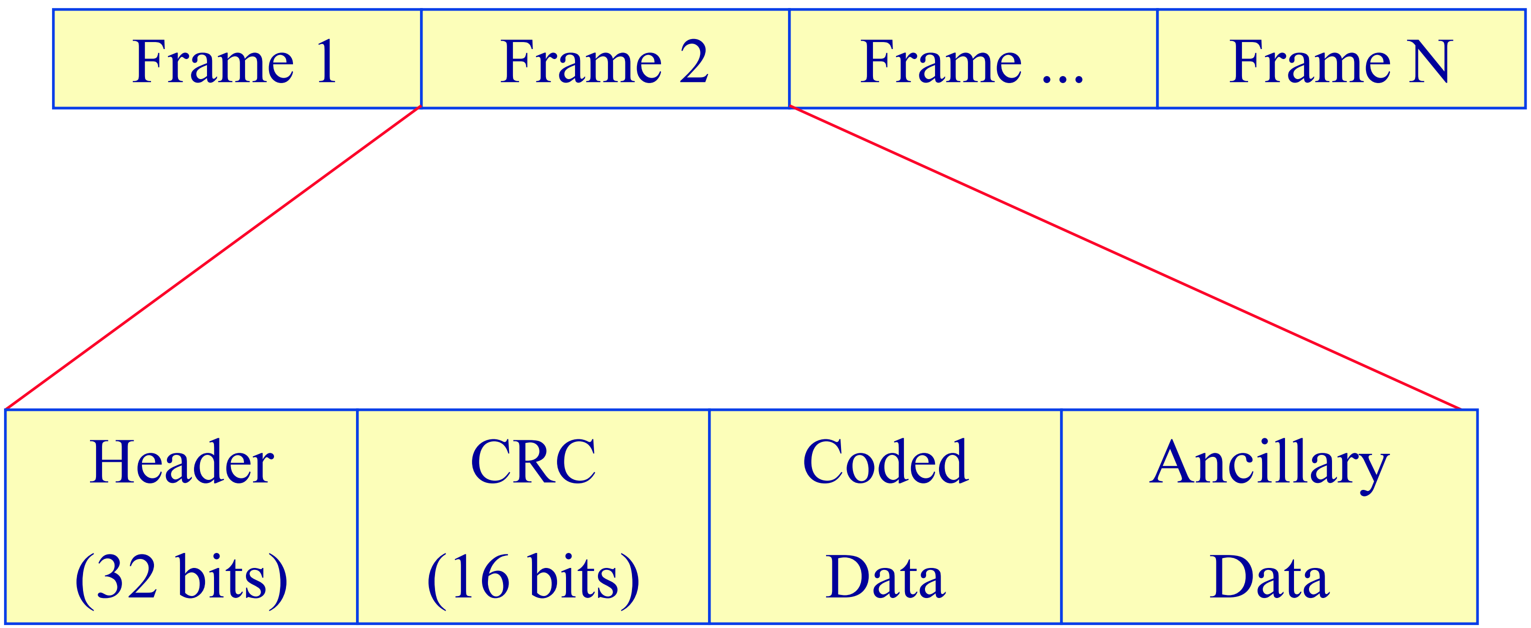
2. A time-dependent mask threshold is estimated

3. Spectral components are quantized and coded, keeping the noise (introduced in quantization) below the mask threshold (Many implementations)

- Bit-Allocation:

1 bit of quantization introduces about 6 dB of noise

Bit-Stream Structure



Header Contents

- *SyncWord* (12 bits)
- *Layer Code* (2 bits): Layer I, II or III
- *Bit-rate Index* (4 bits): according to the table in next slide, 32Kbps up-to 448Kbps.
- *Sampling Frequency* (2 bits): 48, 44.1 or 32kHz.
- *Padding* (1 bit): number of slots, N or $N+1$
- *Mode* (2 bits): Stereo, Joint Stereo, Dual or Single Channel.

Available Bit-Rates (Kbps)

Index	Layer I	Layer II	Layer III
0000	free format	free format	free format
0001	32	32	32
0010	64	48	40
0011	96	56	48
0100	128	64	56
0101	160	80	64
0110	192	96	80
0111	224	112	96
1000	256	128	112
1001	288	160	128
1010	320	192	160
1011	352	224	192
1100	384	256	224
1101	416	320	256
1110	448	384	320

Perceptual Model

- A good estimation of actual masked threshold is essential for a better quality
- A very simple model would allocate bits according to : n_bits=(27dB*lu-lo) / 6.02dB

  lu: upper band limit lo: lower band limit

  (Measured in Bark)
- More advanced models (SMR) are in use

Perceptual Model: reminder

- The masking occurs in each critical band.
- Critical band represents the bandwidth at which subjective response change rather fast.
- The bandwidth of the critical bands varies from 100Hz at low frequencies to about ($0.2 \times f$) for frequencies above 500Hz.
- The loudness of a band of noise at a constant sound pressure remains constant in the critical band.
- The corresponding unit for the critical band is bark.

Filter Banks

- Sub-Band Coders use low number of channels, connected with processing of <u>adjacent samples in time</u>
- Transform Coders use high number of sub-bands and <u>joint</u> processing of <u>adjacent samples in frequency</u>

No Basic difference between both approaches

Quantization

Two basic approaches

- Block Companding (block floating point):

A number of values, ordered either in time domain or in frequency domain are normalized to maximum absolute value (by scale factor)

  - Number of bits **allocated for the block** (derived from the perceptual model) derives the quantization step size

Quantization (Cont'd)

- Noise allocation + Scalar Quant. + Huffman:

Instead of bit allocation, an amount of allowed noise equal to the estimated masked threshold is calculated for each scale-factor sub-band

Quantization noise is colored using scale factors, by changing quantization step size

- Quantized values are Huffman coded
- Process is controlled by iteration loops

MPEG-1 System

- First international standard for digital audio compression
- Joint effort of ASPEC (AT&T, CNET ...)

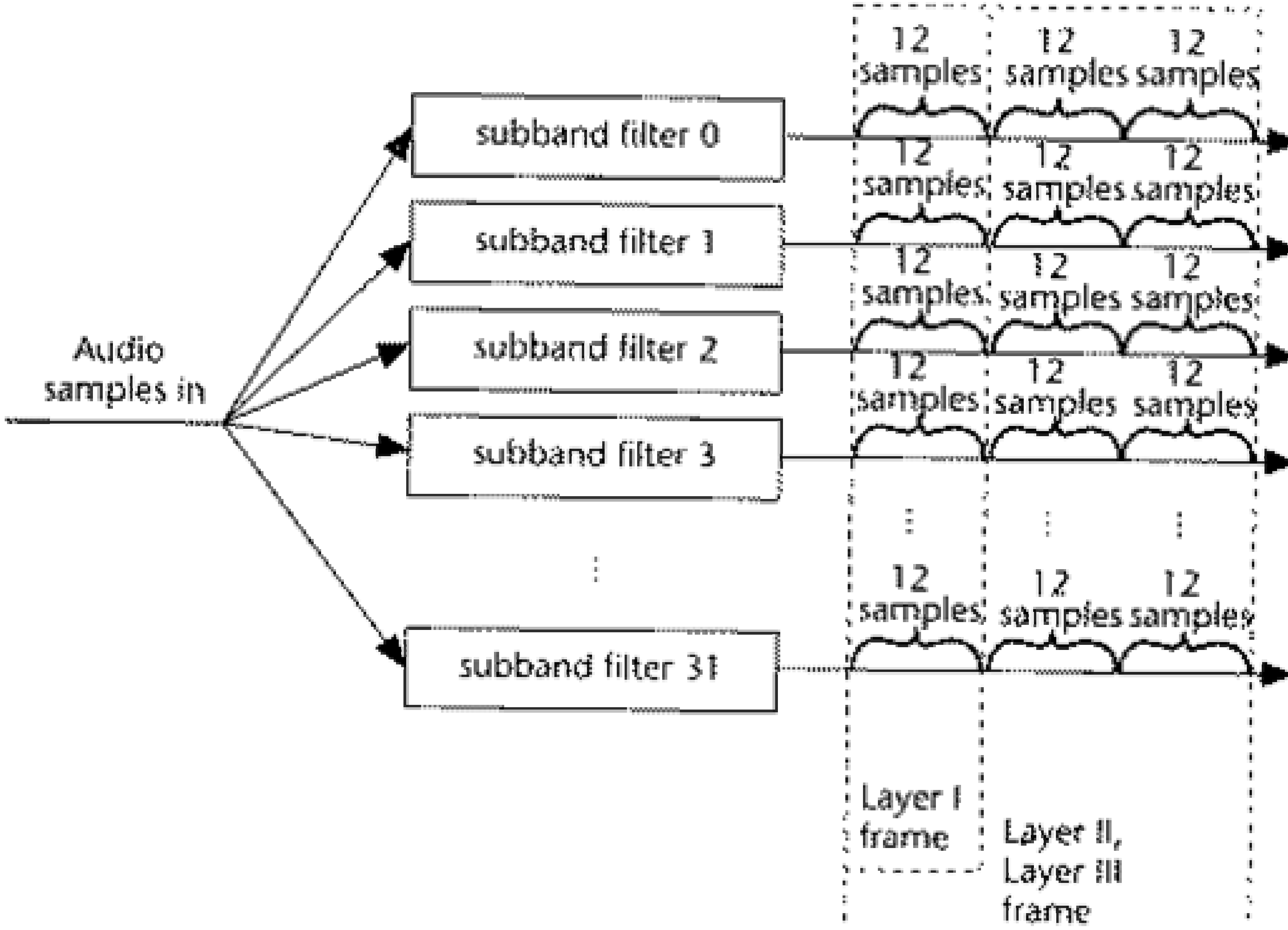
and MUSICAM (Philips, Matsushita ,....)

- A three layer coding algorithms defined with main system properties are increased complexity (encoder mainly) and quality (at low bit-rates)

MPEG-1 Layers

- MPEG defines 3 layers for audio.

  Basic model is same, but Codec complexity increases with each layer.
- Data is divided into frames, each of them contains 384 samples, 12 samples from each of the 32 filtered sub-bands as shown in the next slide
- All layers share definition of basic bitstream format (4 bytes header, sync. Etc.)


Audio samples in
subband filter 0
subband filter 1
subband filter 2
subband filter 3
subband filter 31
12 samples
Layer I frame
Layer II, Layer III frame

MPEG-1 , Layer 1

- Input signal transformed into 32 uniform sub-bands (same frequency width for each band).
- For each sub-band an adaptive bit allocation (based on PA model) and quantization
- Psychoacoustic model uses only frequency masking.
- No control on the amount of noise introduced for each sub-band : bit allocation continues until needed output rate achieved

Layer I : more details

- Filter bank: Equally spaced Polyphase filter: design flexibility of generalized QMF and low computational complexity
- 511 tap prototype used, optimized for very steep response, and stop band attenuation better than 96dB (equivalent to 16 bit resolution)
  - reconstruction error: LSB (of 16) if no quantization
- Impulse response of 10.6ms (@48kHz)
- Time resolution: 0.66ms (@48kHz)
- The prototype filter keeps pre-echo artifacts !

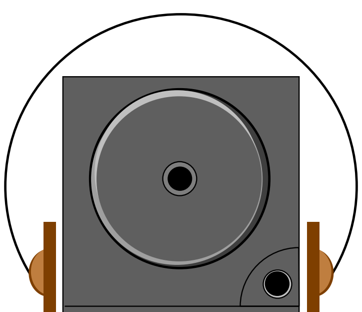
Layer I : more details (Cont'd)

- Quantization step uses block companding of 12 subband samples
  - Basic block length: 12*32=384 samples
- A 4-bit field signals the bit allocation: 0-16 bits for each subband
- A 6-bit field scale factor (G) for each band,
  - the exponent of the block companding quantization
- This method allows changes in the bit allocation procedure

Layer 1 features

- A simplified version of MUSICAM
- Appropriate for consumer applications such as studio use

  (where very low data rated not necessary)
- Compatible with PASC by Philips
- Basic frame length: 8mSec (for 48KHz rate)



MPEG-1, Layer 2

- Further compression, by removing redundancy (a little bit of the temporal masking) and a more precise quantization
- Basic frame length: 24mSec (for 48KHz sampling freq.): 1152 samples
- Identical to MUSICAM (Except frame header)
- Application fields: consumer and professional studio-like broadcasting, recording, multi-media, audio workstations.

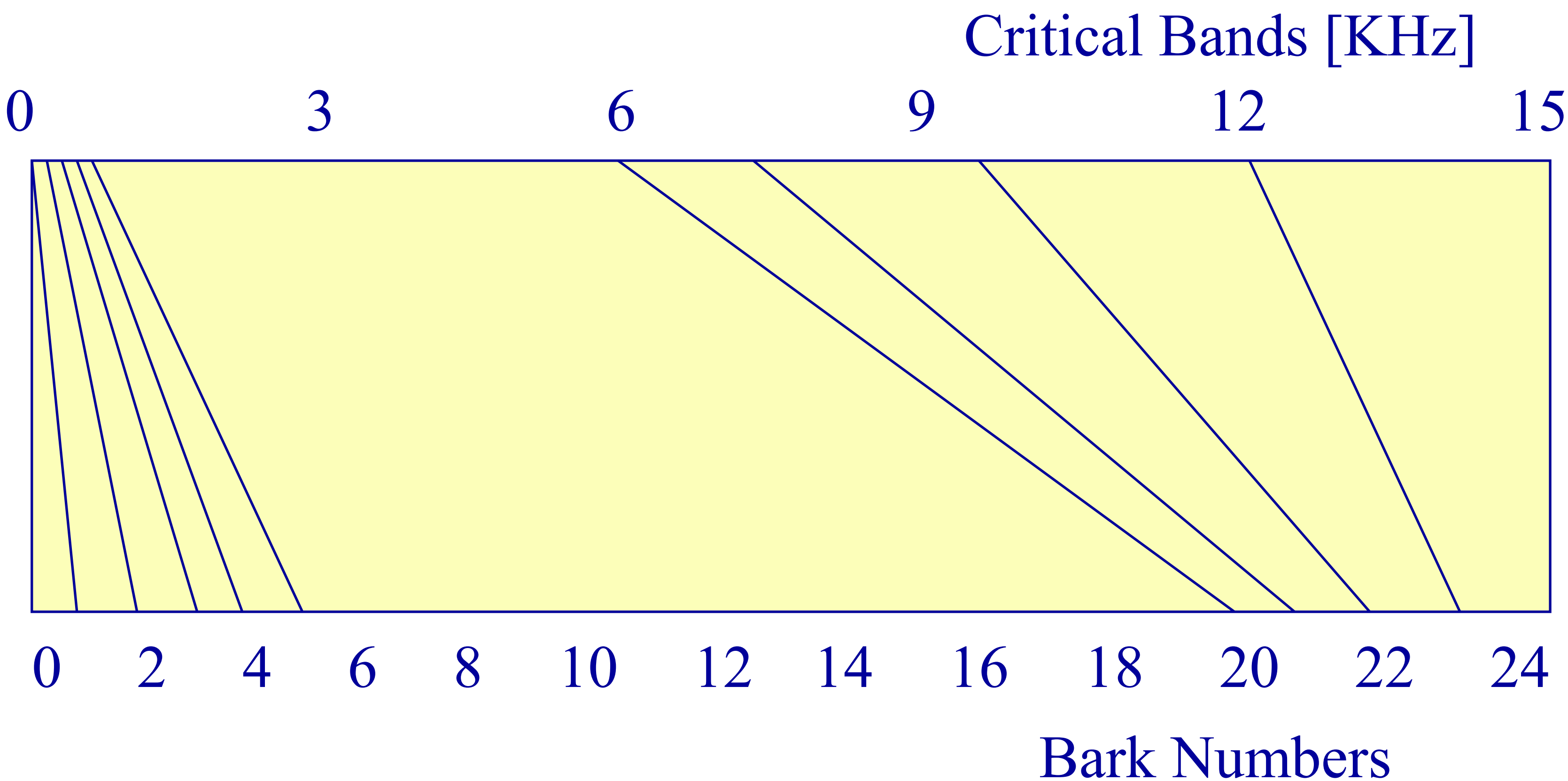
MPEG-1, Layer 2 (Cont'd)

- Additional coding of bit-allocation, scale factors and different frame structure.
- Encoder forms larger groups of 3 blocks, 12 samples/block, and 32 sub-bands (total of 1152 samples per frame).
- One bit-allocation type and 3 scale factors for every 3 blocks frame.
- Radix coding allows allocation of fractional bits for small quantized values

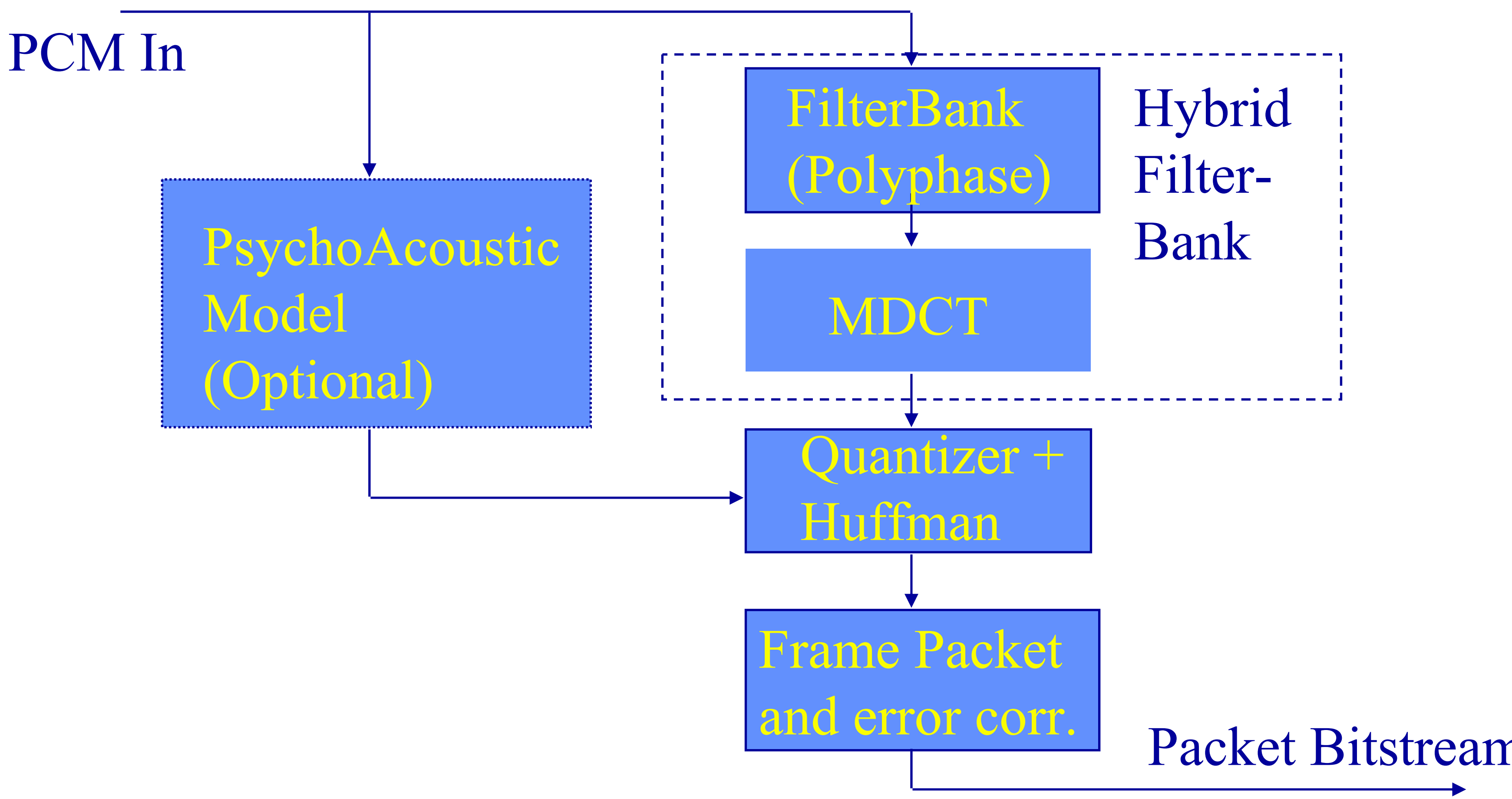
MPEG-1 , Layer 3 (mp3)

- Signal mapping resolution is increased (in freq. domain: non-equal frequencies)
- Signal is divided into “critical bands”, according to human ear resolution
- Adaptive allocation of noise to each critical band, and logarithmic quantization
- Further compression by Huffman coding
- PA model includes temporal masking effects, takes into account stereo redundancy.

Non-Linear Critical Bands



Layer 3 Basic Scheme



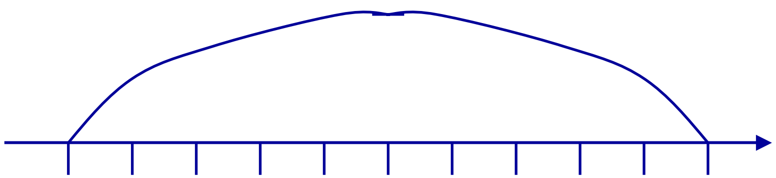
Layer 3 Basic Scheme (Cont'd)

- PolyPhase FilterBank: Filters the input into 32 time domain signals, representing 32 uniform frequency bands.
- MDCT: Transforms each band into freq. domain, getting 576 spectral lines (32*18 samples): additional frequency resolution: 18 sub-subbands
- PA Model: Analyses the input, and controls the quantization step size
- Quantizer + Coder: Quantizes according to PA and needed rate + Huffman coding

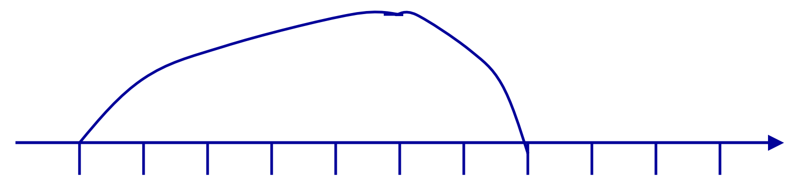
Layer 3 Basic Scheme (Cont'd)

- Adaptive block switching: Dynamic switching of the time-frequency decomposition (filter bank resolution) is allowed
- This is important in order to ensure that the time spread of the filter bank does not exceed the pre-masking period (to avoid pre-echo)
- Adaptive window switching uses four optional windows: normal (long), start, short and stop.

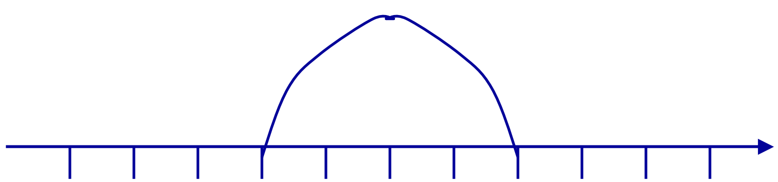
Window types



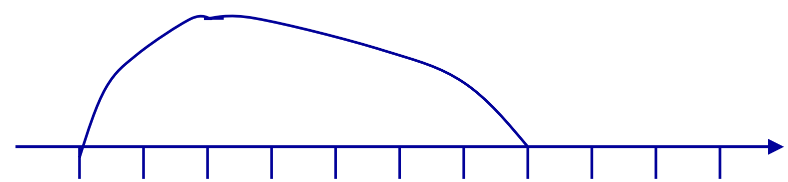
Normal window
for stationary signals
576 spectral lines



Start window
to switch from long to short
right 1/3 is zero to cancel aliasing

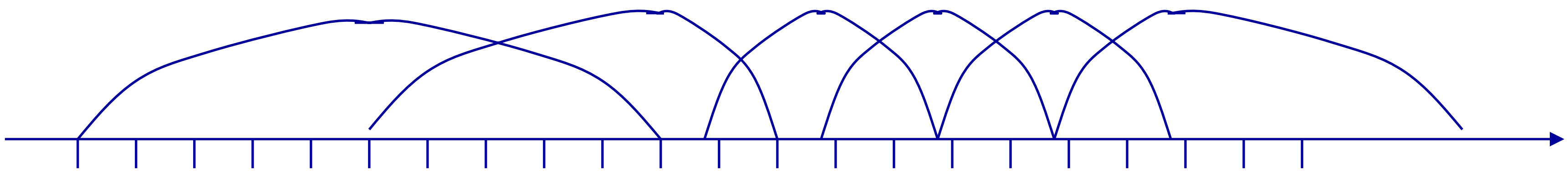


Short window
1/3 length followed by 1/3 length MDCT
time resolution: 4ms (@48KHz)
192 spectral lines



Stop window
to switch from short to long
left 1/3 is zero to cancel aliasing

Example sequence of window forms



Layer III Quantization and Structure

- Non-uniform quantization and VLC (Huffman)
- A-by-S iteration loop
- No bit direct allocation, but ‘Noise allocation’ (indirect allocation) using two iteration loops

Iteration loops

- Two nested iteration loops:
  - Rate loop: Inner iteration for quantization and coding of spectral lines (using Huffman tables) - repeats with increasing step size, until number of allocated bits does not exceed the allowed maximum.
  - Distortion Control loop: Outer iteration, keeping quantization noise below masking threshold according to the PA model
    - Noise coloration: scale factors reduced until injected noise is small enough

Layer 3 Features

- Short Blocks of 12 samples (in addition to regular 36 samples blocks) improve time resolution to cope for transients.
- PA Model and coding technique are NOT part of the standard - related information should be included in bit-stream
- Optional variable rate mode
- Application in Telecommunication, mainly narrow-band ISDN, satellite links , Internet DVD, etc.

Joint Stereo Coding

- Can be used for stereo redundancy reduction.
- Stereo and Dual-Channel signals require twice the bandwidth if we code them separately.
- To decrease bit-rate (or increase quality) we can use intensity stereo mode or Middle/Side (MS) stereo coding.
- MS stereo coding is supported only in Layer III

Joint Stereo Coding (cont'd)

<u>Intensity stereo coding:</u>

- instead of separate L and R subband samples, a single summed signal is transmitted with R and L Scale Factors.
- The frequency spectra of the decoded stereo signals are the same but the magnitudes are different.

Joint Stereo Coding (cont'd)

Middle/Side Stereo Mode:

- Middle (sum of L and R) and Side (difference of L and R) are transmitted instead of L and R.
- M is transmitted in the L channel and S in the R channel.
- R and L channels can be reconstructed using:

$$\boxed{L = (M + S) / \sqrt{2} \qquad R = (M - S) / \sqrt{2}}$$

Effectiveness of MPEG audio

Layer #	Target bitrate	Ratio	Quality @ 64 kbits	Quality @ 128 kbits	Theoretical Min. Delay
1	192 kbit	4:1	---	---	19 ms
2	128 kbit	6:1	2.1 to 2.6	4+	35 ms
3	64 kbit	12:1	3.6 to 3.8	4+	59 ms

5 = perfect, 4 = just noticeable, 3 = slightly annoying, 2 = annoying, 1 = very annoying

More About the PA Model

- The difference between max. signal level and min. masking threshold is used in the bit or noise allocation to determine Q level
- Two models given in the informative part of the standard. model 1 recommended for layers 1,2 and model 2 for layer 3
- PA model output is SMR for each band (L1, L2) or group of bands (L3)

PA Model I (Layer 1,2)

- Transform length (FFT) is 512 samples for layer 1 and 1024 for layer 2.
- The filter bank suffers lack of selectivity at low frequencies.
- To compensate it: FFT in parallel to sub-band filtering.
- Sound Pressure Level (SPL) is computed for each band.
- Tonal and non-tonal components are extracted from the power spectrum.

PA Model I (Cont'd)

- Using "decimation", number of maskers is reduced:

only components (tonal and non-tonal) greater than the absolute threshold are considered.

- Two or more components that are smaller than the highest power within the distance of 0.5 bark are removed from the list of tonal components.
- Masking thresholds (both *t* and *non-t*) are defined by adding the masking index and masking function to the masking component (both index and function are provided in the standard as formal equations)

PA Model I (Cont'd)

- Global masking threshold, *LTg*, (for the frequency component) is derived by summing the powers of the individual masking thresholds (tonal: LT_{tm}, *non-tonal* LT_{nm}) and the threshold in quite:

$$LTg(i) = 10\log_{10}\left[10^{LT_q(i)/10} + \sum_{j=1}^{m} 10^{LTtm(j,i)/10} + \sum_{j=1}^{n} 10^{LTnm(j,i)/10}\right] \text{[dB]}$$

PA Model I (Cont'd)

- To determine the Signal-to-Mask Ratio (SMR) in sub-band *n*, the minimum global masking threshold *LTmin* is used:

$$SMR_{SB}(n) = L_{SB}(n) - LT_{\min}(n) \quad [\mathrm{dB}]$$

Where $L_{SB}(n)$ is the signal component in sub-band *n* .

PA Model II: Layer 3

- The size of FFT (+ *Hann* window) can be varied. In practice: model is computed twice in parallel (192 samples for short block and 576 samples for long block).
- Masking in time (forward and backward) is taken into calculations (spreading energy).
- Final energy threshold obtained by the convolution (via FFT) of "spreading" energy and partitioned original energy.

PA Model II (Cont'd)

- SMR is calculated by the ratio between energy in the "scale factor" band (e_part_n) and the noise level in the scale factor band (n_part_n):

$$SMR_n = 10\log_{10}(e\_part_n / n\_part_n) \quad [\mathrm{dB}]$$

n: index of coder partition

<u>Scale factor:</u> the maximum of the absolute values of 12 samples in a sub-band is determined. (6 bits)

MPEG-2 Audio

- Backwards compatible - defines extensions:
  - – MultiChannel coding
    - 5 channel audio (L, R, C, LS, RS)
  - – Multilingual coding
    - 7 multilingual channels
  - – Lower sampling frequencies (LSF)
  - – Optional Low Frequency Enhancement (LFE)

MultiCahnnel Coding

- Up to 5 audio channels Matrixation of channels for compatibility:

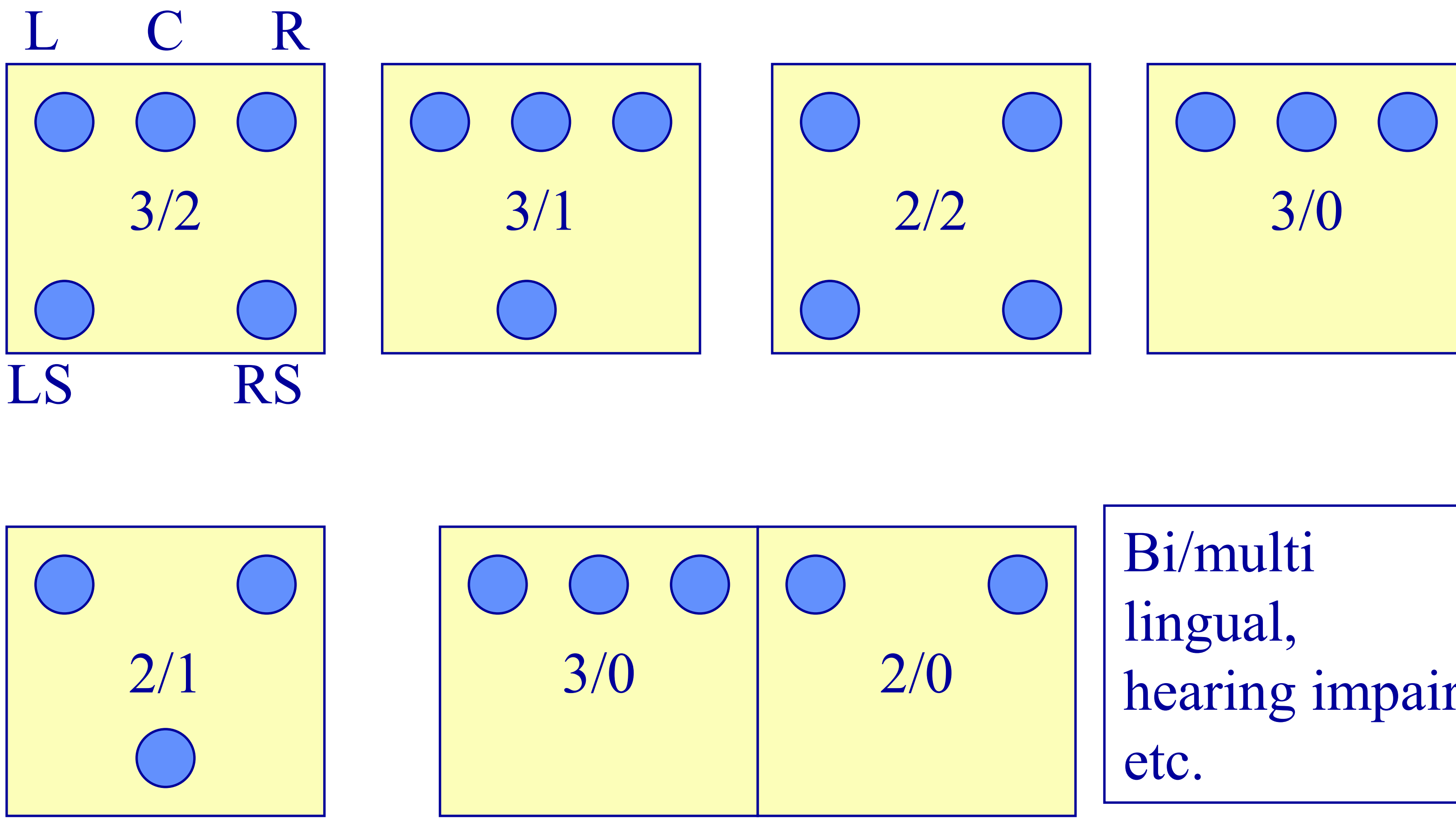
$$Lc = b\left[L + C/\sqrt{2} + a \cdot Ls\right] \qquad Rc = b\left[R + C/\sqrt{2} + a \cdot Rs\right]$$

$$b = \frac{1}{1 + 1/\sqrt{2} + a} \qquad a = 1/\sqrt{2};\ 1/2;\ 1/2\sqrt{2};\ 0$$

C: center Ls,Rs: surround

- Lc and Rc are MPEG-1 encoded
- Layer 1,2: Use syntax of MPEG1-L2
- Layer 3: flexible number of extension channels

Multi-channel Configurations

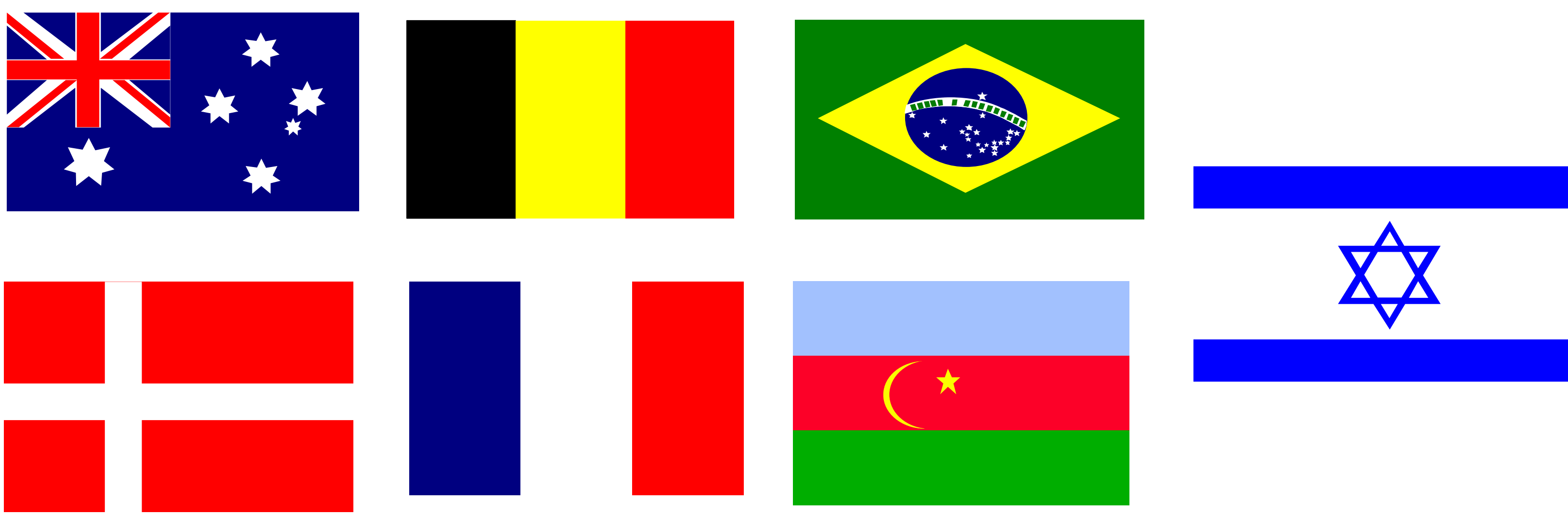


Bi/multi lingual, hearing impaired etc.

And more options...

Multilingual Coding

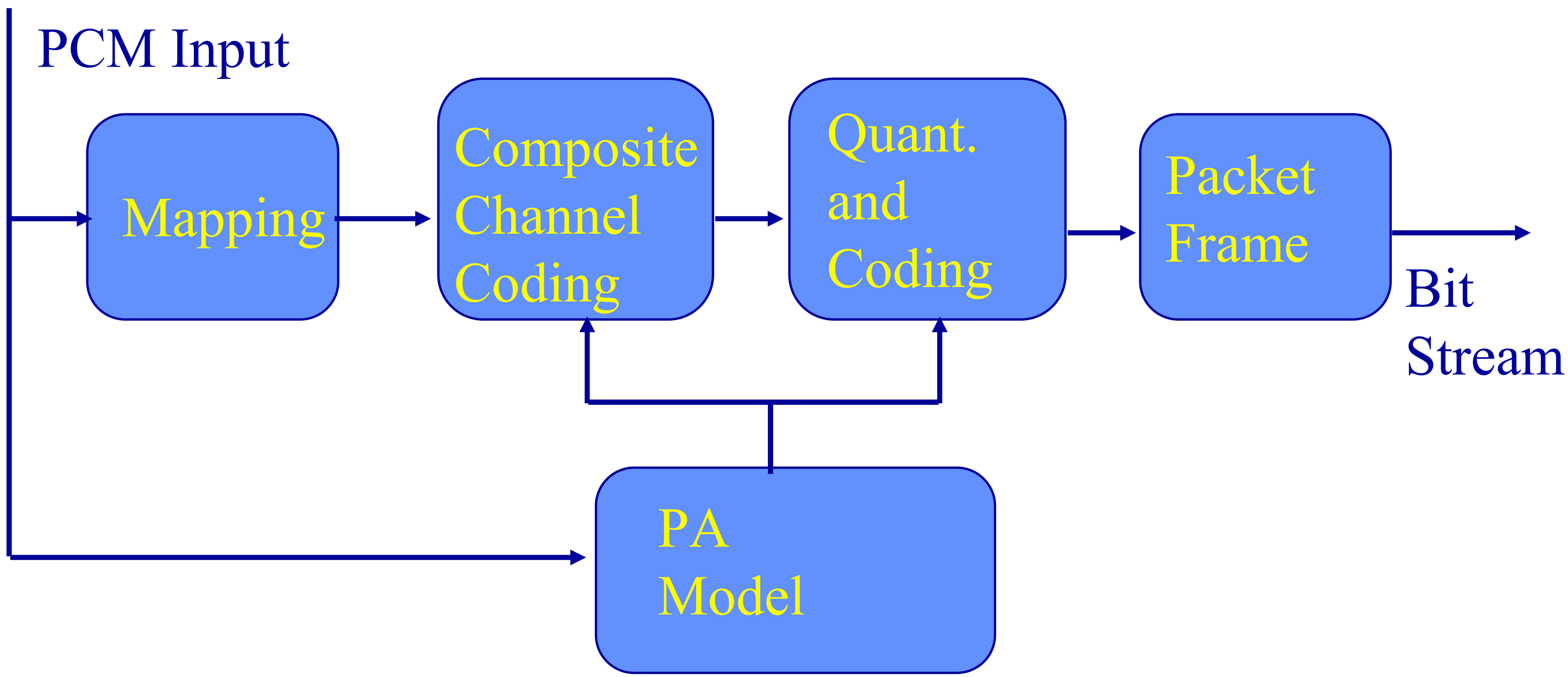
- Up to 7 additional channels for multi-lingual purposes



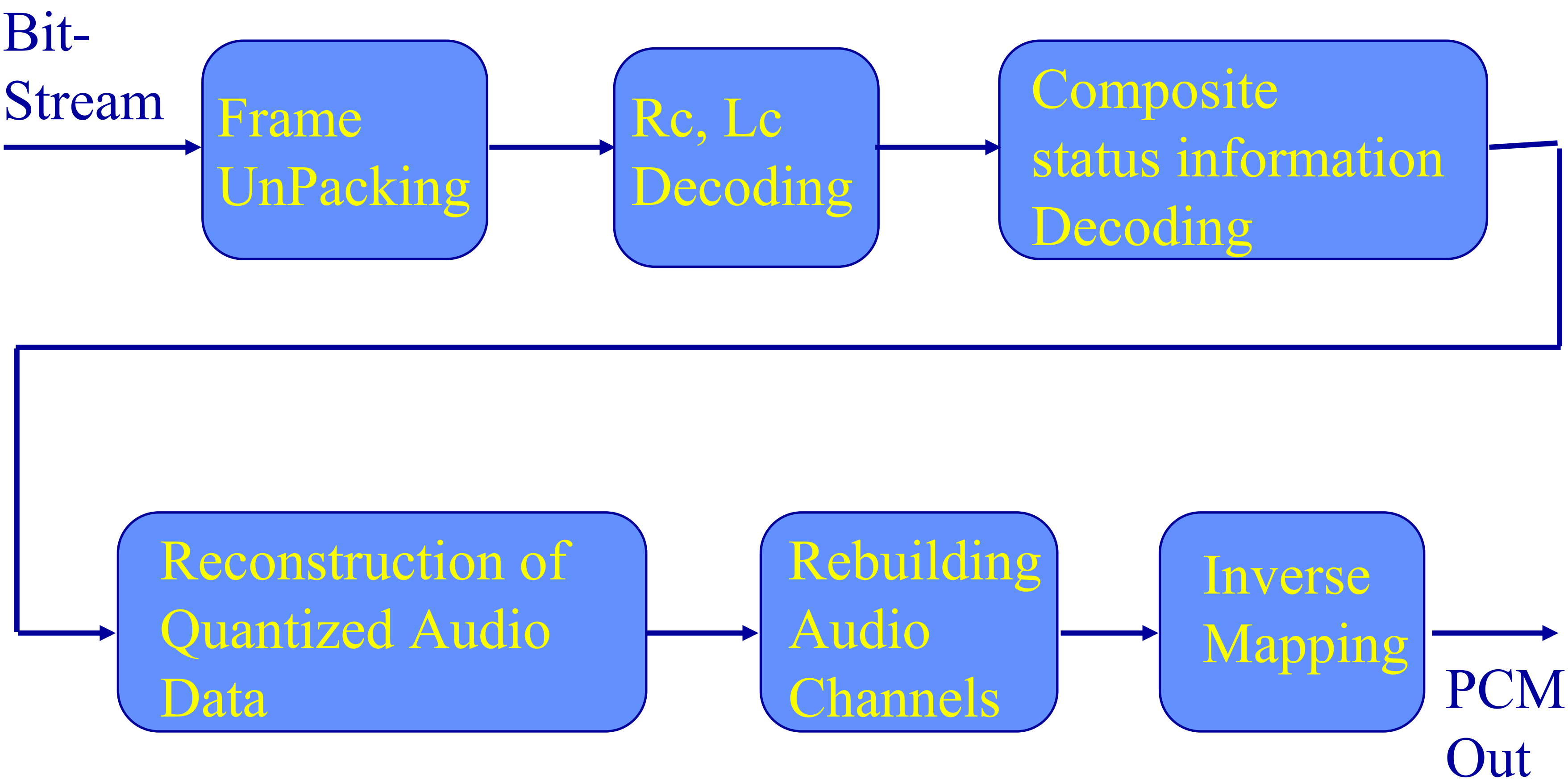
Low Sampling Frequency Coding

- For narrow band frequencies, no need for high sampling rates (wide-band speech and medium quality audio)
- Added sampling rates are the halves of the MPEG-1 rates: 16K, 22.05K and 24KHz
- Need to change PA model tables
- Optional sixth channel: LFE capable of handling signals from 15Hz to 120Hz (Sub-Woofer), added to 5 regular channels

Encoder Scheme



Decoder Scheme



MPEG-2, Layer-I extensions

- A “slot” consists of 32 bits.
- The number of slots in frame depends on the sampling frequency and bit-rate.
- Each frame contains information on 384 samples of the original input signal.
- $Frame\_size = 384 * (1/f_s)$ (16mSec for fs=24KHz)
- $Num\_of\_slots = bit\_rate * (384/32)/f_s$

  (32 for fs=24KHz)

MPEG-2, Layer II extensions

- Difference from MPEG-1 only in formatting, possible quantization and PA model.
- A slot consists of 8 bits.
- Each frame contains information on 1152 samples of the original input signal.

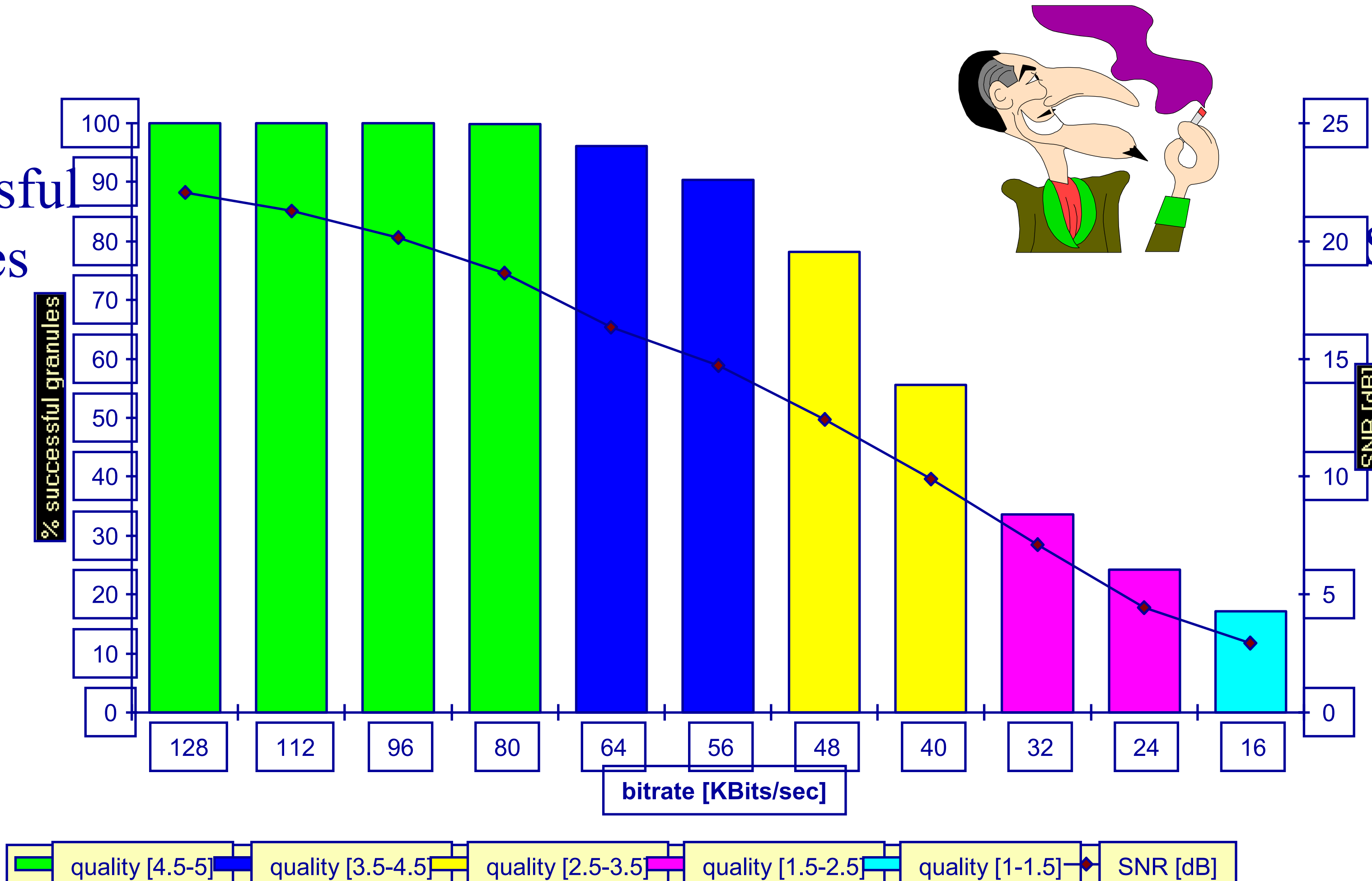
MPEG-2, Layer III extensions

- Different scale factor band tables.
- Omission of some side information (due to changed frame layout).
- Some changes in PA model tables.
- 21 Scale factors bands for each f_s (long windows)
- 12 Scale factors bands for each f_s (short windows)

- Scale factor band: a set of frequency lines that are scaled by the same scale factor.

Witches: “Dingo” at 22.05Khz

Successful granules



SNR