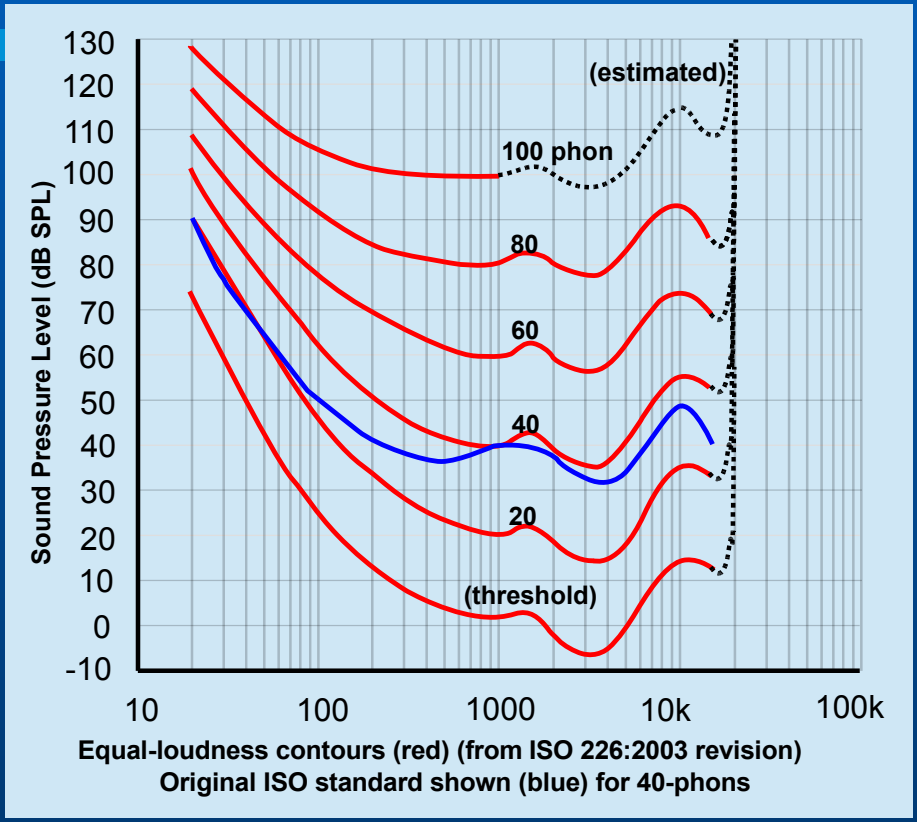
# MPEG audio compression and description

**Georgios Tziritas**
**Computer Science Department**
**http://www.csd.uoc.gr/~tziritas**

# Loudness relations

Equal loudness curves 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 range of human hearing is about 20 Hz to about 20 kHz

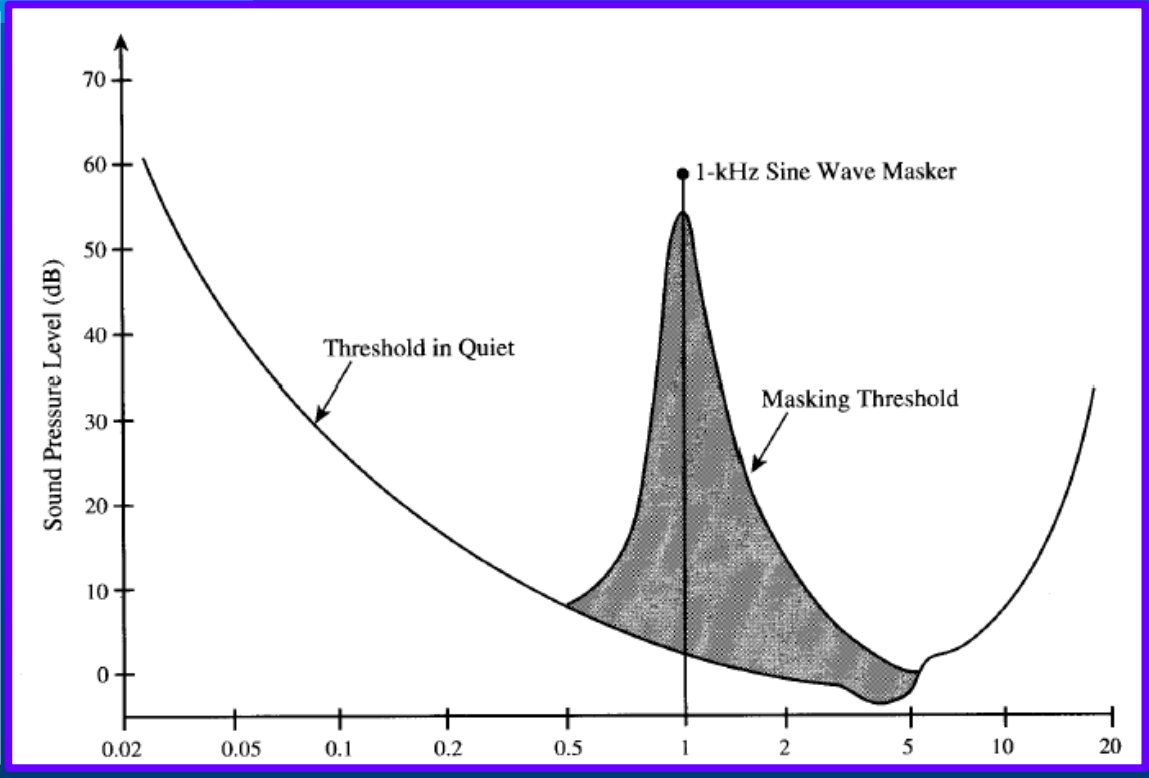
# Sound masking

A lower tone can effectively mask (make us unable to hear) a higher tone

A higher tone does not mask a lower tone well

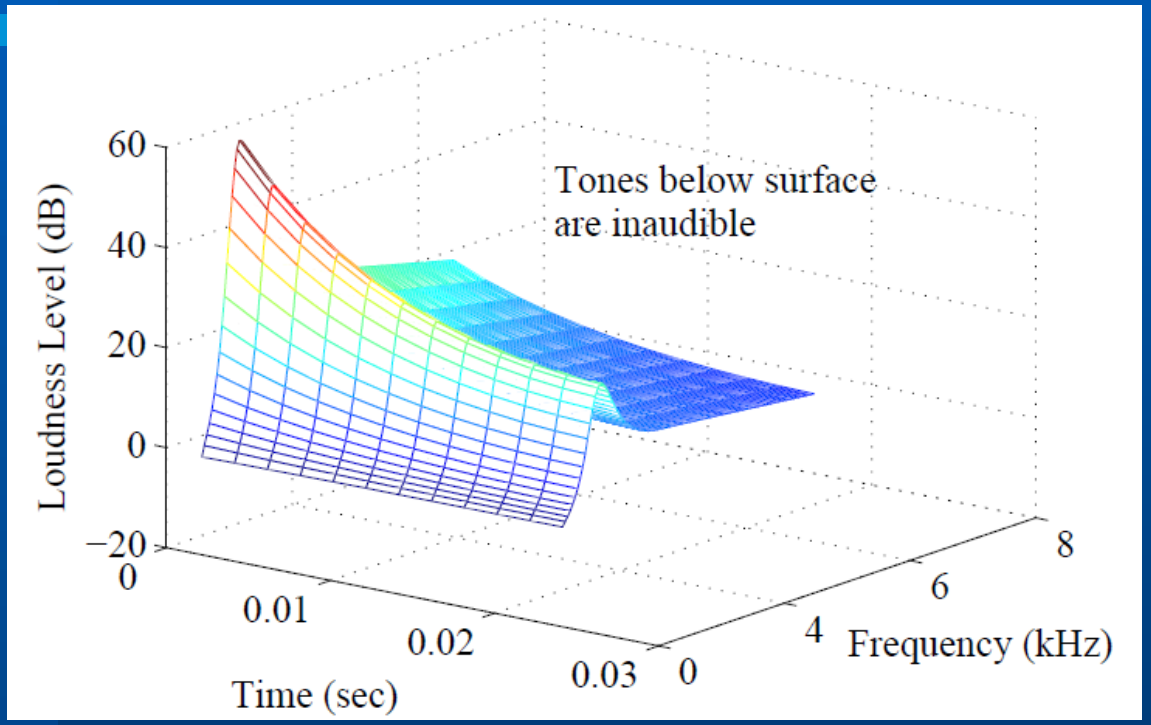
The greater the power in the masking tone, the wider is its influence – the broader the range of frequencies it can mask.

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




Frequency masking is studied by playing a particular pure tone at a loud volume, and determining how this tone affects our ability to hear tones nearby in frequency

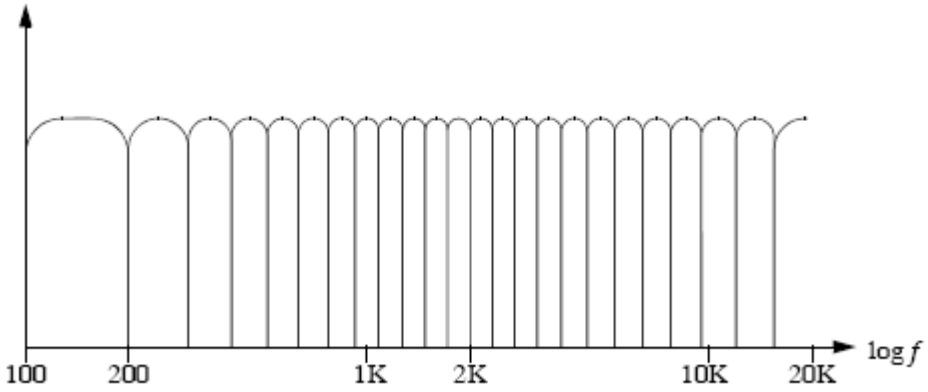
# Temporal masking




Effect of temporal masking depends on both time and closeness in frequency

# 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 <500Hz: remains approximately constant in width (about 100 Hz)
  - For masking frequencies >500Hz: increases approximately linearly with frequency




$$b = 13\arctan(0.76f) + 3.5\arctan\left(f^2/56.25\right)$$

$$\Delta f = 25 + 75\left(1 + 1.4f^2\right)^{0.69}$$

***f* kHz, *Δf* Hz**

# Critical bands

| Band No. | Center Freq. (Hz) | Bandwidth (Hz) | Band No. | Center Freq. (Hz) | Bandwidth (Hz) | Band No. | Center Freq. (Hz) | Bandwidth (Hz) |
|---|---|---|---|---|---|---|---|---|
| 1 | 50 | -100 | 10 | 1175 | 1080-1270 | 19 | 4800 | 4400-5300 |
| 2 | 150 | 100-200 | 11 | 1370 | 1270-1480 | 20 | 5800 | 5300-6400 |
| 3 | 250 | 200-300 | 12 | 1600 | 1480-1720 | 21 | 7000 | 6400-7700 |
| 4 | 350 | 300-400 | 13 | 1850 | 1720-2000 | 22 | 8500 | 7700-9500 |
| 5 | 450 | 400-510 | 14 | 2150 | 2000-2320 | 23 | 10,500 | 9500-12000 |
| 6 | 570 | 510-630 | 15 | 2500 | 2320-2700 | 24 | 13,500 | 12000-15500 |
| 7 | 700 | 630-770 | 16 | 2900 | 2700-3150 | 25 | 19,500 | 15500- |
| 8 | 840 | 770-920 | 17 | 3400 | 3150-3700 | | | |
| 9 | 1000 | 920-1080 | 18 | 4000 | 3700-4400 | | | |

# Key technologies

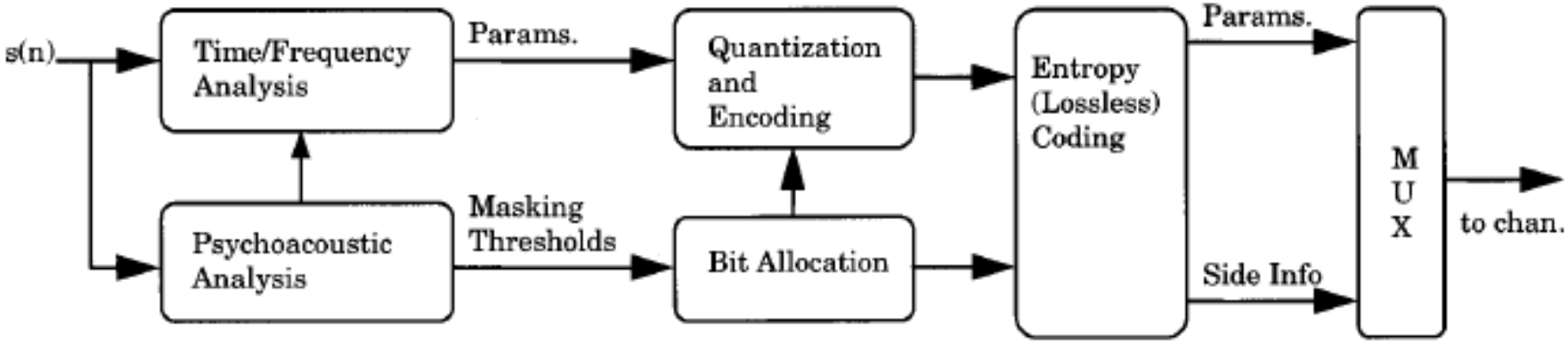
**Acoustic masking**

**Computation as a function of frequency of signal / masking level**

**Quantization step and bit allocation**

**Transparent quality = inaudible error**

**Transform / subband coding**



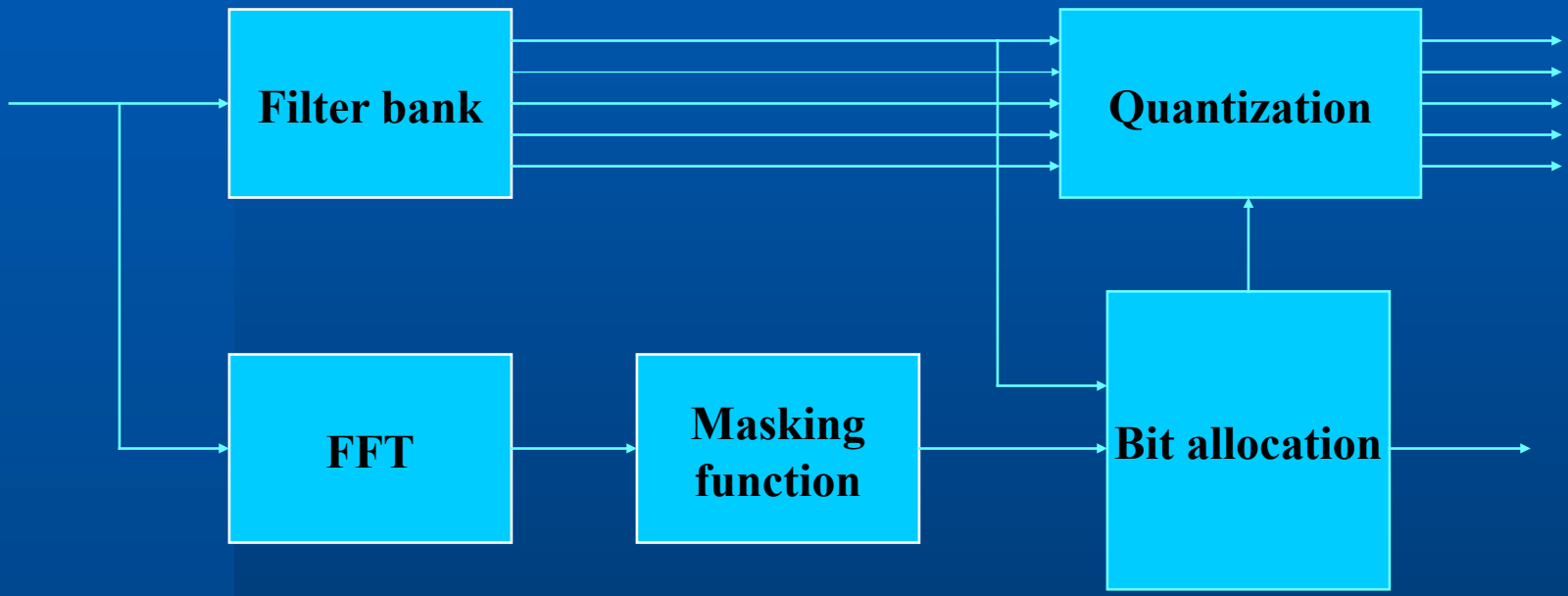
# MPEG-1 audio compression

**CD-Audio : 2 x 44100 samples/sec x 16 bits/sample = 1.41 Mbits/sec**

| | Transparent quality | | Compression ratio |
|---|---|---|---|
| Layer I | 384 kbits/sec | DCC | 4 |
| Layer II | 192 kbits/sec | DAB, CD-I, DVD | 8 |
| Layer III | 128 kbits/sec | ISDN, Internet, Satelite | 12 |

**Sampling rate : 32 kHz, 44.1 kHz, 48 kHz**

# MPEG-1 compression (Layers I and II)




**32 filters : 750 Hz για 48 kHz**
**Low-pass filter and modulation using DCT**
**Quantizer controlled by dynamic bit allocation**
**Signal frames (8 ms / 24 ms for 48 kHz)**

# MPEG-1 compression (Layer II)

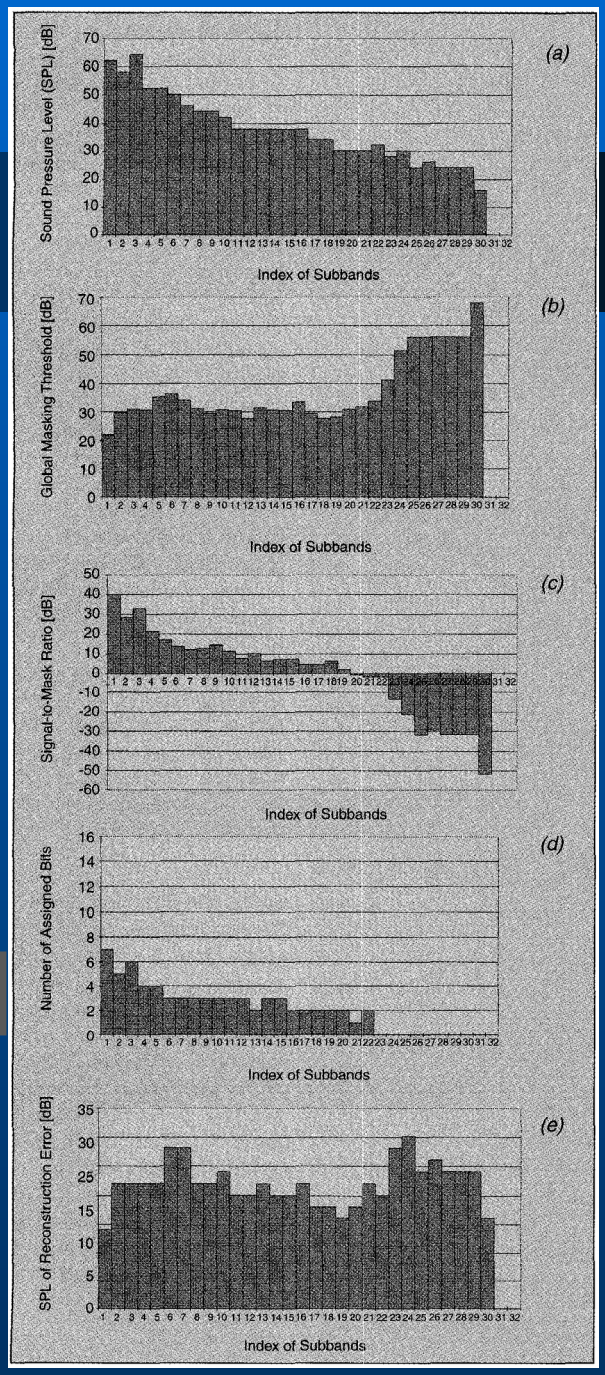
**Pressure level**

**Masking threshold**

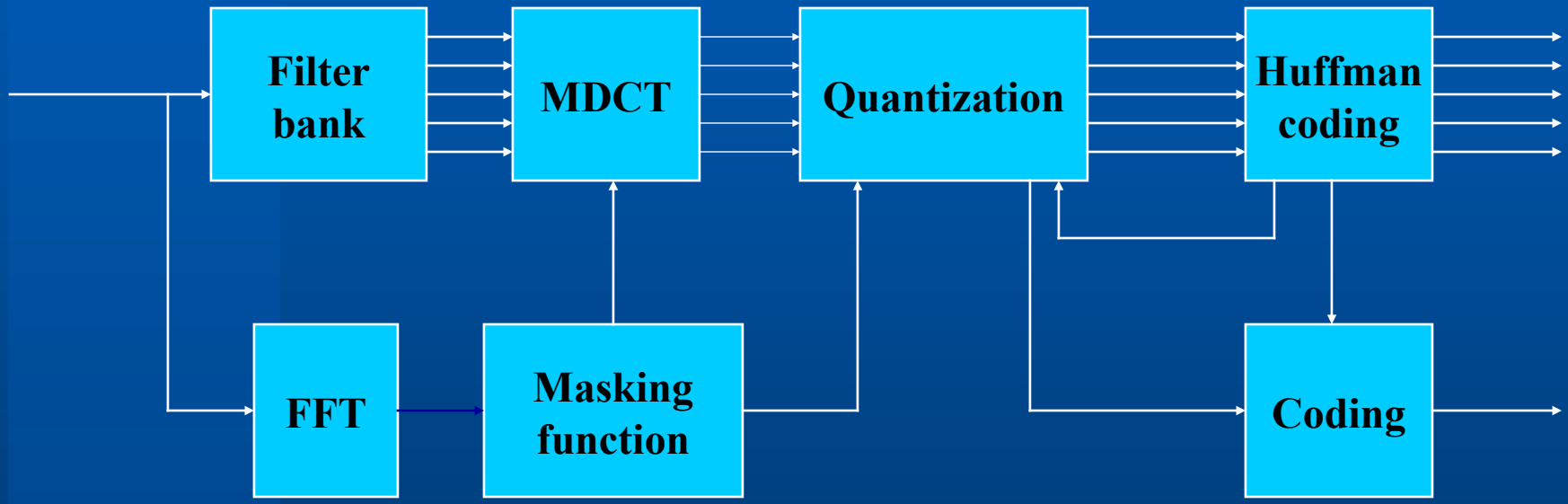
**Signal/mask**

**Bits allocated**

**Level reconstruction error**



# MPEG-1 compression (Layer III) MP3

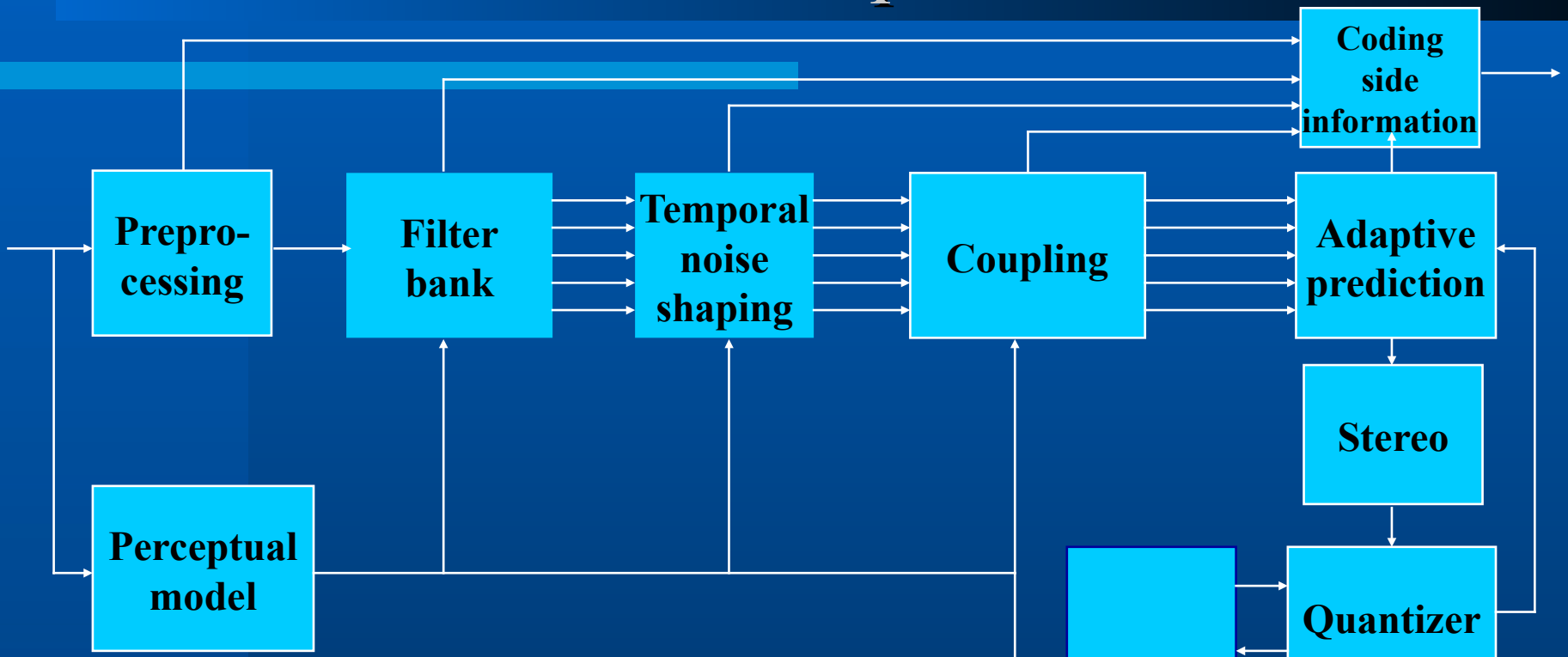



- Modified DCT (50% block overlap, 6/18 coefficients)
  - improved frequency analysis
- Non-uniform quantization
- 32 Huffman code tables and “zero” intervals

# MPEG-1 compression (Layer III) MP3

| 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 compression




Modified DCT (50% block overlap)

Non-uniform quantization

Huffman coding

# MPEG-2 AAC compression

**Sampling rate**
**8 kHz – 96 kHz**

**Multi-channel audio signals**

**Limitation of the reconstruction error using an adaptive distribution of quantiztion errors by predictions in the frequency domain**

# MPEG-4 compression

- **Text-to-Speech**
  - **200-1200 bits/sec**
    - **international phoneme alphabet**
- **Structure Audio Orchestra Language**
  - **synthetic music and sound effects**
- **Harmonic Vector Excitation Coding (speech LPC)**
  - **2-4 kbits/sec at 8 kHz**
  - **4-16 kbits/sec at 8 kHz or 16 kHz**
- **Speech coding (CELP)**
  - **6-24 kbits/sec at 8 kHz or 16 kHz**
- **Audio coding (MPEG-4 AAC)**

# MPEG-4 AAC

Based on MPEG-2 AAC

Target bit rate : 24 kbits/sec/channel

Scalable encoding

24 kbits/s (mono), 40 kbits/s (stereo), 56 kbits/s (stereo)

Vector quantization for better compression ratio

Long delay prediction (for harmonic signals)

'Noise-like' signals are not transmitted

Transmission error resilience

# Speech signals

| | Frequency bandwith Hz | Sampling rate kHz | bits/sample | kbits/sec |
|---|---|---|---|---|
| Telephone | 200-3400 | 8 | 8 | 64 |
| Wideband telephone | 50-7000 | 16 | 8 | 128 |

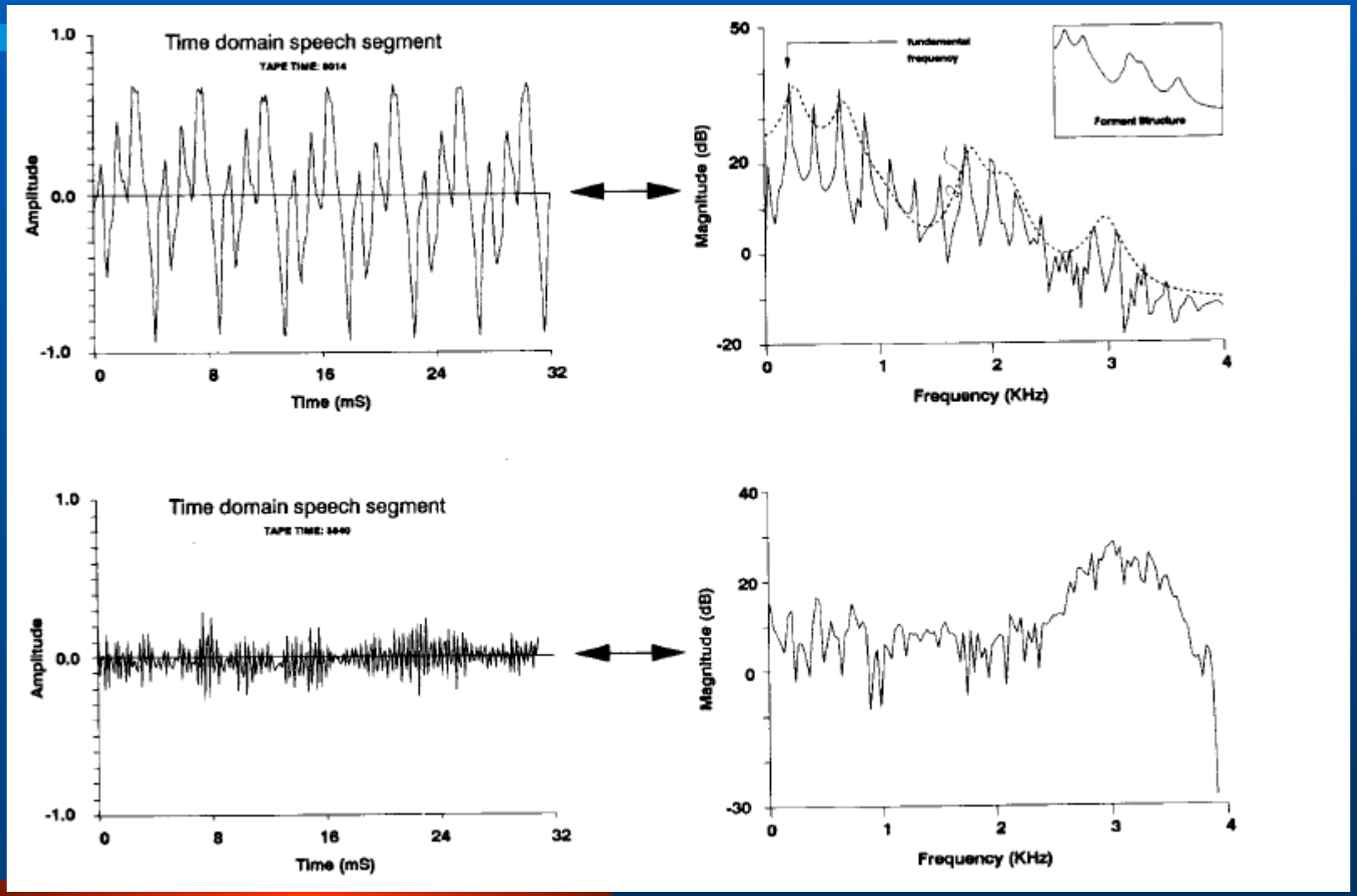
- High bit rate : 32 kbits/sec
- Medium bit rate : 8 kbits/sec
- Low bit rate : 4 kbits/sec
- Very low bit rate : 2 kbits/sec

# Speech signals

**Non-stationary signal,**
**stationarity for short time inervals, 5-20 msec**

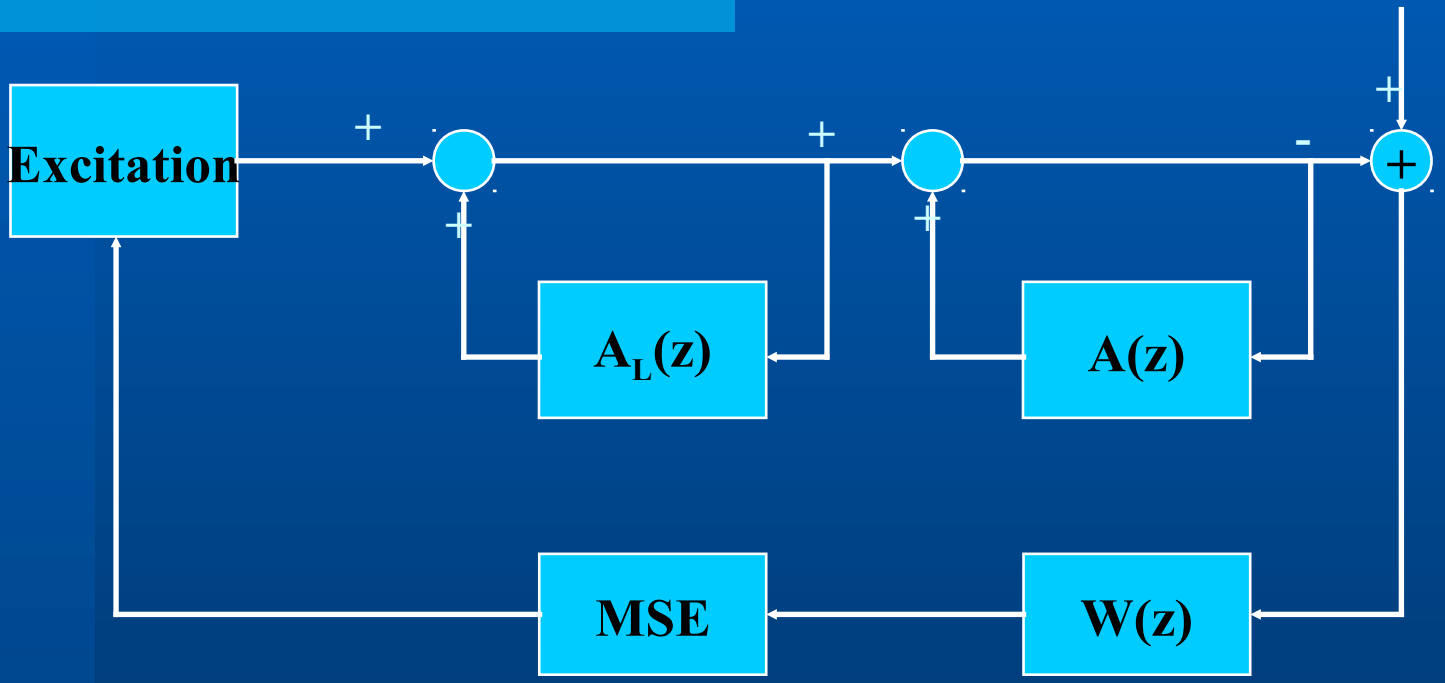
- **Voiced : periodic signals**
  **modulated pure frequencies (harmonics)**
- **Unvoiced : wideband, 'noise-like'**

# Speech signals



# Analysis synthesis

## Closed-loop optimization for excitation signal




Short delay prediction : speech modulation
Long delay prediction : fundamental frequency
Perceptual error weighting

Coded signals excitation (5 ms)
Dictionary : 1024 vector of 40 components

# MPEG-7 audio description

**Segmentation**

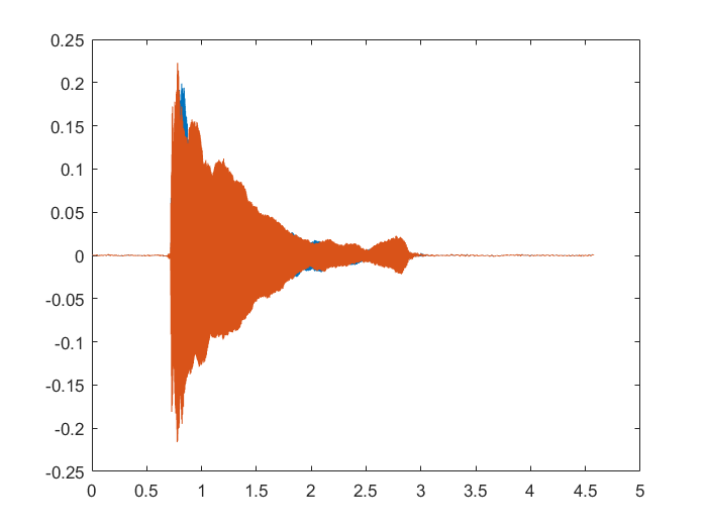
**Scaling**

**Low level descriptors (Signal / spectrum / timbre)**

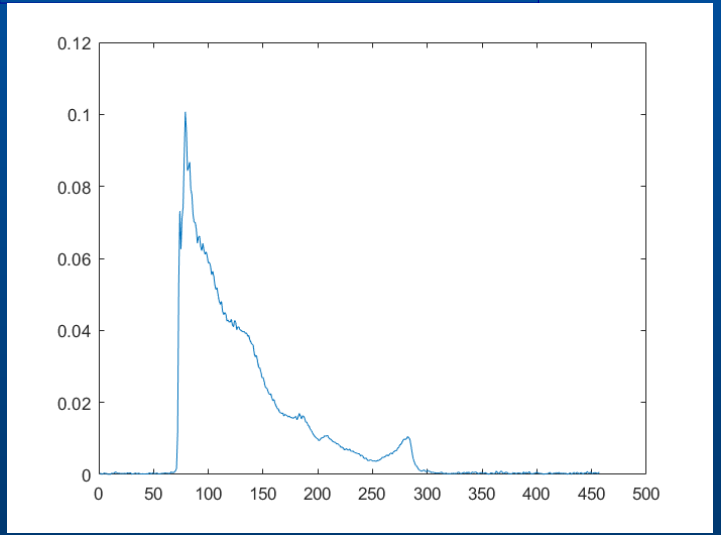
**High level descriptors (Schemes / tools)**

# MPEG-7 basic audio descriptors

**Waveform Envelope**




**Power**
**Temporally-smoothed**



# MPEG-7 basic spectral descriptors

**Spectrum envelope**
- **Logarithmic scale, spectrogram**

**Spectrum centroid**

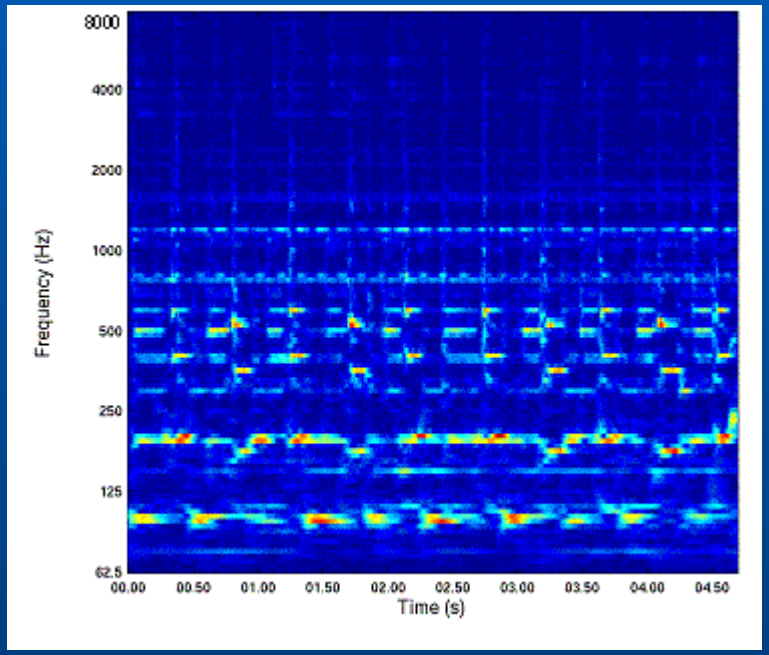
**Spectrum spread**

**Spectrum flatness**

**Fundamental frequency**

**Harmonicity**
- **musical tones, vowels**
- **percussive sound**
- **noise, consonants**

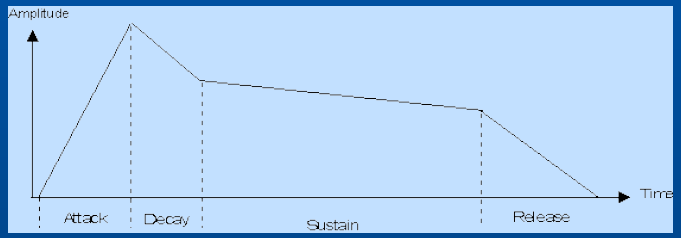
**Silence**



# MPEG-7 timbral audio descriptors

**Log attack time**

**Temporal centroid**




**Spectral centroid (linear frequency scale)**

**Harmonic spectral centroid**

**Harmonic spectral deviation**

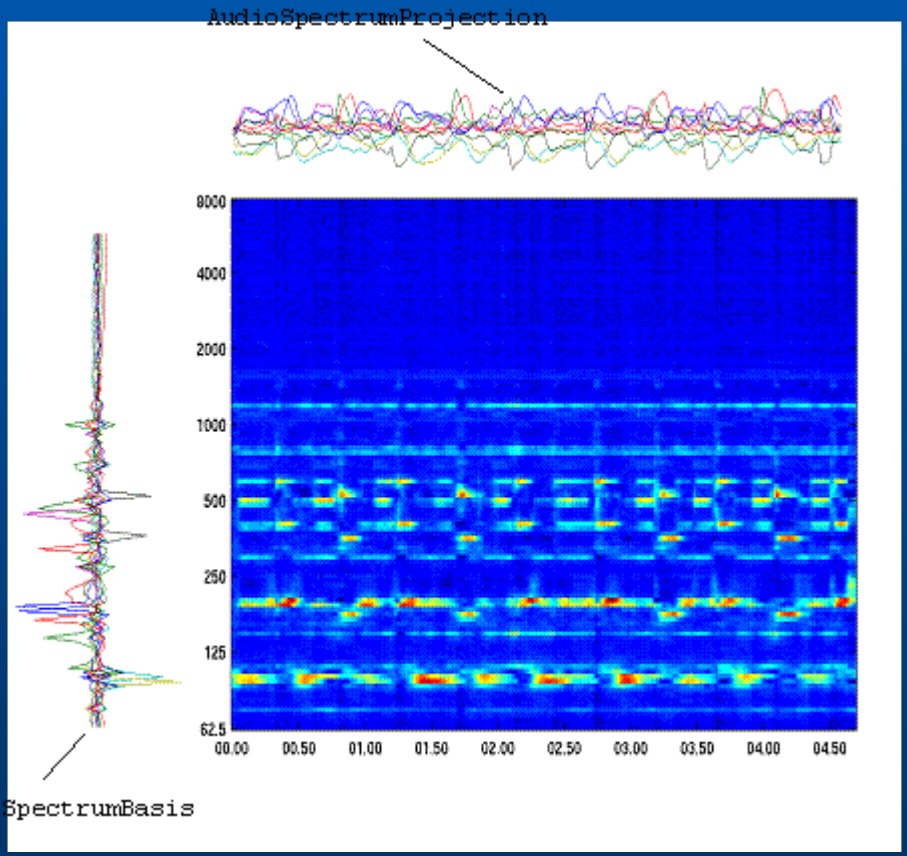
**Harmonic spectral spread**
**(normalized standard deviation)**

**Harmonic spectral variation (temporal intervals)**

# MPEG-7 timbral audio descriptors

**Spectrum basis**
**Power spectrum representation**

**Spectrum projection**



# MPEG-7 high level descriptors

**Signature scheme**

**Musical instrument timbre**

**Harmonic instrument: harmonic spectral centroid, harmonic spectral deviation, harmonic spectral spread, harmonic spectral variation, attack time**

**Percussive instrument: attack time, temporal centroid, spectrum centroid**

**Melody (monophonique, notes and rythm)**

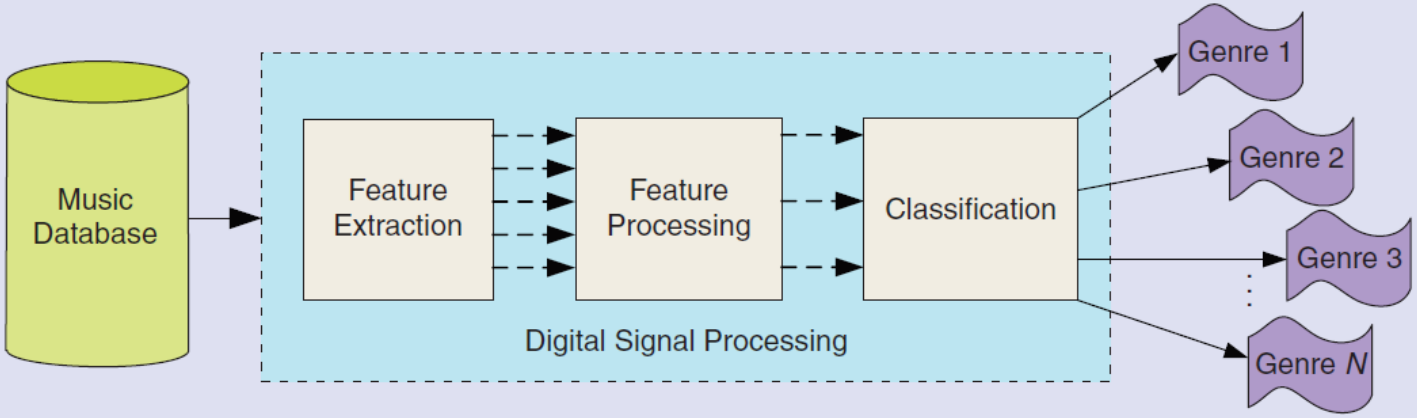
**Melody contour description scheme**

**Melody sequence description scheme**

**General Sound Recognition and Indexing : spectral basis statistical classification model**

**Spoken content description tools**

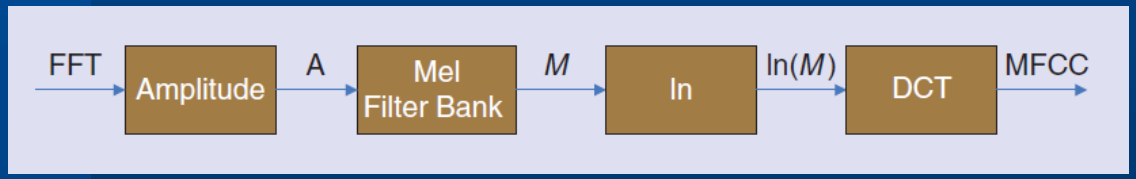
# Music classification




**Zero-crossing rate**

$$r_{zc} = \frac{1}{2(N-1)} \sum_{n=0}^{N-2} |\operatorname{sgn} x(n+1) - \operatorname{sgn} x(n)|$$

**Frequency band coefficients**




**Fundamental frequency**

H. Blume et al., **Huge music archives on mobile devices**, IEEE Signal Processing Magazine, 2011.