UNIVERSITÀ DEGLI STUDI
DI MODENA E REGGIO EMILIA

Dispense del corso di
Sistemi di Elaborazione Multimediali

Advanced Audio Coding

Last update: 28/05/2018

Advanced Audio Coding

- AAC is a lossy audio coding format designed to be the MP3 successor.
- It is defined in ISO documents ISO/IEC 13818-7 and ISO/IEC 14496-3 as part of the MPEG2 and MPEG4 standards.

- It introduces many improvements but the overall structure of the encoder is still very similar to that of an MP3 encoder.

- Both MP3 and AAC are *perceptual audio coding systems*.

Main features

- The AAC format can work with audio signals sampled using sampling rates ranging from 8 kHz to 96 kHz.
- It is a multichannel format that supports up to 48 regular audio channels and up to 16 additional LFE channels (Low Frequency Effect) dedicated to audio signals in the 3 – 120 Hz band.
- AAC supports both CBR (Constant Bit Rate) coding and VBR (Variable Bit Rate) coding.
- The encoding exploits many perceptual and psychoacoustics techniques to achieve the best compression results, such as:
  - Block switching
  - Prediction
  - Temporal Noise Shaping
  - Masking effects
  - Non-uniform quantization
- Huffman coding is used in the last encoding steps (noiseless coding).

Perceptual audio coding

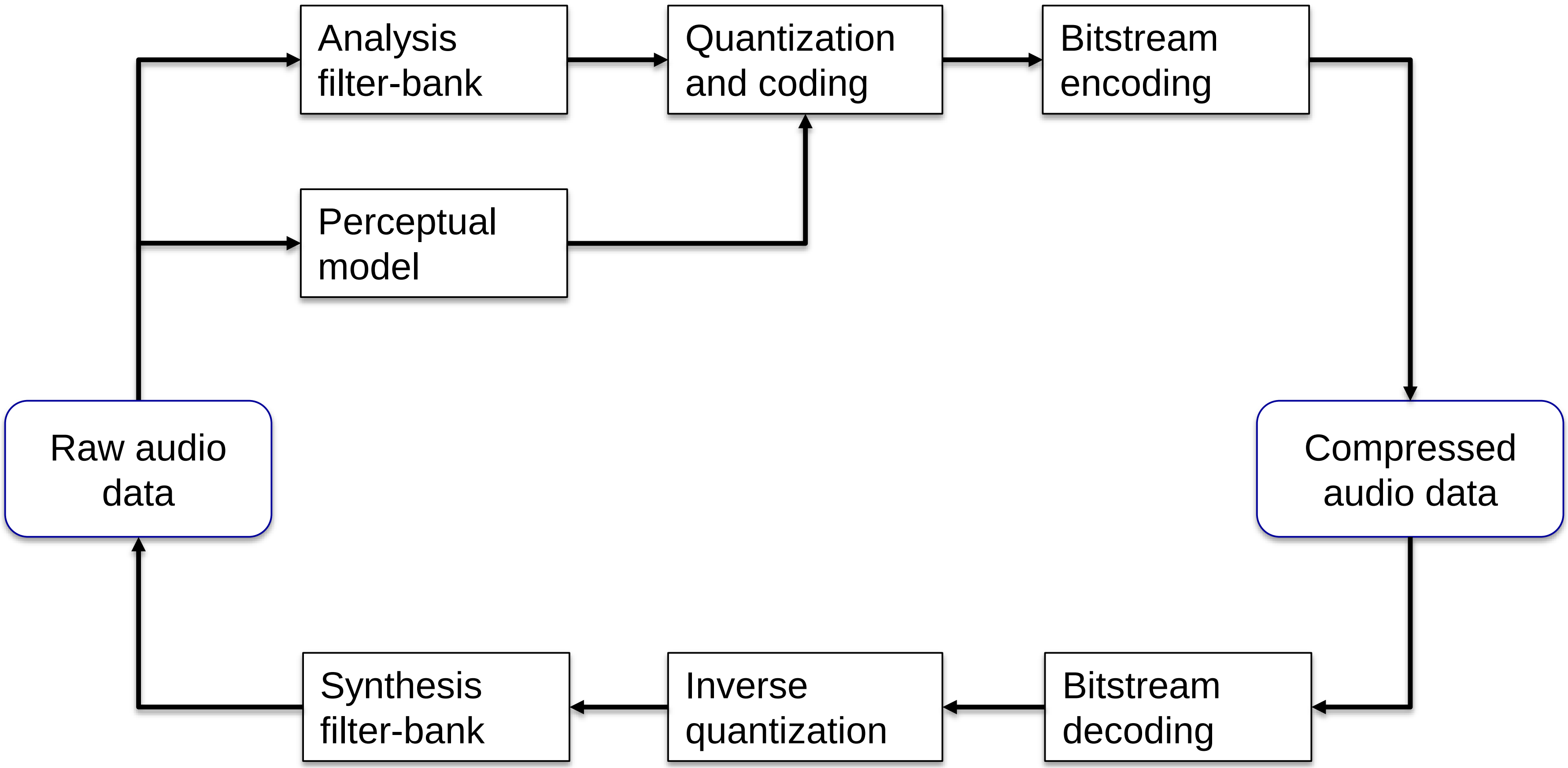
- The basic task of a perceptual audio coding system is to compress the digital audio data in a way that:
  - the compression is as efficient as possible, i.e. the compressed file is as small as possible and
  - the reconstructed (decoded) audio sounds exactly (or as close as possible) to the original audio before compression.

- Other requirements may include:
  - Low complexity (to make encoding and decoding as inexpensive as possible for both software and hardware).
  - Flexibility (to make the encoder usable in different scenarios).

- Perceptual audio coding is a **lossy** compression technique.

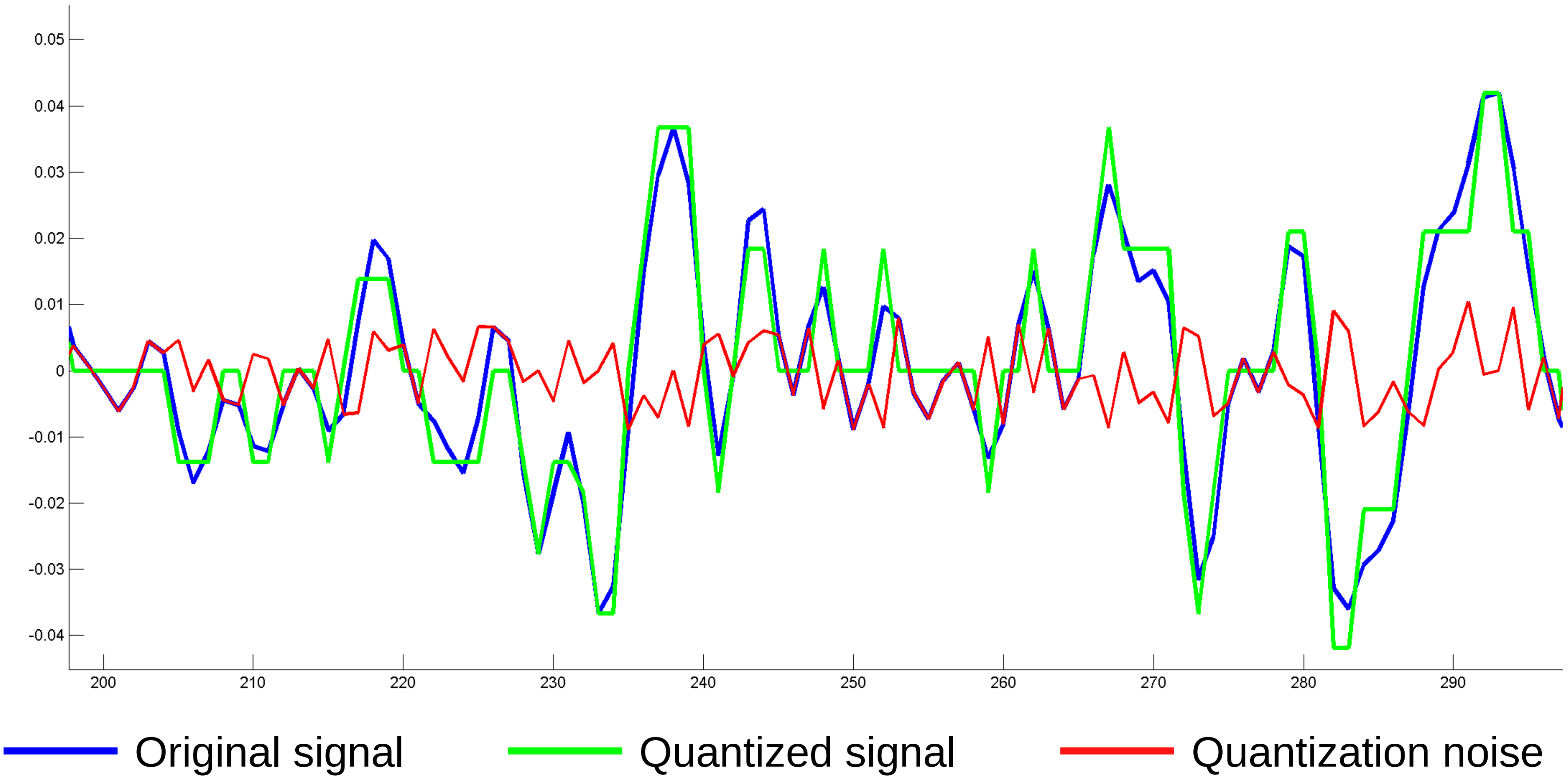
Perceptual encoder/decoder basics

- A basic perceptual coding system is composed of the following blocks:



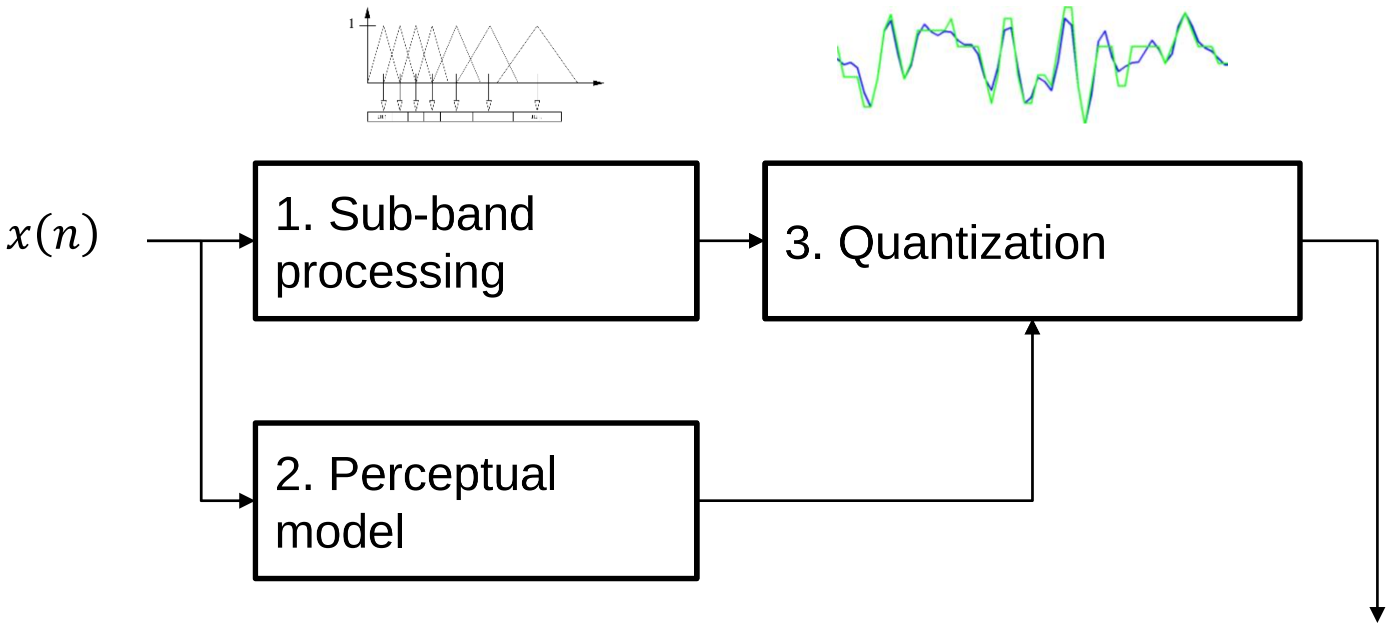
Quantization noise

- The quantization process introduces noise in the reconstructed signal
- The goal is to *shape* the introduced quantization noise to reduce the noise *perceived* by the human auditory system.
- Psychoacoustics and perceptual models are employed.



Problem Definition

- **Goal**: minimize perceptual error (PE) by shaping the noise



$$x_q(n) = x(n) + e(n)$$

$$\arg\min_{e(n)} PE(x(n), e(n))$$

Question

- Let us listen to two different error signals

A

B

- Can you guess which one is associated to the best compression algorithm? Why?
- Keep in mind our goal:

$$\arg\min_{e(n)} PE(x(n), e(n)) \qquad x_q(n) = x(n) + e(n)$$

- Answering is easier when listening to the compressed signal

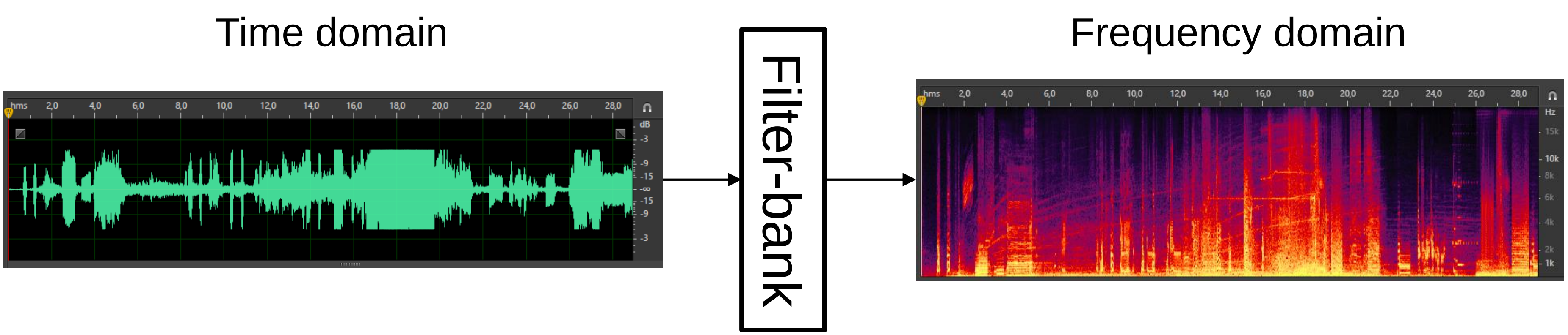
A

B

SUB-BAND PROCESSING

Sub-band analysis filter-bank

- The analysis filter-bank transforms the input audio signal in its spectral representation.
- The input audio is analyzed by means of a sliding window of N samples in order to produce N spectral coefficients.
- The spectral analysis has many purposes:
  - Spectral coefficients are used to calculate masking information.
  - Spectral coefficients will be quantized and coded in the final bit stream (as in a JPEG encoder).



MDCT and IMDCT

- AAC uses the *Modified Discrete Cosine Transform* (MDCT) to calculate spectral coefficients (in other words: MDCT is used as a filter-bank):

$$X_k = \sum_{n=0}^{2N-1} x_n w_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2} + \frac{N}{2}\right)\left(k + \frac{1}{2}\right)\right] \qquad k \in [0, N-1]$$

- The inverse transform used to obtain samples from the spectral coefficients is called *Inverse MDCT* (IMDCT):

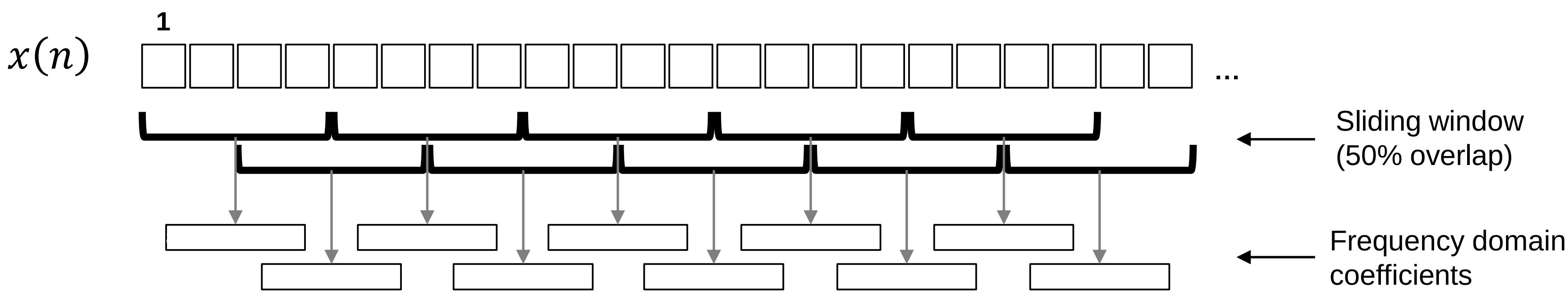
$$y_n = \frac{2}{N} w_n \sum_{k=0}^{N-1} X_k \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2} + \frac{N}{2}\right)\left(k + \frac{1}{2}\right)\right] \qquad n \in [0, 2N-1]$$

- The MDCT produces N coefficients from 2N input values.
- Viceversa the IMDCT produces 2N values from N spectral coefficients.

- To obtain N coefficients from N samples the *overlap-add* technique is used.

MDCT and IMDCT

- MDCT is a modified version of the type IV of the Discrete Cosine Transform (DCT-IV).
- It is also a *lapped transform*, which means that it is performed over 50% overlapped blocks of the input signals.

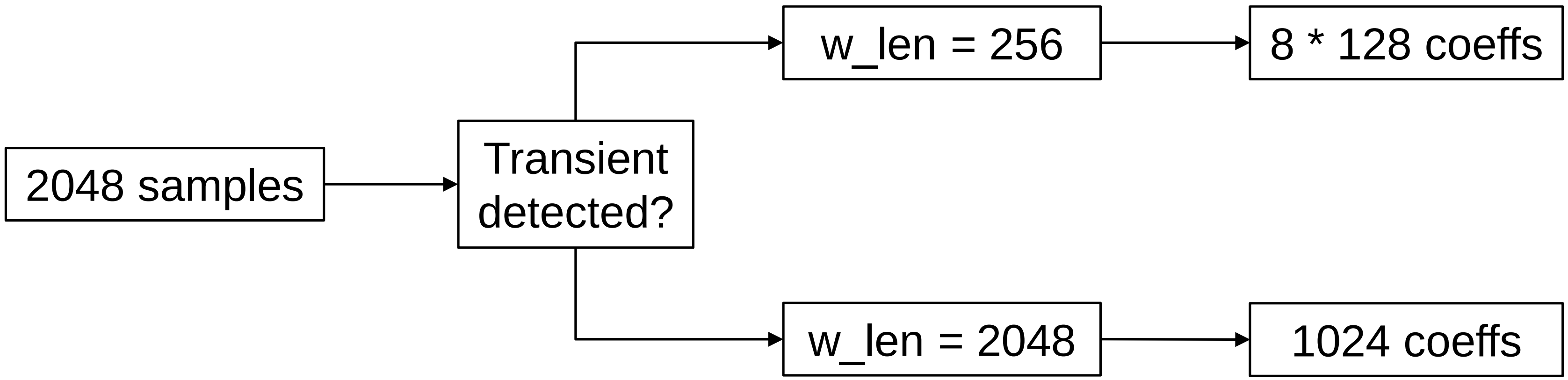


- Why MDCT?
- Being a lapped transform blocking artifacts are reduced.
- Has a property called *time-domain aliasing cancellation* (TDAC), the aliasing introduced is removed by the inverse transform, hence it permits *perfect reconstruction* of the original signal.

Window length

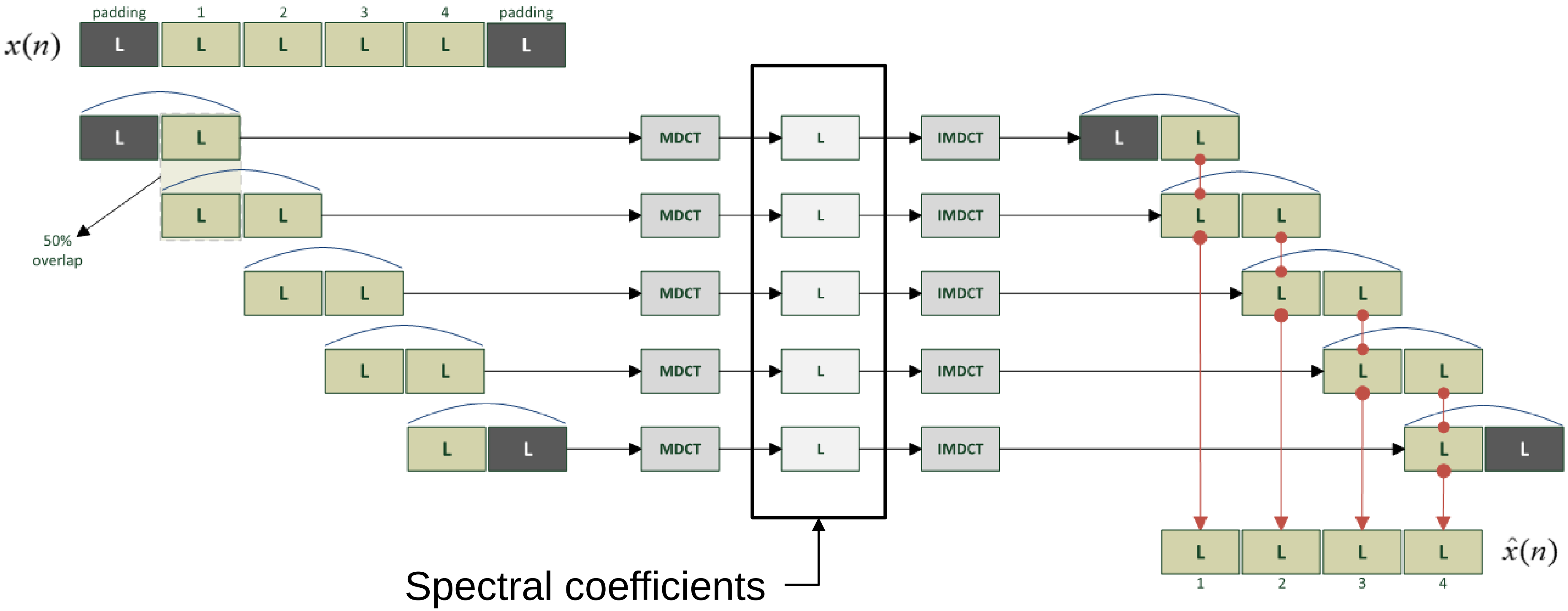
- An AAC encoder always processes a 2048 audio samples at a time. A block of 2048 samples is called *frame*.
- The MDCT can be applied on sample windows of different length.
- AAC supports two different window lengths:
  - 2048: the regular size
  - 256: used when a transient is detected in the audio samples, i.e. a drum hit.
- This technique is called *block switching*, it is employed to better represent short and sudden variations in the audio samples.

- For each frame processed by the encoder two situations may occur:



Analysis/synthesis via MDCT

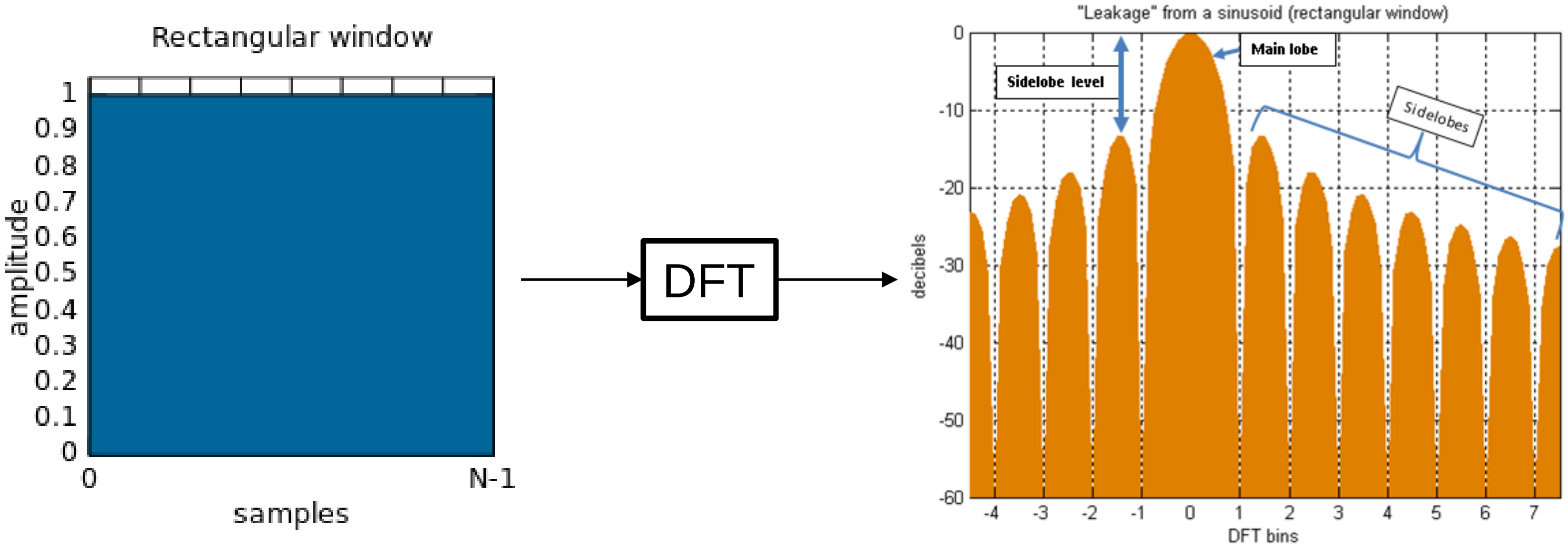
- The input audio signal is padded with L zeros at the beginning and at the end in order to not lose any information around the edges.
- The MDCT is applied on subsequent couples (*overlap* operation) of L samples each in order to produce L coefficients.
- Audio samples values are reconstructed by *adding* the second half of the previous reconstructed window with the first half of the current one.



Blocking artifacts

- Processing 2L samples at a time is the same as transforming the input signal multiplied by a rectangular signal centered on the current 2L samples.
- This introduces unwanted variations in the resulting coefficients. This phenomenon is called *spectral leakage*.

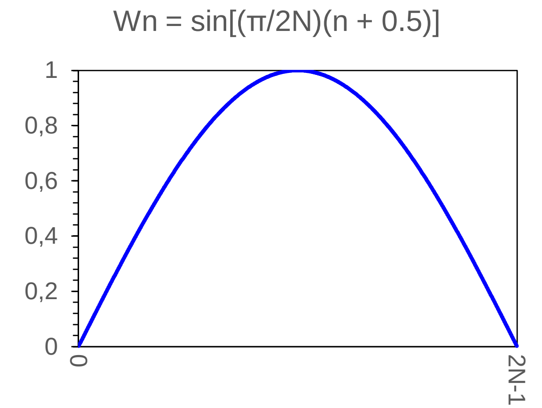
$$\mathrm{rect}(n) \underset{\mathcal{F}}{\Leftrightarrow} \mathrm{sinc}(f)$$



Minimum blocking artifacts

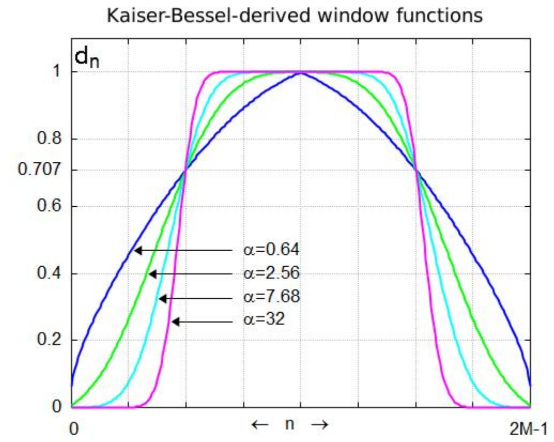
- AAC supports two different window functions to reduce spectral leakage:
- Sine window:

$$w_n = \sin\left[\frac{\pi}{2N}\left(n + \frac{1}{2}\right)\right], \qquad n \in [0,2N-1]$$



- Kaiser-Bessel-Derived (KBD) window:

$$w_n = \begin{cases} \sqrt{\dfrac{\sum_{p=0}^{n} W'(p,\alpha)}{\sum_{p=0}^{N} W'(p,\alpha)}}, & n \in [0,N) \\ \sqrt{\dfrac{\sum_{p=0}^{2N-n-1} W'(p,\alpha)}{\sum_{p=0}^{N} W'(p,\alpha)}}, & n \in [N,2N) \end{cases}$$

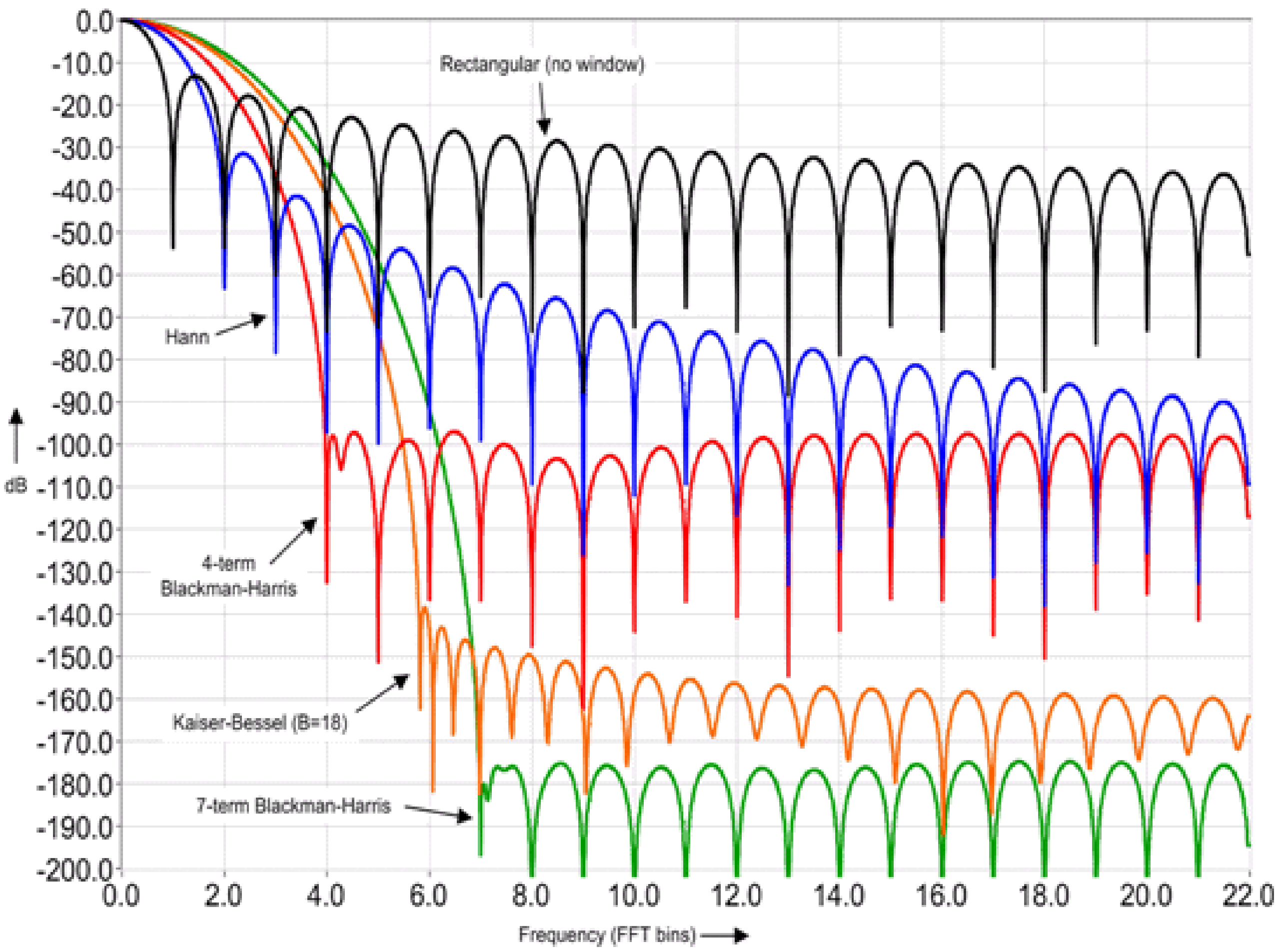


- Where W' is a Kaiser-Bessel window function.
- Alpha depends on the window length.

$$\alpha = \begin{cases} 4 \text{ for } N = 1024 \\ 6 \text{ for } N = 128 \end{cases}$$

Minimum blocking artifacts

- KBD functions still introduce unwanted frequencies but their amplitude is extremely reduced compared to those introduced by a rectangular window:

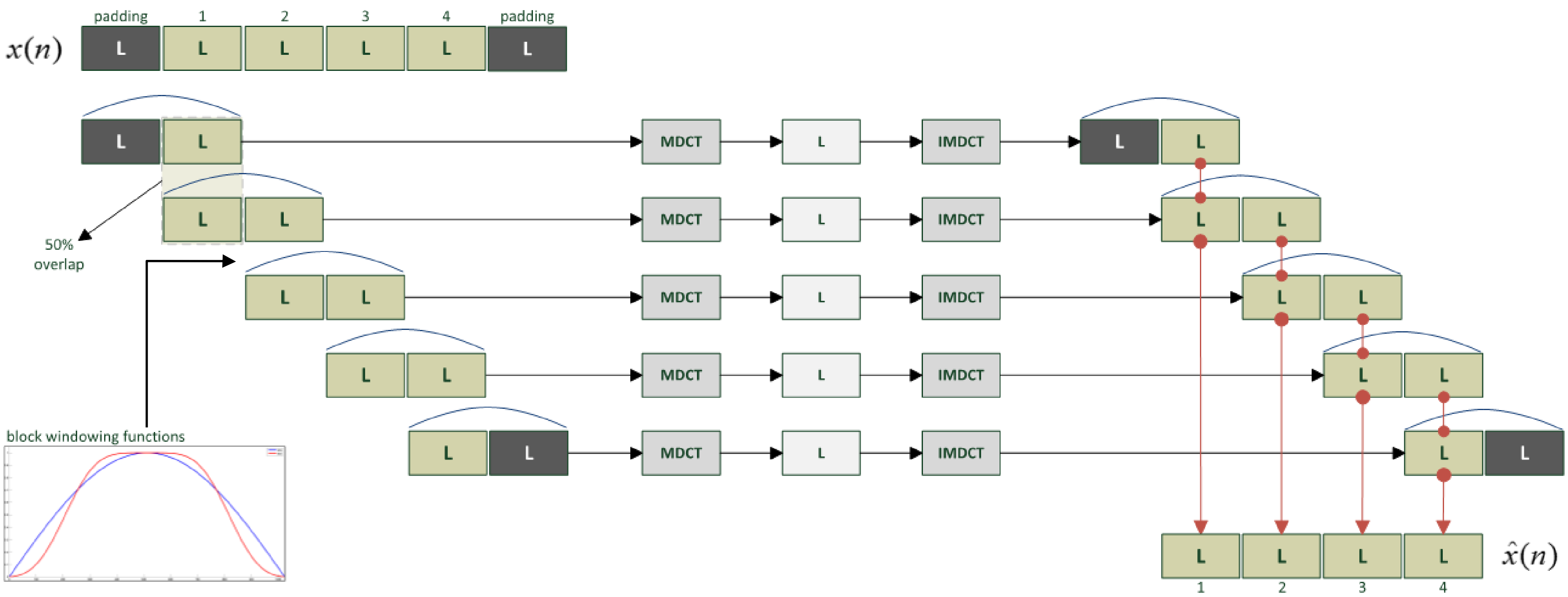


→ **Rectangular window:** -13 dB unwanted lobes

→ **KBD window:** lobes below -140 dB

Windowing

- When calculating the MDCT and the IMDCT the window function is applied on each frame (2L audio samples):



PERCEPTUAL MODEL

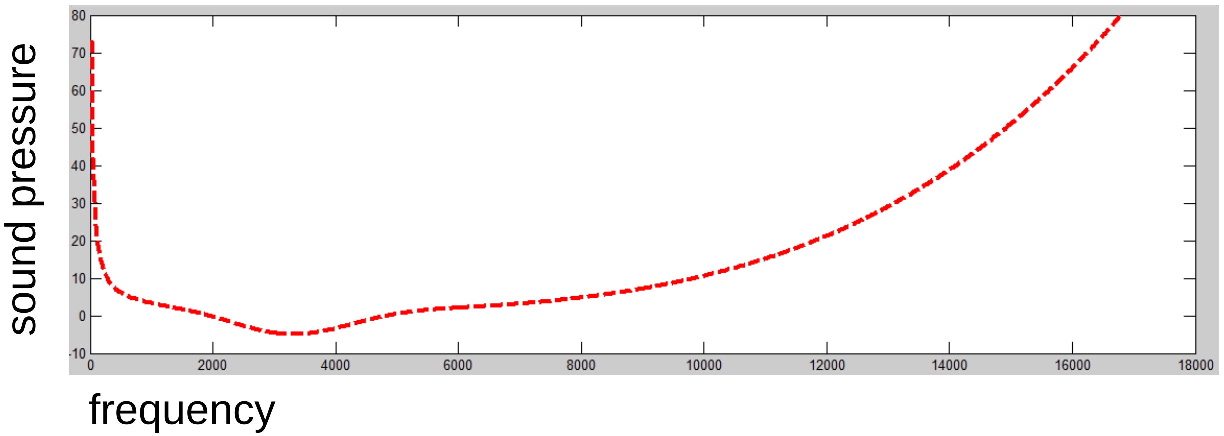
Perceptual model

- Exploited to (implicitly) compute the perceptual error.

- Embedding priori knowledge about human perception:
  - Absolute threshold
  - Critical bands
  - Tone-masking-noise
  - Noise-masking-tone
  - Spread of masking

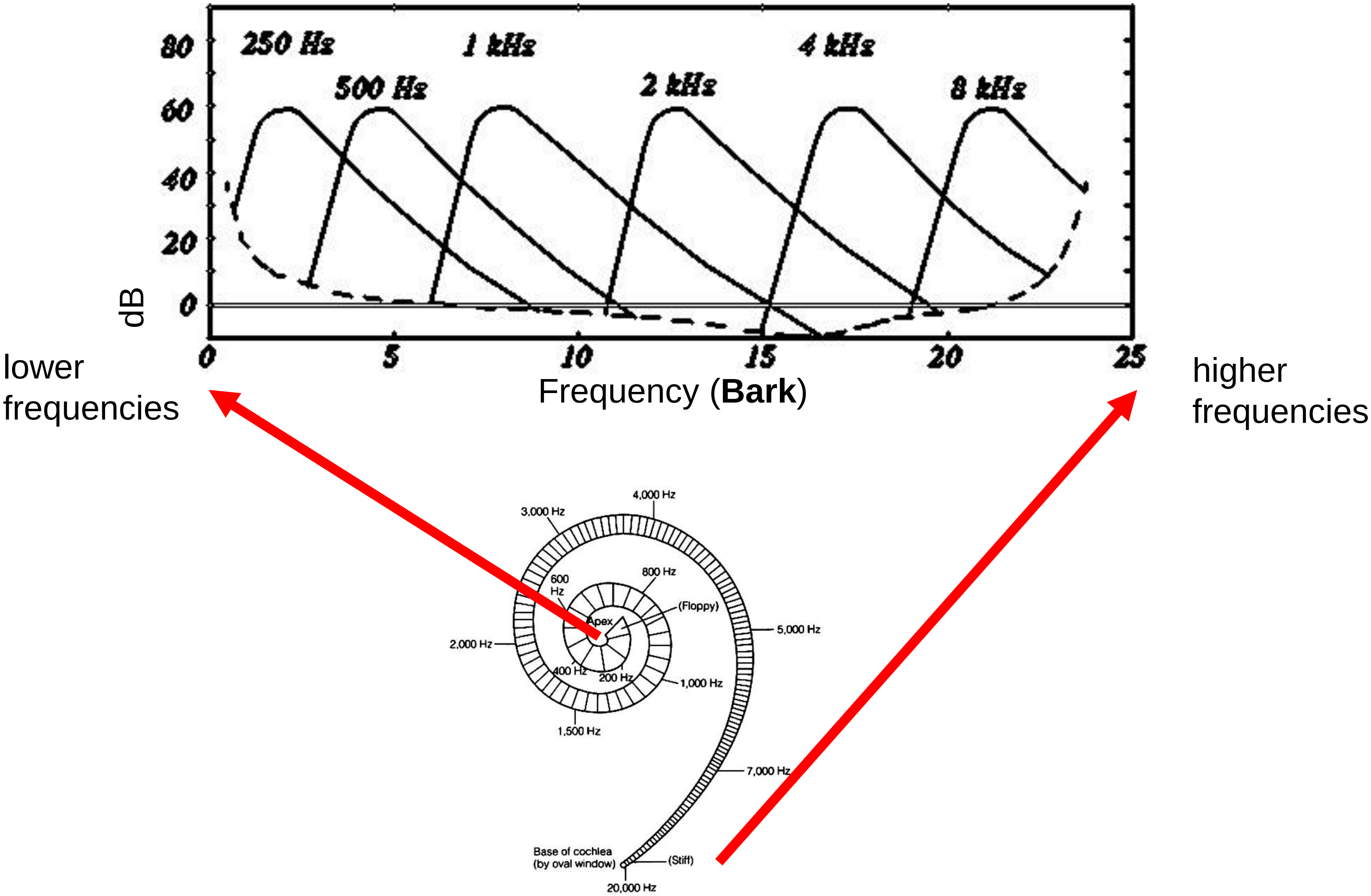
Absolute threshold of hearing

- Minimum sound level of an audible pure tone, below that level, the tone is not audible.
- Of all the equal-loudness contours, the absolute threshold of hearing is the lowest one.



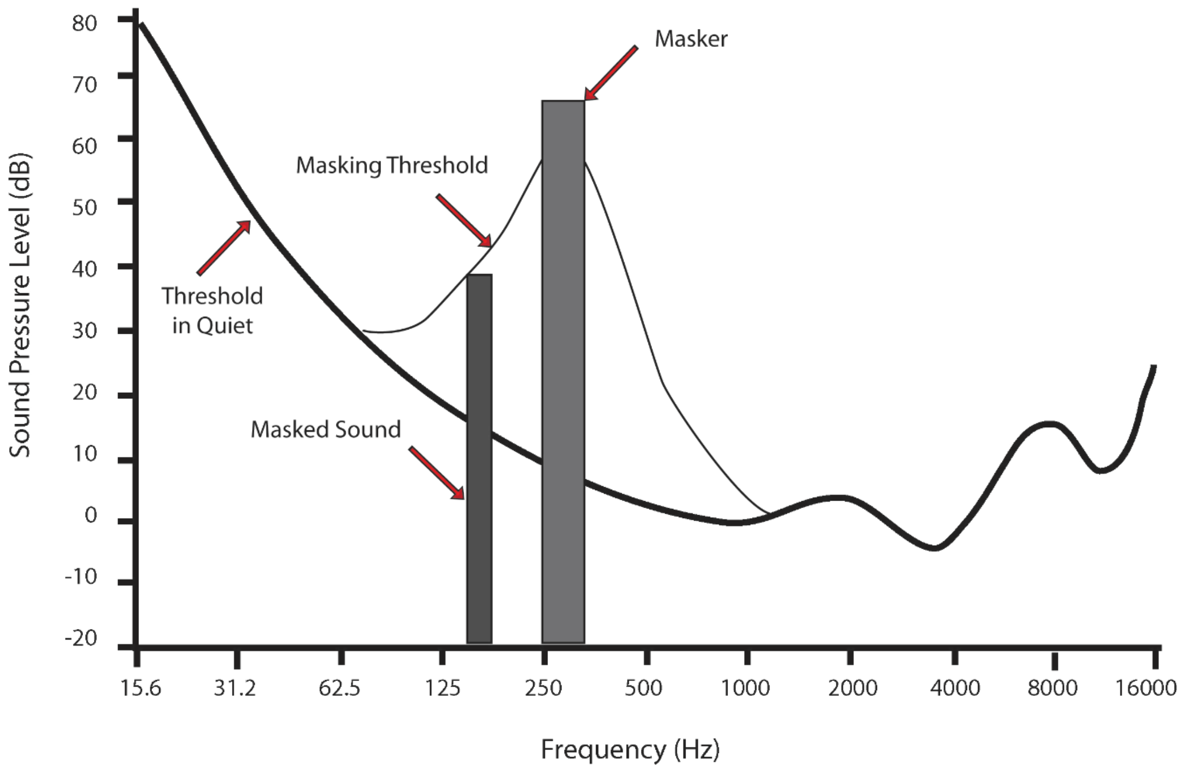
- The threshold depends on the subject, in particular on its age and its hearing conditions.
- It has been determined by statistical studies conducted on large numbers of patients.

Critical bands



Masking

- It is the reduction of the response of the human auditory system to a signal due to the presence of another (stronger) signal.

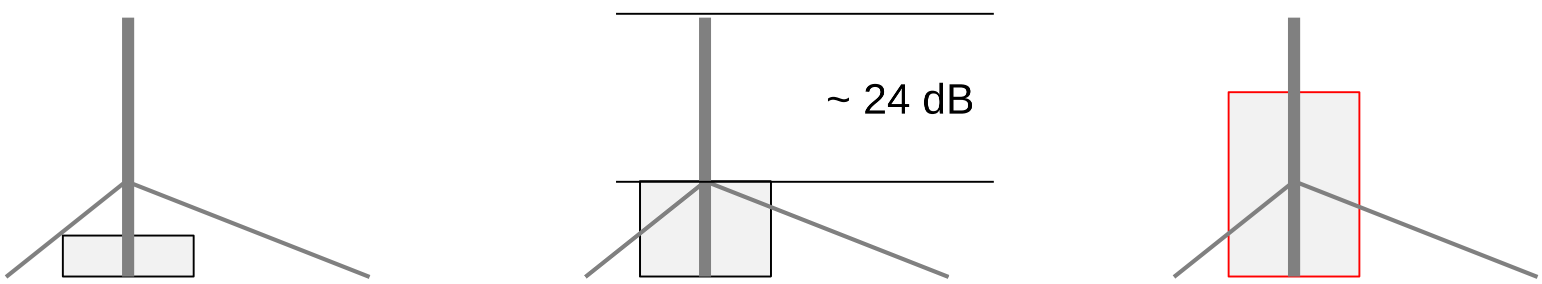


Tone-Masking-Noise (TMN)

- **Signal**: narrowband* noise ±400 Hz
- **Masker**: pure tone at 4 kHz
- (*) narrow means within the critical band of the tone

- Example, SMR (Signal to Mask Ratio) steps:

-40 -35 -30 -25 **-20** -15 -10 -5 0 -5 -10 -15 **-20** -25 -30 -35 -40

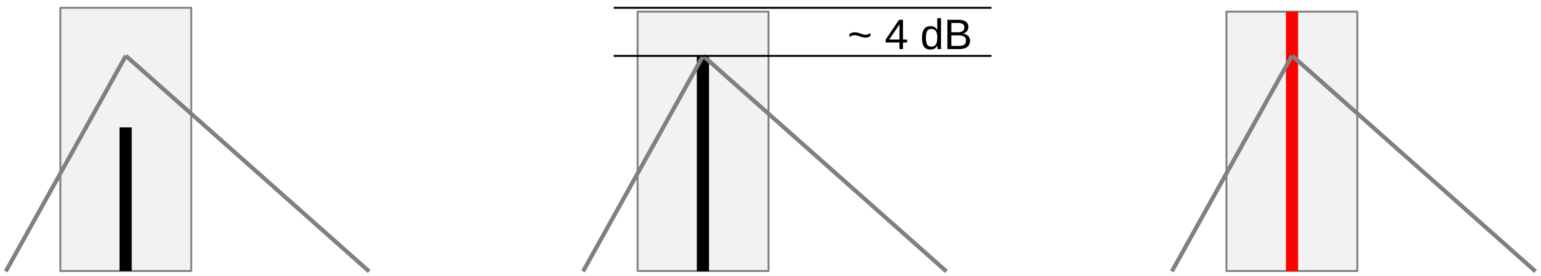


Noise-Masking-Tone (NMT)

- **Signal**: pure tone at 4 kHz
- **Masker**: narrowband* noise ±400 Hz
- (*) narrow means within the critical band of the tone

- Example, SMR steps:

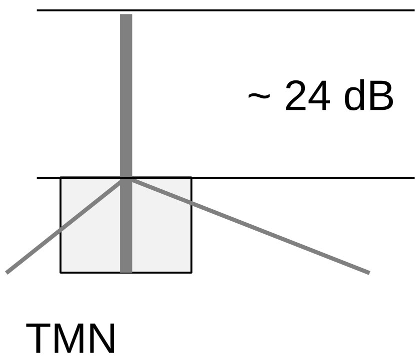
-15 -10 -5 -2.5 0 +2.5 **+5** 10 **+5** +2.5 0 -2.5 -5 -10 -15

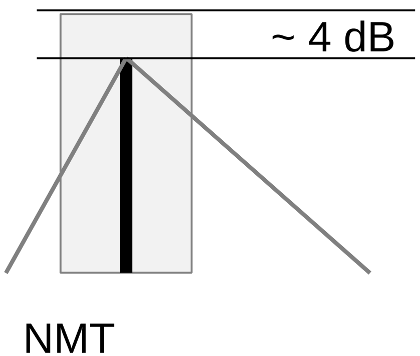


Spread of masking

- The masking effect is not band-limited since it also affects adjacent critical bands.

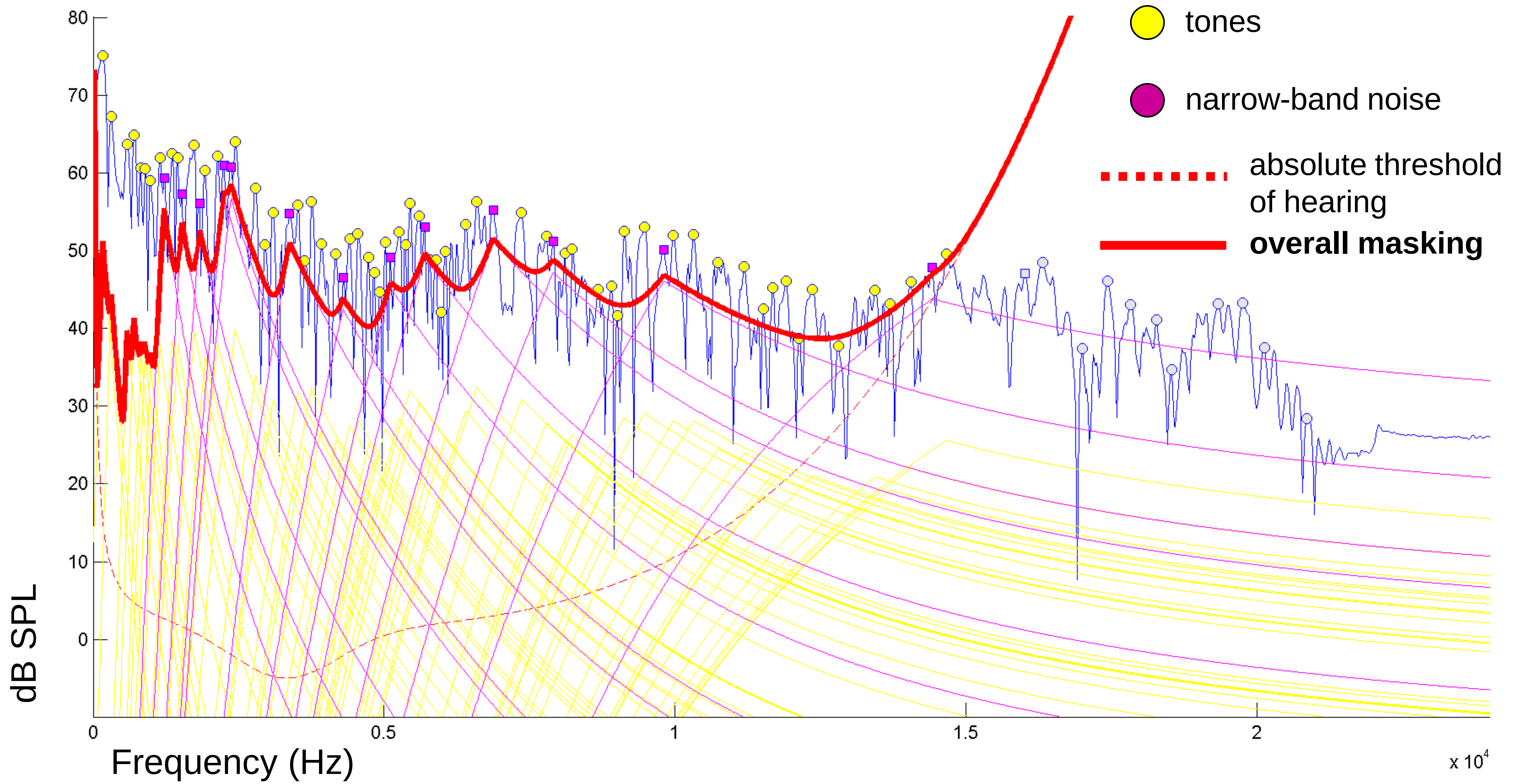
- The effects of a masking signal can be approximated by triangular function with different slopes on each side:
  - Left-slope: +25 dB/Bark
  - Right-slope: -10 dB/Bark





Overall masking function

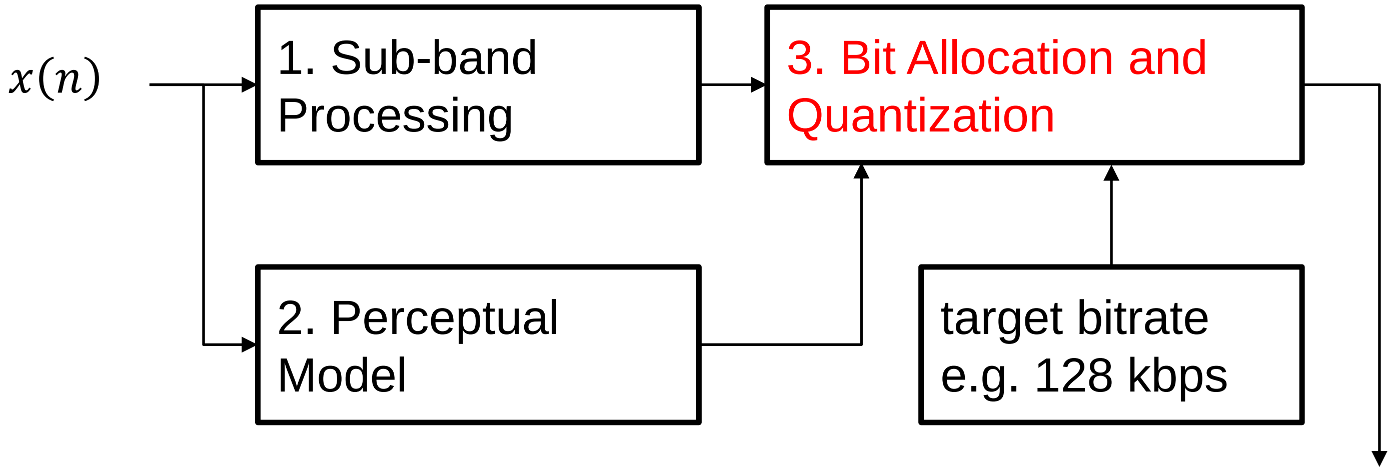
- Considering all the psychoacoustic effects, for a given signal it is possible to compute the overall masking curve.
- Coefficients below the curve will not be heard.



BIT ALLOCATION AND QUANTIZATION

Bit allocation and quantization

- For every frame:
  - Frequency coefficients are extracted.
  - Masking function is computed.
  - Bit allocation and quantization are performed.

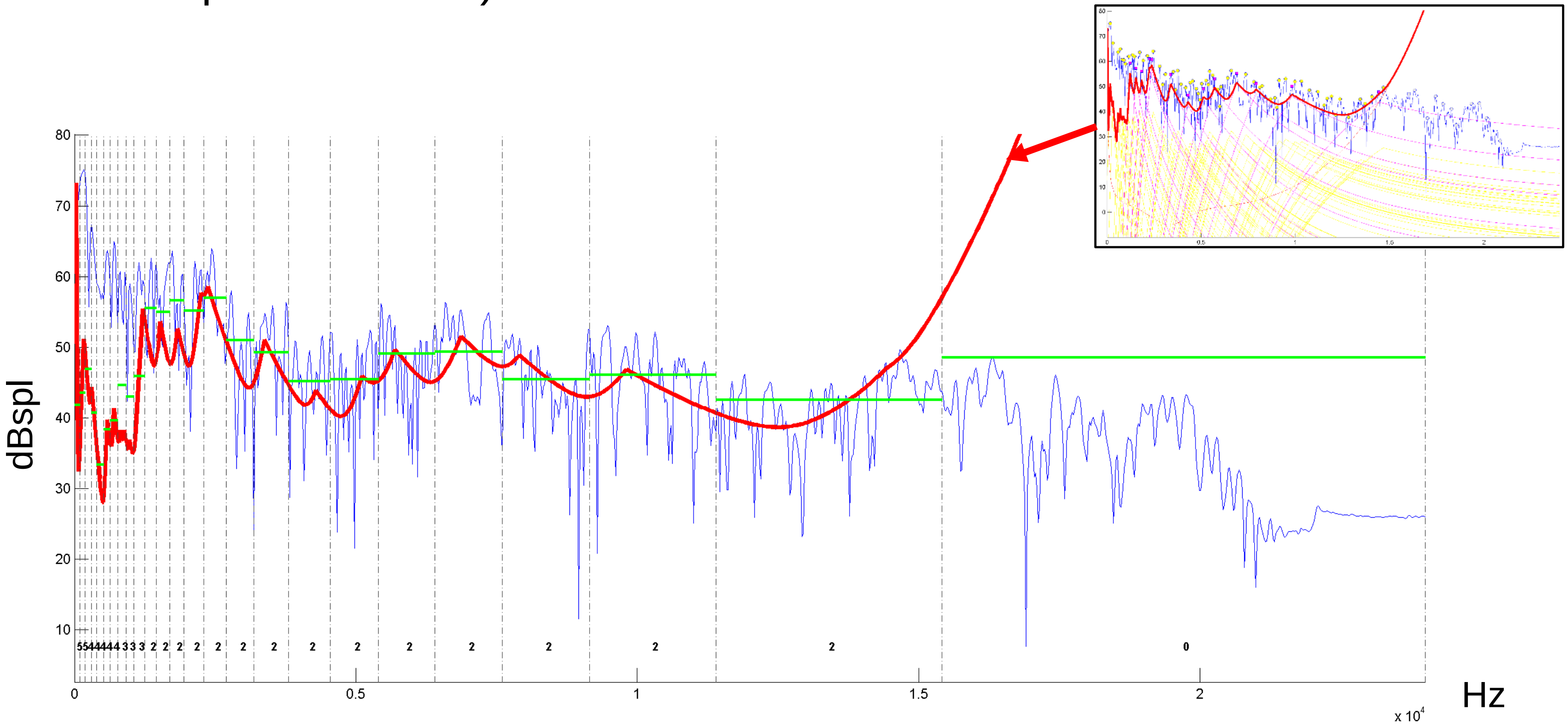


$$x_q(n) = x(n) + e(n)$$

$$\arg\min_{e(n)} PE\big(x(n), e(n)\big)$$

Bit allocation and quantization

- The power spectrum and the calculated masking threshold are used to allocate bit to each band.
- MDCT coefficients are quantized according to the band to which each coefficient belongs (using the number of bits assigned to the correspondent band)



Bit allocation algorithm

- Usually composed of two nested loops.
- Progressively allocates bits to each band trying to minimize the total perceptual error.
- This nested loop combination is also called *rate/distortion loop*. In this case the word *rate* refers to the number of bits used to quantize the coefficients and the word *distortion* refers to the noise introduced by the quantization.

- Within each (critical) band some values are calculated:
  - SNR: Signal to Noise Ratio
  - SMR: Signal to Mask Ratio
  - NMR: Noise to Mask Ratio
- These values depend on:
  - maxSPL: the maximum power spectrum value within the band.
  - minMT: the minimum masking threshold value within the band.

Bit allocation algorithm - initialization

- the total number of available bits is computed as:

$$available\ bits = \frac{Br\ \cdot N}{Sr}$$

  Where *Br* is the desired bitrate, *N* is the number of frequency coefficients and *Sr* is the sampling rate.

- A vector containing the number of bits assigned to each band is created and all its values are set to 0, this vector is called *bits*:

  bits(band) ← *0*

Bit allocation algorithm - loop

- The rate/distortion loop, for each iteration, calculates the perceptual error for each band and consequently assigns more bits to the band that has the highest error.
- This loop stops when all the available bits have been assigned to a band.
- Elements used in the algorithm:
  - bits: vector of assigned bits (one value per band).
  - NMR: vector of perceptual errors (one value per band) (Noise to Mask Ratio).
  - PE(b): function that evaluates the perceptual error for a specific band.
  - num_coeff(b): returns the number of spectral coefficients for a specific band.
- Pseudo-code:

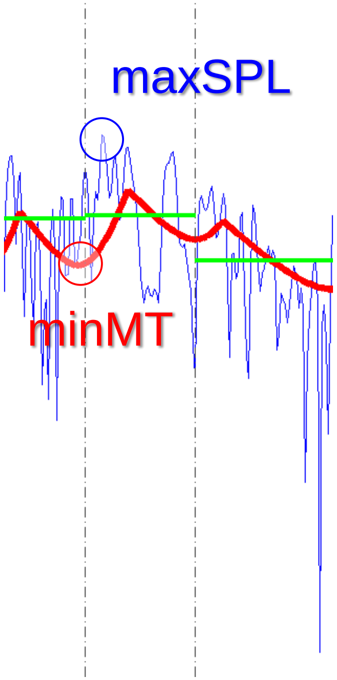
```
while (available_bits > 0):

    for band in bands:
        NMR(band) <- PE(band)

    worst_b <- max(NMR)
    bits(worst_b) += new_bits
    available_bits -= (new_bits * num_coeff(worst_b))
```

Computing the perceptual error

- Given a band having width of 1 Bark:

- maxSPL ← max power spectrum value (within the band).
- minMT ← min masking threshold value (within the band).

- **SMR ← maxSPL – minMT**
- SMR = Signal to Mask Ratio.
  If the SMR is high, it means that the coefficients within the band are far from the masking function.



Computing the perceptual error

- Given a band having width of 1 Bark:

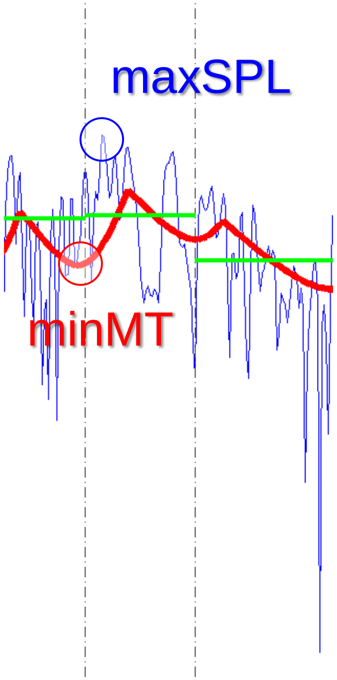
- maxSPL ← max power spectrum value (within the band).
- minMT ← min masking threshold value (within the band).

- SMR ← maxSPL – minMT
- **SNR (signal-to-noise ratio) ← SNR reference value for bit(band)**.
- The SNR in dB for a digital signal is evaluated as:

$$\mathrm{SNR_{dB}} \cong 6.02 \cdot n + 1.761$$

where n is the number of bits used.

- The Signal to Noise Ratio is determined by the number of bits used to quantize the coefficients, a strong quantization introduces more noise, hence the lower SNR value.
- SNR values for number of bits used:



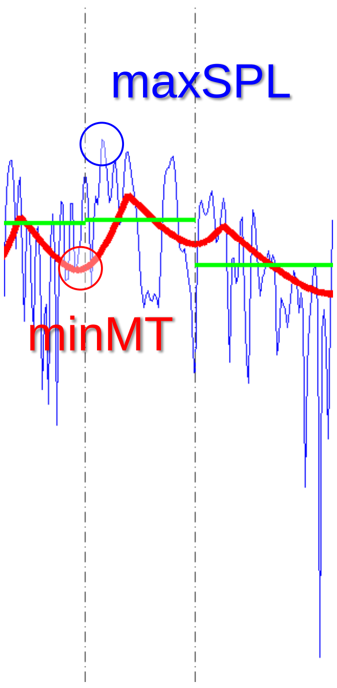
# bits	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\mathrm{SNR_{dB}}$	0.00	7.78	13.80	19.82	25.84	31.86	37.88	43.90	49.92	55.94	61.96	67.98	74.00	80.02	86.04	92.06

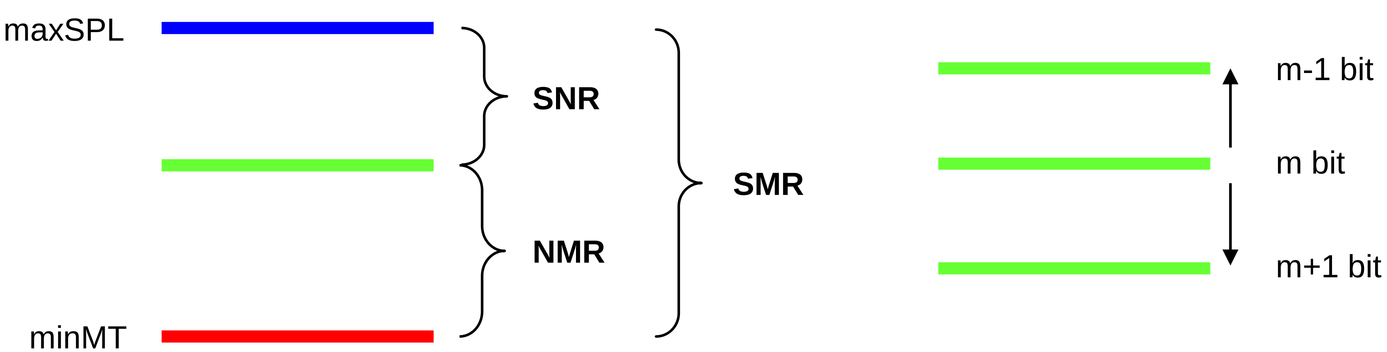
Computing the perceptual error

- Given a band having width of 1 Bark:

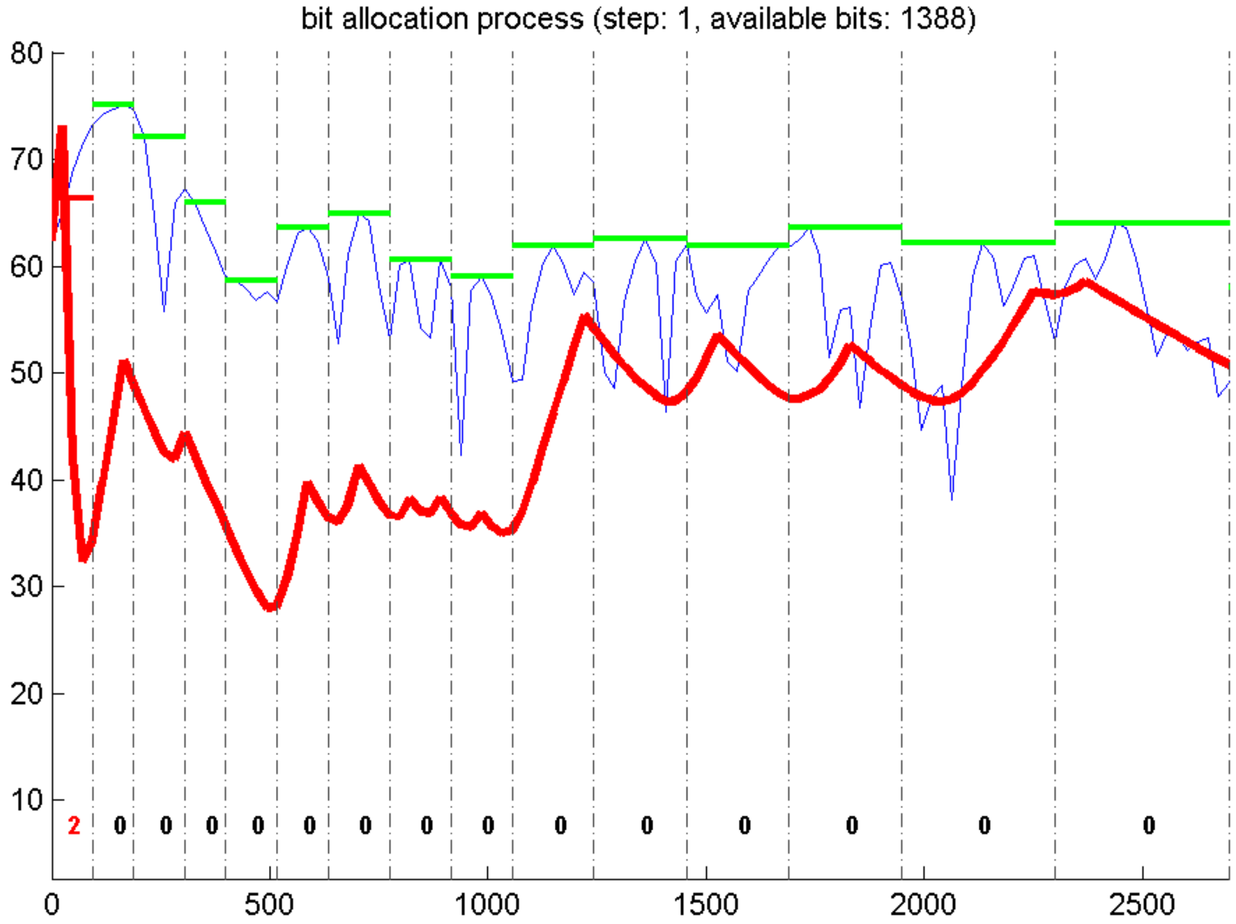
- maxSPL ← max power spectrum value (within the band).
- minMT ← min masking threshold value (within the band).

- SMR ← maxSPL – minMT
- SNR ← SNR reference value for bit(band).
- **NMR(band) ← SMR – SNR**
- The Noise to Mask Ratio is the difference between the SMR and the SNR.
- Increasing the number of bits increases the SNR and lowers the NMR.

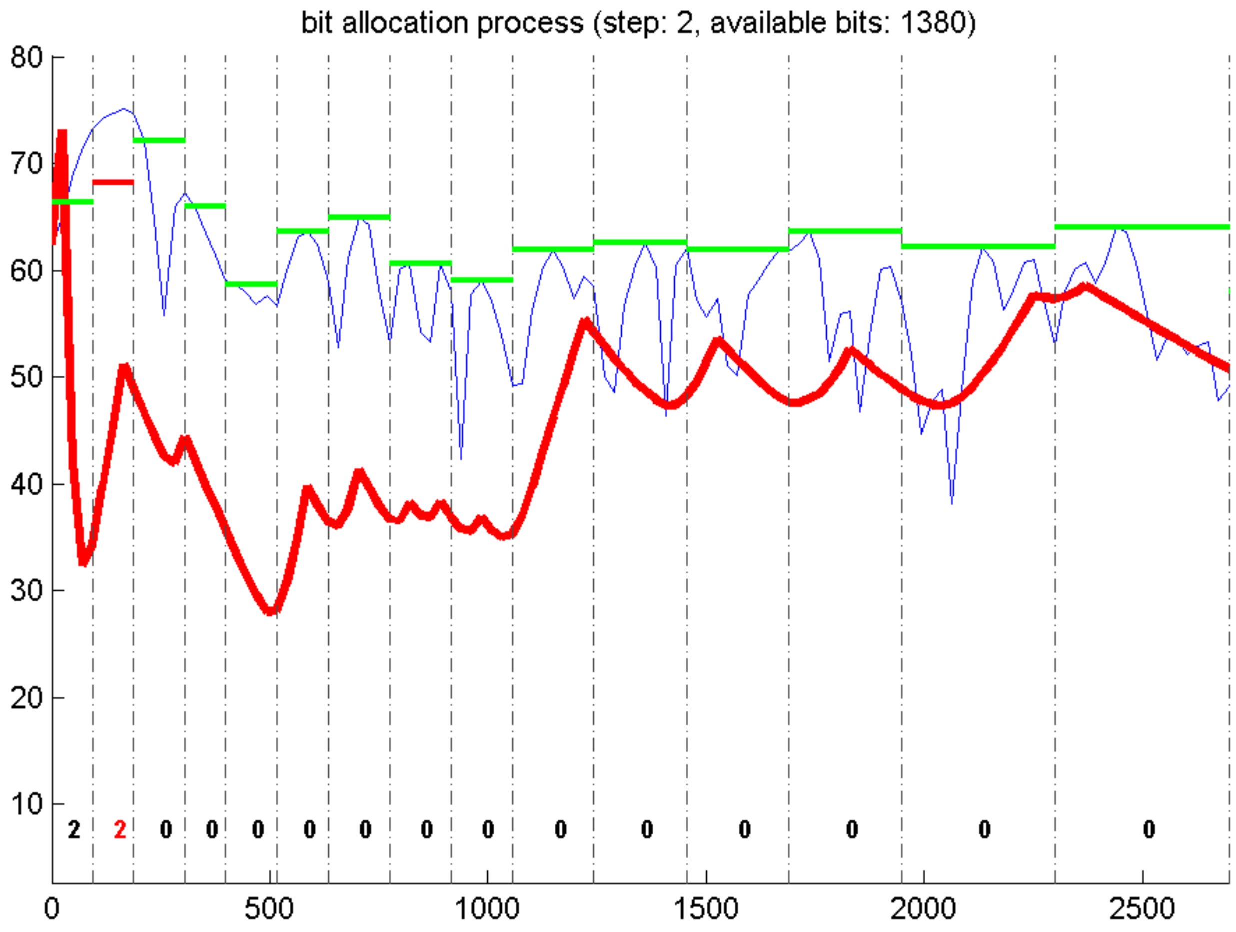




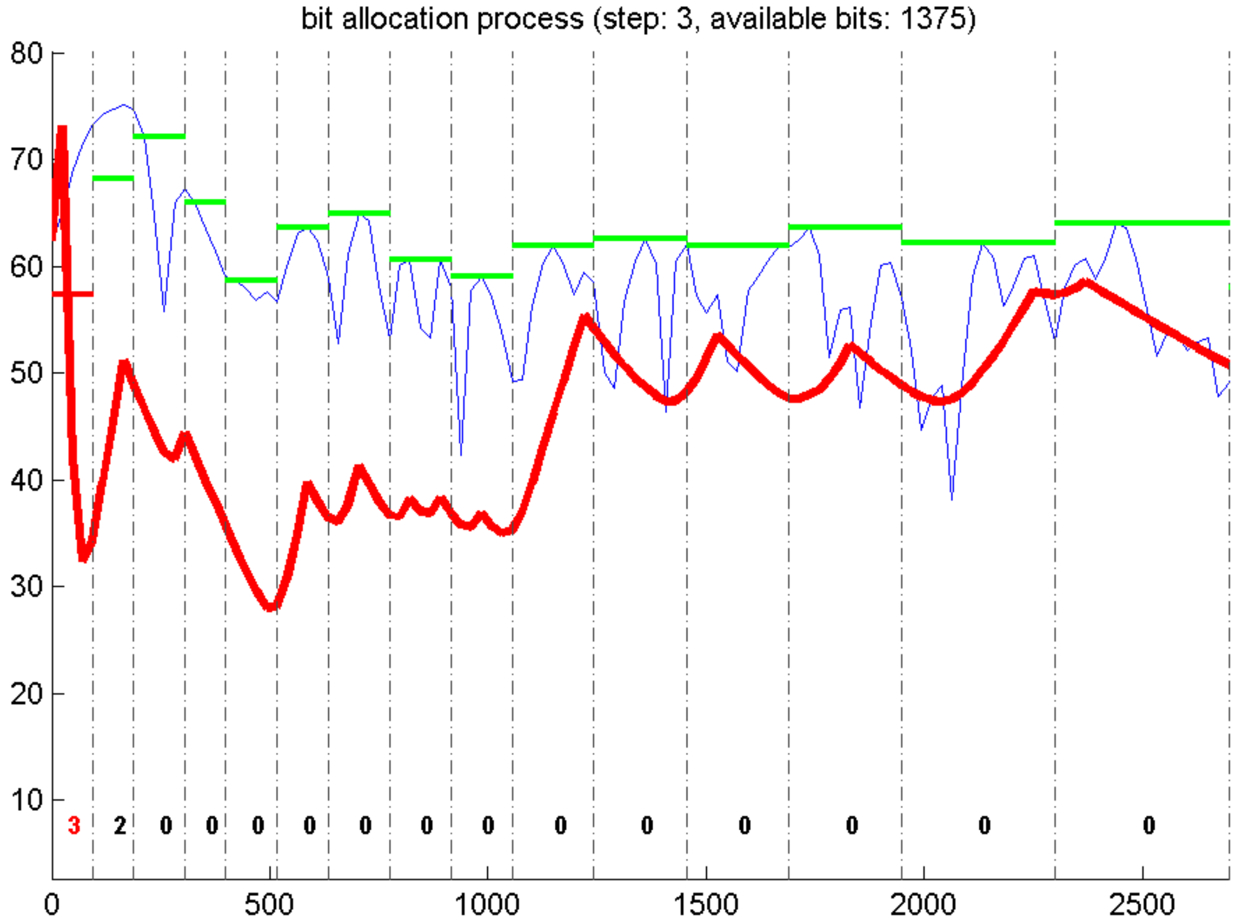
Bit allocation example



Bit allocation example



Bit allocation example



Bit allocation example