Multimedia

§7 Audio compression

Prof. Dr. Georg Umlauf

Hochschule Konstanz
University of Applied Sciences

Content

§7.1 The ear

Sound (1)

- Vibrations of the air pressure in the audible frequency band are called sound.
- Pressure is the force, that is applied to a certain surface area.
  - The unit of pressure is Pascal [Pa], where 1 Pa is defined as a force of 1 N (Newton) applied to a surface of 1 m².
  - In acoustics the pressure is measured in Micro-Pascal (µPa) :

$$1\ \mu\text{Pa} = 10^{-6}\ \text{N/m}^2.$$

  - English: Sound pressure = volume
  - German: Schalldruck = Lautstärke

§7.1 The ear

Sound (2)

- Perception of sound at optimal conditions:
  - Smallest sound pressure: $20\ \mu\text{Pa} = 0\ \text{dB}$
  - Largest sound pressure (pain level): $10^8\ \mu\text{Pa} = 150\ \text{dB}$

  ➡ **Sound pressure level (SPL, German: Schalldruckpegel):**

  Logarithmize the sound pressure p relative to the threshold of hearing $p_0 = 20\ \mu\text{Pa}$ measured in *dB SPL*

$$L(p) = 20 \cdot \log_{10}(p/p_0).$$

  - Examples:

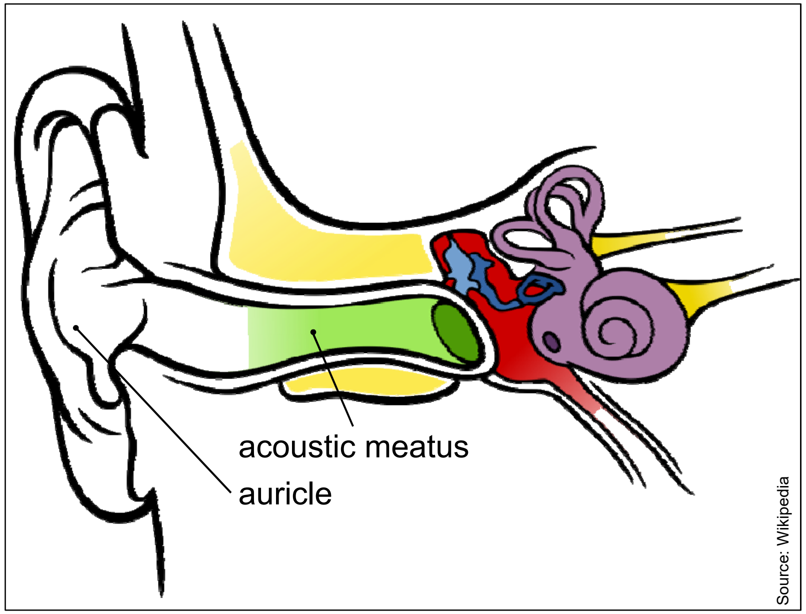
Sound	Sound pressure level [dB]
Rustling of leaves	20
Talk	60
Subway	100
Launching jet plane	140

§7.1 The ear

Outer ear

- Components: Auricle and acoustic meatus.
- Function: Sound conduction from the environment to the eardrum.
- Properties: Sound conduction depends on frequency and direction.

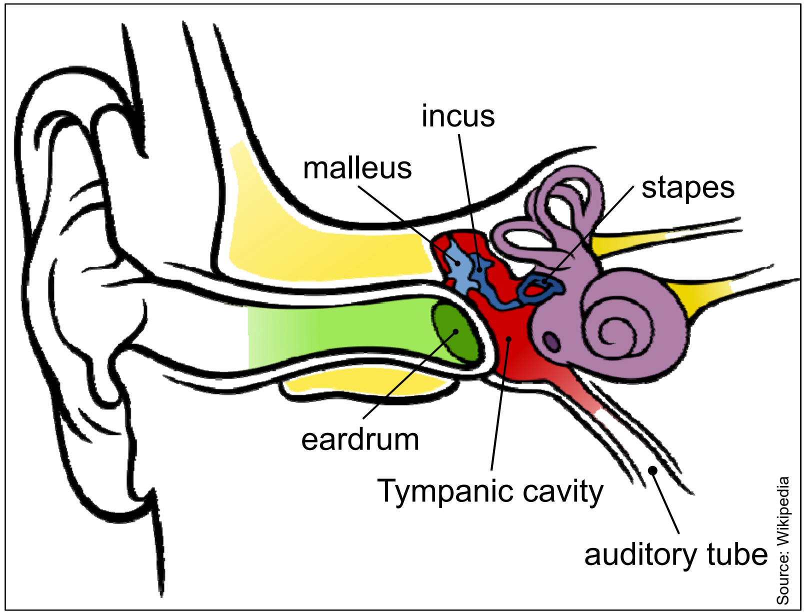
➡ Allows for spatial hearing.



§7.1 The ear

Middle ear

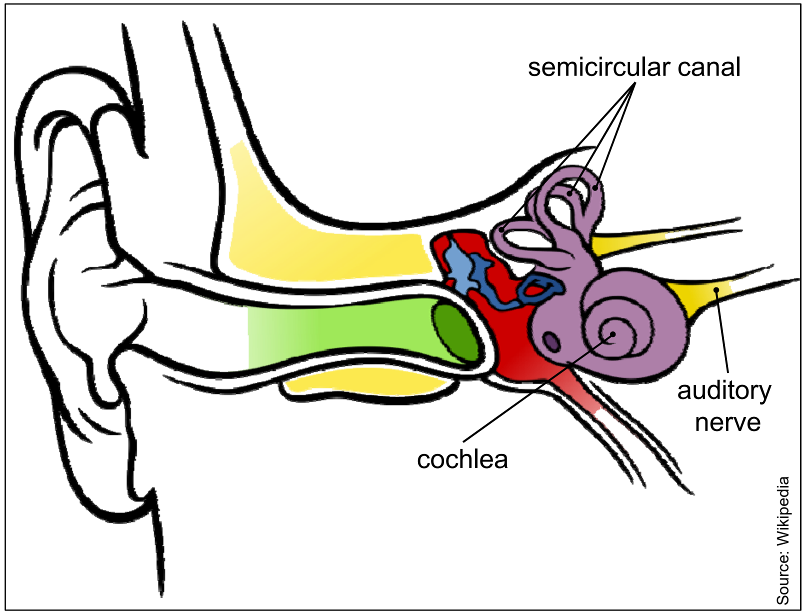
- Components: Eardrum, tympanic cavity (filled with air), two middle ear muscles, auditory ossicles (malleus, incus, stapes).
- Functions:
  - Transmission of vibrations from outer to inner ear.
  - Impedance adjustment between middle and inner ear.
  - Extension of the dynamic range of the ear.
  - Frequency dependent sensitivity shift of the ear.
  - Protection of the inner ear against excessive vibrations.



§7.1 The ear

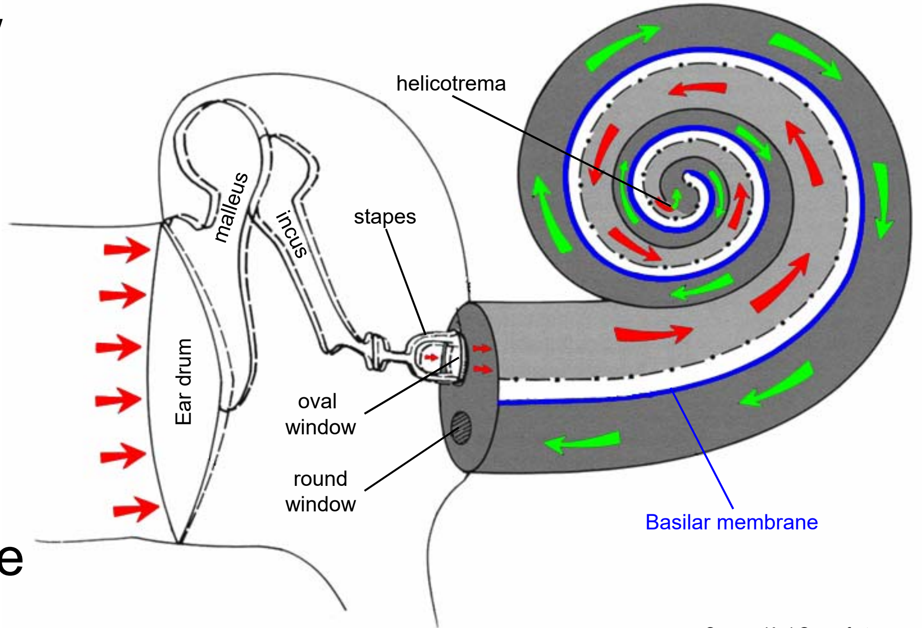
Inner ear

- Components: Cochlea, equilibrium organ (semicircular canal).
- Functions:
  - **Distribution of stimulus** to sensory cells (traveling wave).
  - **Transformation of stimulus** from mechanical vibrations to nerve impulses.



Distribution of the stimulus (1)

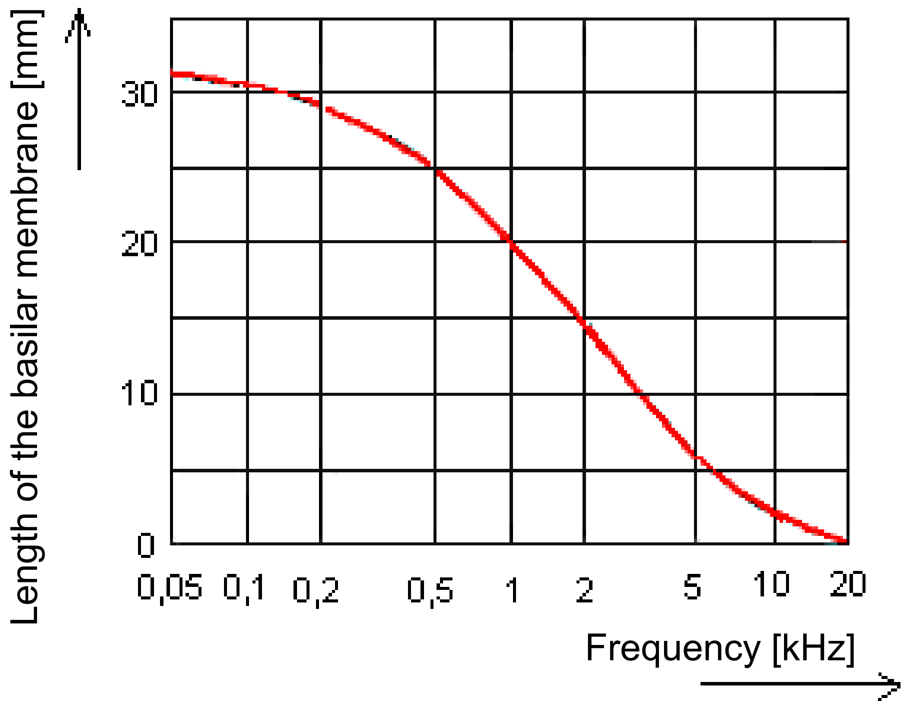
- Movement of the stapes results in a fluid movement and pressure change in the cochlea.
- The **basilar membrane** oscillates.
- A travelling wave forms on the basilar membrane.
- It propagates from the oval window along the membrane and yields its maximal amplitude at a frequency-dependent location on the membrane.
- Sounds with high frequencies are mapped to locations close to the oval window, sounds with small frequencies to locations close to the helicotrema.



Source: Karl Gegenfurtner

Distribution of the stimulus (2)

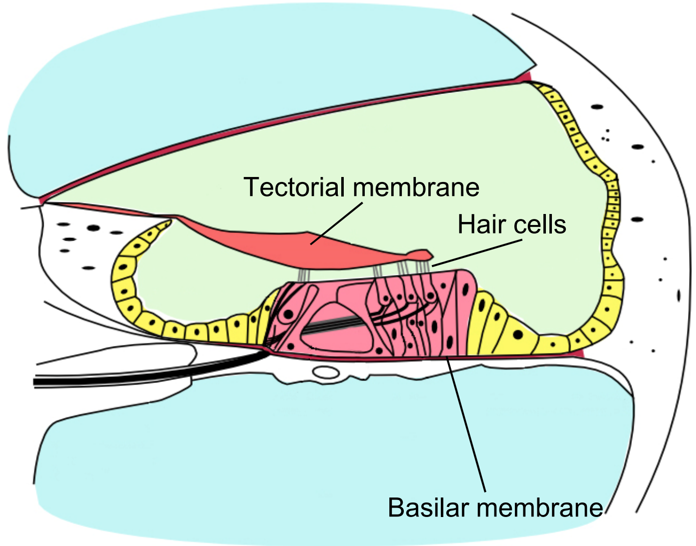
- Mapping of excitation frequency to the location on the basilar membrane with maximal amplitude.



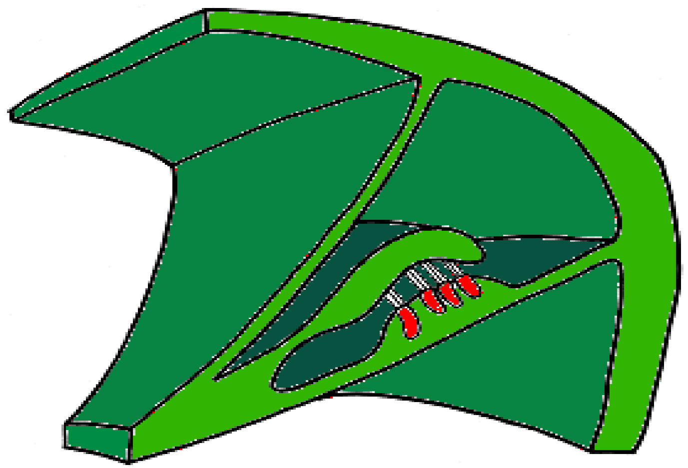
§7.1 The ear

Transformation of the stimulus

- This causes at the location of the maximal amplitude a relative movement of the basilar membrane to the tectorial membrane.
- ➡ Tangential shearing of the hair cells.
- ➡ This triggers the nerve impulse in the hair cells.
- ➡ Transmission via the acoustic nerve to the neural processing stages in the brain.



Source: wikipedia



§7.1 The ear

Conclusion

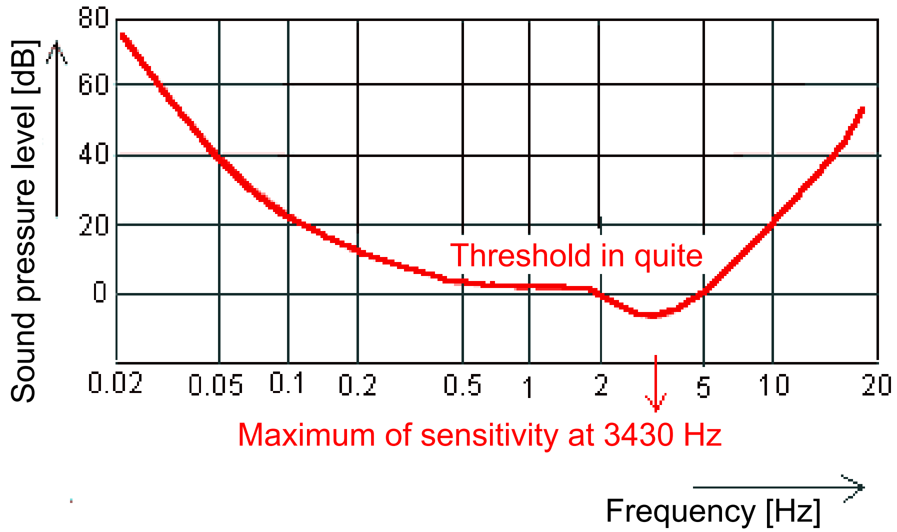
- The hearing organ transforms acoustic signals to the frequency domain.
  - ➡ The ear is a Fourier analyzer!
- This transformations of acoustic signals by the hearing organ yields a effective simplification of the acoustic pattern and reduces the amount of acoustic data.

Contents

§7.2 Psycho acoustics

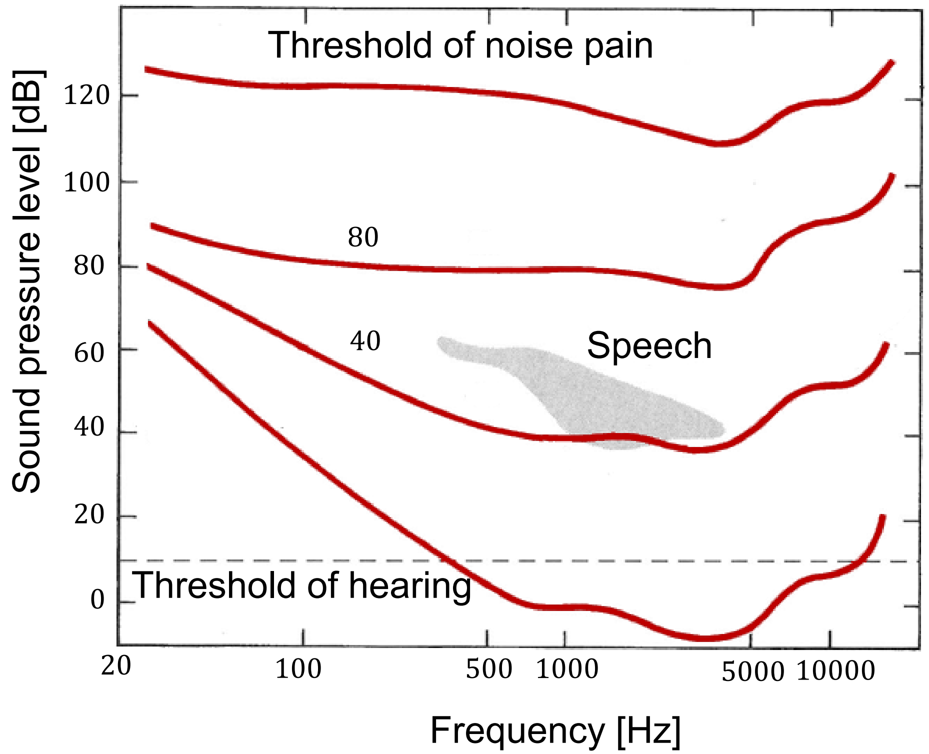
Perception of der acoustic intensity (1)

- The ear can perceive only acoustic stimuli within a certain frequency and sound pressure level range.
  - Frequencies in the range from 20 Hz to 20 kHz.
  - Sound pressure from 20 µPa or sound pressure level from 0 dB required.
  - **Threshold in quiet (Ruhehörschwelle):** Sound pressure level that is necessary to only just hear a sound depending on its frequency.



Perception of der acoustic intensity (2)

- The perception of the sound pressure depends on the frequency.
  - Physical quantity: Volume, sound pressure, sound pressure level.
  - Perceived quantity: Loudness.
  - German: Lautheit



Masking effects (1)

- In a mixed sound individual frequency components are perceived with different sensitivity.
  - Example: In the presence of loud bass sounds quite sound with middle or high frequencies cannot be perceived.

➡ The **masking threshold** (**Mithörschwelle**) is raised in the presence of background noise.

  - The masking threshold has a maximum at the location of the mid-frequencies of the background noise.

  ➡ The masking signal has only little influence on the perception of sound, whose frequency differs significantly from the masking noise mid-frequency.

  ➡ Sounds below the masking threshold can be omitted.

Masking effects (2)

- **Audio example:**
  - **Reference:** Sine wave at frequency 2kHz at threshold of quite ca. 0dB.
  - **Original:** Reference signal at 11 different sound levels, decreased by 3dB each.
  - **Example 1:** Original sequence distorted by band noise with mid-frequency 2kHz and band width 700Hz.
  - **Example 2:** As example 1 with band width 100Hz.

➡ In Example 2 fewer sounds are audible than in example 1.

Reference

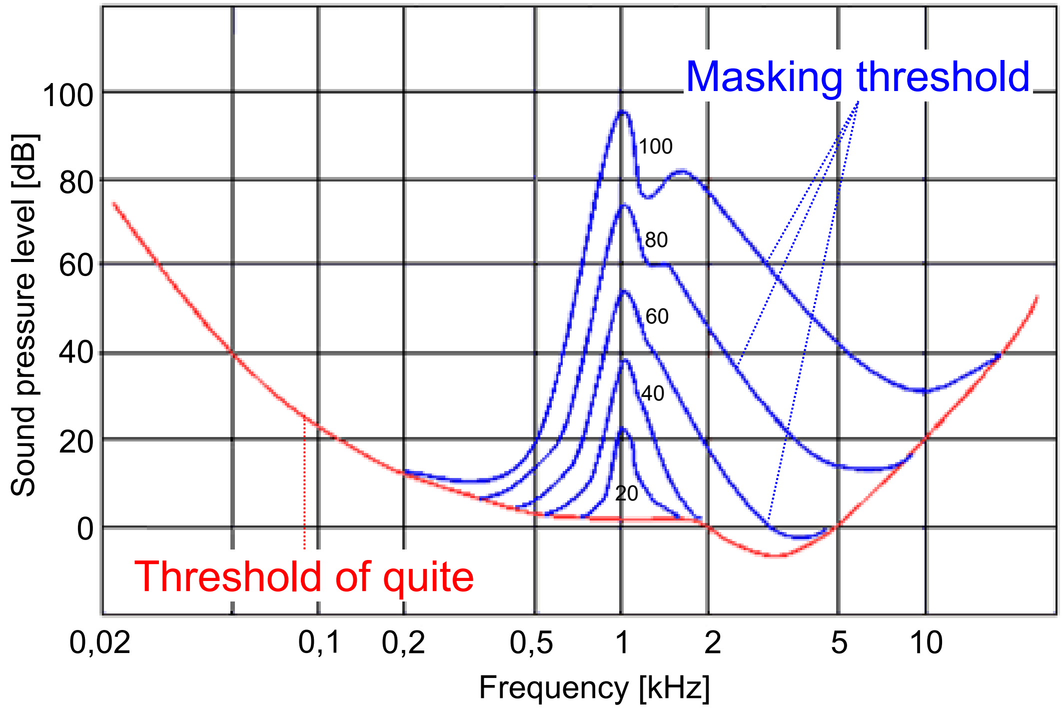
Original

B=700Hz

B=100Hz

Masking effects (3)

Example: Masking thresholds of sine waves that are masked by a narrow-band noise with mid-frequency at 1 kHz and band width 160 Hz for varying sound pressure levels of the masking noise.



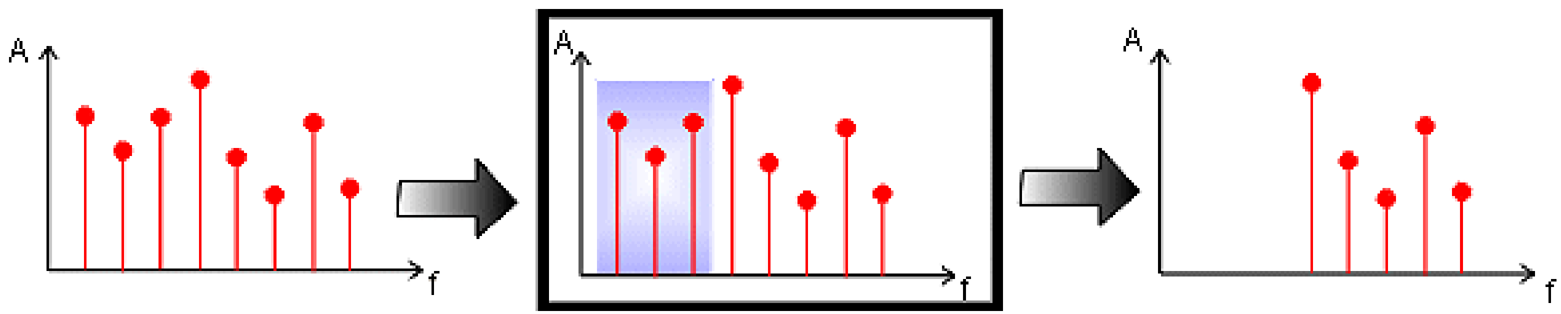
§7.2 Psycho acoustics

Masking effects (4)

- Frequency masking
  - A certain frequency component masks neighboring frequency components.
- Temporal masking
  - Two sounds, played in quick succession, can mask each other.

Virtual pitch level and residuum (1)

- The perceived pitch level of a sound corresponds usually to the pitch level of the fundamental oscillation (1st harmonic).
- The virtual pitch level arises, if from a broad-band line-spectrum only the high frequencies are transmitted.
- The resulting “residual sound", where the harmonic with low order are removed, is the so-called **residuum**.



Original sound signal

Filtering:
removal of the fist three harmonics

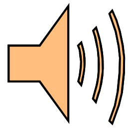
Spectrum of the residuum

§7.2 Psycho acoustics

Virtual pitch level and residuum (2)

- **Audio example (1):**
  - Remove from a sound of 15 harmonics at fundamental oscillation 200 Hz successively the first three harmonics.
  - After each removal of a spectral component a sine-sound at 200 Hz is played, in order to illustrate the constant pitch level.
  - ➡ The musical pitch level of the residuum does not change.

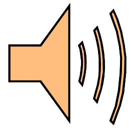
Sound with 15 harmonics



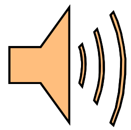
Virtual pitch level and residuum (3)

- **Audio example (2):**
  - The perception of the virtual pitch level is important for speech intelligibility at the phone or musical transmissions via channels, that transmit only a limited spectral range.
  - The frequency range of the telephone is limited from 300 Hz to 3400 Hz.
    - ➡ The first two or three harmonics of a sound signal are suppressed.
  - This has no influence on the perception of the pitch level.
    - ➡ The sense of hearing generates a real pitch level corresponding to the virtual pitch level.

Without filtering



With filtering



Directional hearing and sound source localization (1)

- Horizontal sound source localization is based on the difference of the sound signals in both ears.
- There are **temporal** and **level differences** in the ear signals.
- For spatial orientation of natural sound signals always both are used.
  - But each types of signal differences alone can also lead to a sound source localization.

§7.2 Psycho acoustics

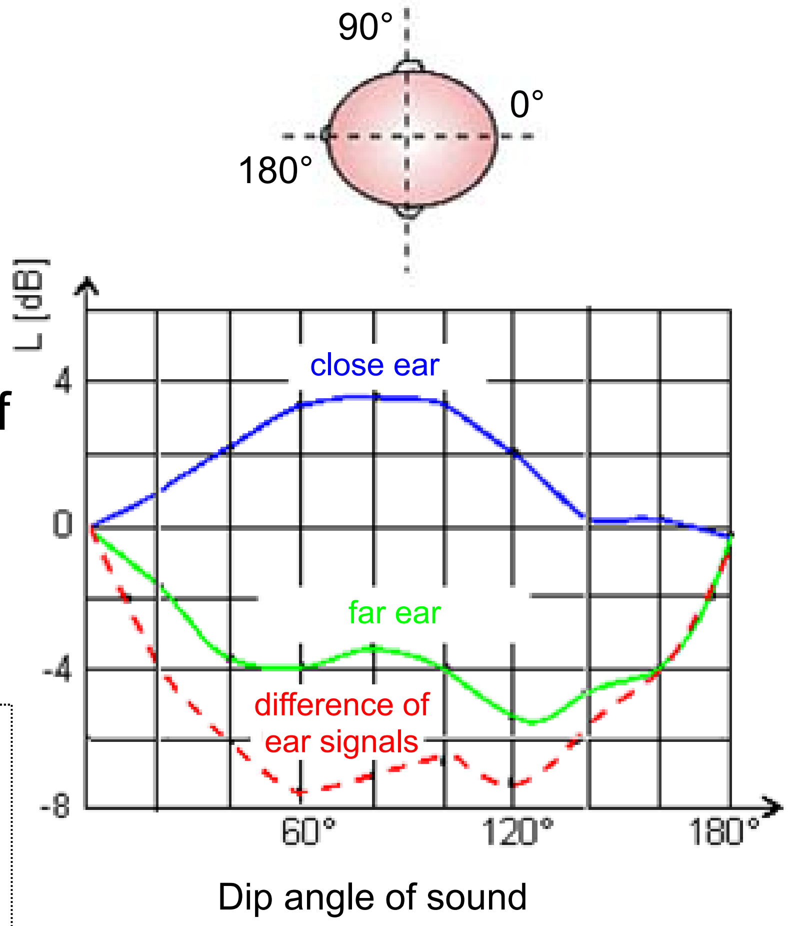
Directional hearing and sound source localization (2)

- The differences between the individual ear signals is described by the **interaural transmission function**.
  - If a sound source is immediately in front of or behind a person, both eardrums perceive the same ear signal.
  - If a sound source in not immediately in front of a person each eardrum perceives a different ear signal, because
    - of the different geometric location of the ears relative to the sound source and
    - of the head, which is an acoustic obstacle.

Directional hearing and sound source localization (3)

Example

- Interaural transmission function of sound events between 500 Hz and 2500 Hz.
- For higher frequencies the transmission functions are inclined towards the low levels because of the acoustic shadow of the head.
- There is hardly any direction information in very low/high frequencies perceivable.
- Audio examples: **plane** **birds**



Contents

Parameters for audio coding

- Frequencies in the signal
- Number of possible channels
- Sampling rate
- Sampling depth: Quantization

Type	Frequency [Hz]	Sampling rate [Hz]	Sampling depth [bit]	Channels
Phone	200 − 3.400	8.000	8	1
Radio	50 − 11.000	22.050	8	2
CD	20 − 20.000	44.100	16	2
Studio	20 − 20.000	48.000	24	n

Overview

- Un-compressed formats:
  - Linear Pulse-Code-Modulation (LPCM) in various forms
  - E.g.: wav
- Lossless compressed formats:
  - m4a (aka Apple Lossless, MPEG-4 ALS)
- Lossy compressed formats:
  - Adaptive Differential Pulse Code Modulation (ADPCM)
    - DECT (cordless phone)
  - mp3
  - m4a (aka MPEG-4 AAC)

History

- **MPEG – Moving Picture Experts Group:** workgroup of the ISO/IEC for audio- and video-coding, since 1988.
- **MPEG-1 (1992):** contains e.g. MPEG-1 audio layer III (MP3), video CD.
- **MPEG-2 (1994):** contains e.g. DVD, DVB (digital tv), DAB (digital radio)
- **MPEG-4 (1998):** Multimedia-Standard, object-oriented, scalable
  - MPEG-4 Part 2 (Video): H.263
  - MPEG-4 Part 10 (Video): AVC, H.264 ➡ HD-DVD, Blu-ray Discs
  - MPEG-4 Part 3 (Audio): AAC, ALS ➡ MP3-sucessor (.m4a)
- **MPEG-7 (2002):** Multimedia Content Description (Meta-data –"Bits about the bits")
- **MPEG-21 (2001):** "Framework for multimedia delivery and consumption"
- **MPEG-DASH (2012):** „Dynamic adaptive streaming over HTTP".
- **MPEG-H (2013):** „High efficiency coding and media delivery in heterogeneous environments", H.265

MP4

Container format for MPG-4 contents

- **Video:** MPG-4/Part 2 (Video) MPG-4/Part 10 (AVC), MPG-2, MPG-1
- **Audio:** MPG-4/Part 3 (Audio), AAC, MP3, MP2, MP1
- **Image:** PNG, JPG
- **Graphics:** BIFS (**BI**nary **F**ormat for **S**cences) (MPG-4 Part 11)

QuickTime

Container format for multimedia contents by Apple:

- **Video:** animated GIF, H.26x, etc.
- **Audio:** AAC, MP3, Apple lossless (.m4a) etc.
- **Image:** BMP, GIF, TIFF, PNG, JPG(2000), etc.

Flash Video

Container-Format for

- **Video:** H.264
- **Audio:** AAC, MP3

MPEG-1 (1)

- First international standard defined 1992 as ISO/IEC 11172-3 (1993).
- Bit rats up to 1.5 Mb/s for video with audio (Video-CD).
- Defines only the function of the decoder the format of the audio-bit-stream, but not the encoder to allow for later improvements.
- Format suitable for speech and music.
- No assumptions about the source, but uses a psycho-acoustic model instead to use masking effects.
- The variants (layers), each in mono / stereo / joint stereo.
- Sampling rates 32, 44.1, 48 KHz.
- Bit rates 32 .. 224 kb/s/channel.

MPEG-1 (2)

Layer 1:
- simplest algorithm
- for bit rates larger than 128 kb/s per channel

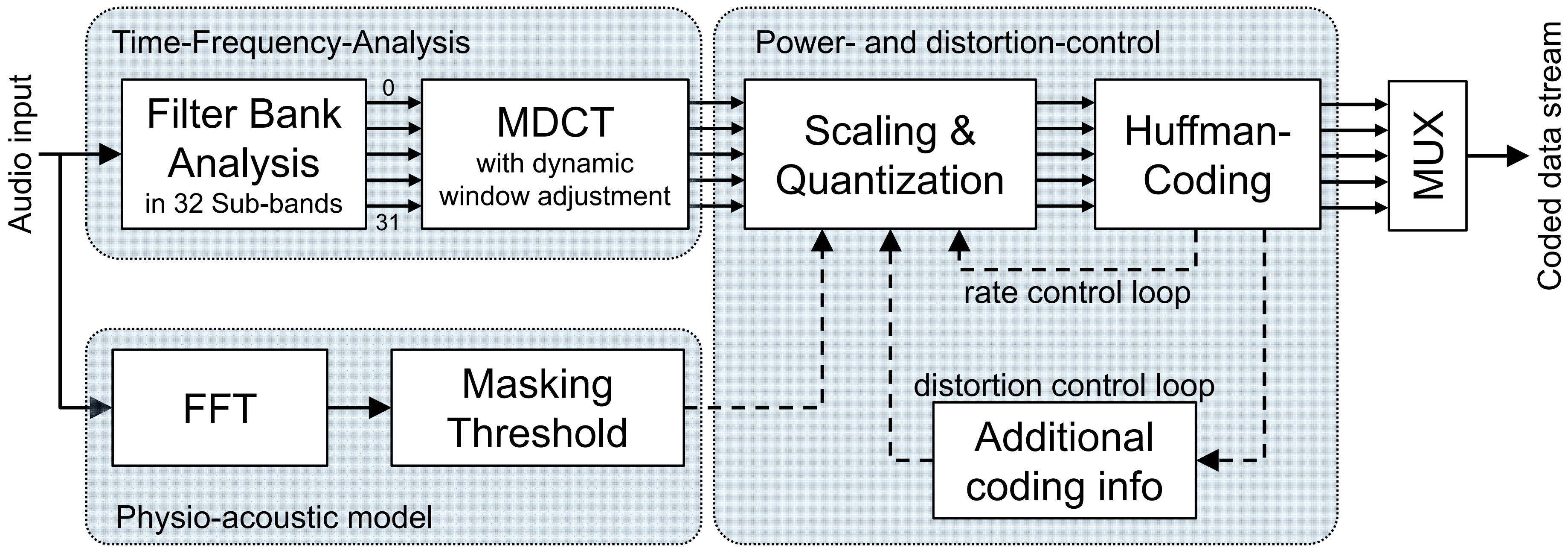
Layer 2:
- middle complexity
- used for CD-I and Video-CD
- for bit rates of 128 kb/s per channel

Layer 3:
- Better quality, but significantly more complex
- Starting at 64 kb/s per channel, good quality above 128 kb/s
- "MP3", ISDN-Transmission

MP3 (1) (MPEG-1 Audio Layer III)

- Sampling rates: 32kHz, 44.1kHz, 48 kHz
- Supports 2-channel stereo signals
- Supports CRC-checksum for error detection
- Compression rate around 10:1.
- A MP3-file has no explicit header, it is a list of subsequent data blocks, each having its own header and audio information
  - ➡ Streaming

MP3 (2)



- rate control loop: Control of chosen data rate
- distortion control loop: Control quantization noise below the hearing threshold.

MP3 (3)

1. **Filter bank Analysis (sub-band-coding)**
   - FIFO-Buffer of 512 samples, adding only 32 new samples per step.
   - Audio spectrum is partitioned by filter bank into 32 uniform and overlapping frequency bands.
2. **MDCT** for each frequency band
   - **Dynamic window adjustment (block length switching):** Dependent on temporal variation of the signal.
     - *Stationary signals:* One time frame for 36 samples, yielding 18 DCT-coefficients each.
     - *Highly varying signals:* Three time frames of 12 samples, yielding 6 DCT-coefficients each.

MP3 (4)

3. **Psycho-acoustic model**
   a) **Fourier-transformation** of the signal.
   b) Computation of der **masking thresholds:**
      - Between frequency bands (frequency/temporal) masking occurs.
      - Almost all frequency bands carry less relevant information than the loudest frequency band.

      ➡ Yields control parameters for the non-uniform quantization.
4. **Scaling & Quantization**
   - Non-uniform quantization of groups of frequency bands.

   ➡ Lossy compression step!
5. **Huffman coding** using a fixed code table.
6. **Multiplexer**

MP3 (5)

Coding of stereo channels

- **Intensity Stereo:** Coding of certain frequency ranges only mono and enrich these with „direction information“ from the other frequency bands.
  - ➡ Phase differences are neglected.
  - ➡ Only amplitude differences are coded.
- **Mid/Side-Stereo:** If left L and right channel R are very similar, transmit $\mathrm{L}+\mathrm{R}$ and $\mathrm{L}-\mathrm{R}$ instead of L and R .
  - Switch transmission depending on coding efficiency.

MP3 development

MPEG 2 AAC (Advanced Audio Coding)

- Improved prediction algorithms in the psycho acoustic model.
- Up to 48 regular channels + 16 low frequency channels.
- Sampling rate up to 96 kHz.
- Frame size up to 2048 samples: improved temporal and frequency resolution.
- Temporal Noise Shaping: Control of the quantization noise.
- Quality like MP3 with only 70% of necessary bit rate.

MPEG 4 AAC

- Specialized for **Mobile Computing** and voice transmission.
- From 4 kbps on intelligible voice transmission.
- Perceptual Noise Substitution (PNS) and Long Term Prediction (LTP).

MPEG-4 ALS (1) (Audio Lossless Coding)

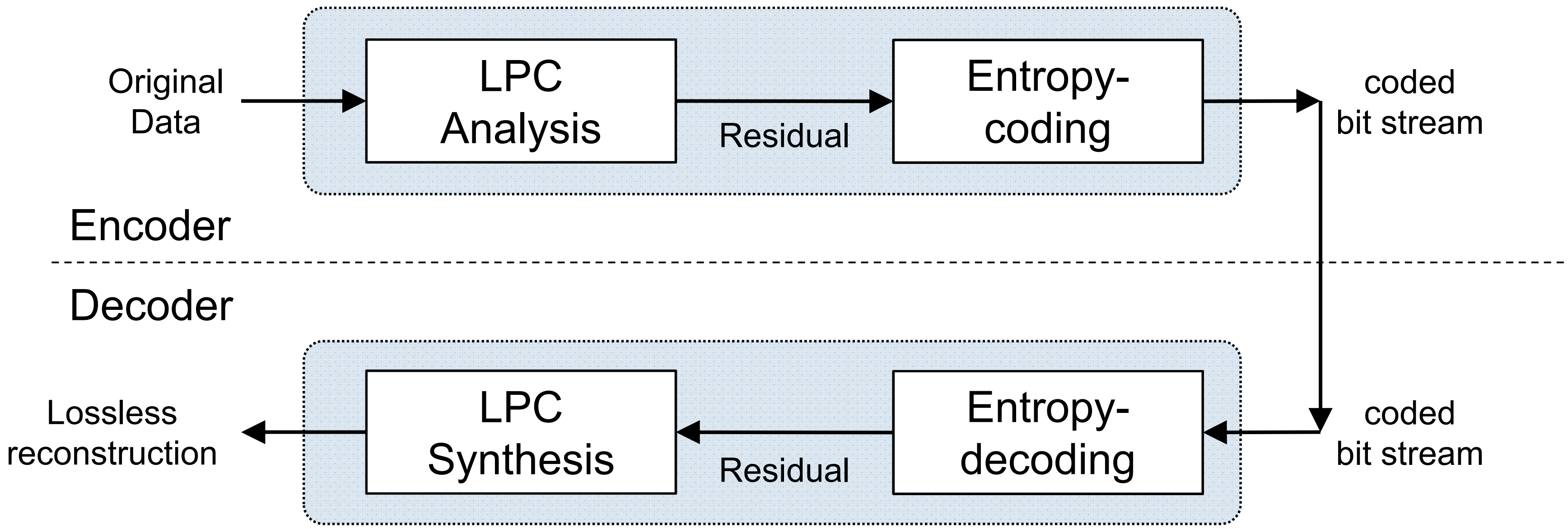
- Audio-Compression with perfect (bit-identical) reconstruction.
- **But:** Universal compression methods fail for audio signals.

➡ **MPEG-4 Audio Lossless Coding**

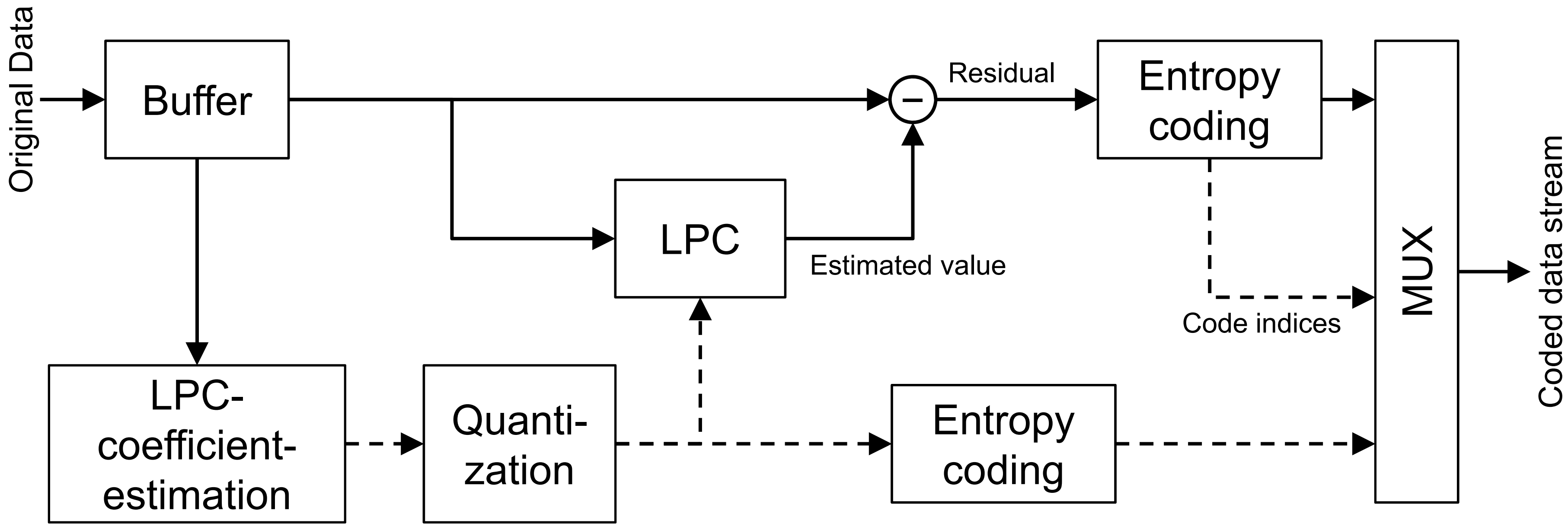
- Lossless coding of audio signals with very large resolutions (16 to 24 bit, 44,1 to192 kHz)
- Random access (direct access) to individual code segments.
- **Applications**
  - Professional: archiving, recording studios, distributed sound editing, …
  - Private users: high-resolution disks, online music shops, …

MPEG-4 ALS (2)

- **Typical approach**
  - Decorrelation of the audio signal (Prediction / Transformation) using linear prediction (LPC=linear predictive coding).
  - Entropy-coding of the de-correlated samples.



MPEG-4 ALS (3)


Original Data
Buffer
Residual
Entropy coding
LPC
Estimated value
Code indices
MUX
Coded data stream
LPC-coefficient-estimation
Quanti-zation
Entropy coding

MPEG-4 ALS (4)

1. **Buffer**
   - Input buffer for one frame of samples
2. **LPC-coefficient-estimation** (see page §7/43)
   - Compute the optimal LPC-coefficients
3. **Quantization** of the LPC-coefficients (see page §7/47)
4. **LPC (linear predictive coding)**
   - Compute the residual error signal (Restfehlersignal)
5. **Entropy-Coding:** Use various Rice-codes
6. **Multiplexer:** residual error signal, code-indices and coefficients

Linear predictive coding (1)

- The actual value $x(n)$ is estimated from the previous values $x(n-i)$, $i = 1, \dots, K$, from then past

$$\hat{x}(n) = \sum_{i=1}^{K} h_i \cdot x(n-i)\,.$$

- The residual error signal (residual) is

$$e(n) = x(n) - \hat{x}(n).$$

- Which predictor coefficients h_i yield a minimal residual?

➡ The previous values and the actual value are highly correlated.

➡ Minimize e.g. the mean square error $E[e^2(n)]$ (MSE).

Linear predictive coding (2)

- Compute the minimum of $E[e^2(n)]$ depending on the coefficients $\boldsymbol{h} = (h_1, \dots, h_K)^t$.

➡ $\frac{\partial}{\partial h_j} E\left[\left(x(n) - \hat{x}(n)\right)^2\right] = \frac{\partial}{\partial h_j} E\left[\left(x(n) - \sum_{i=1}^{K} h_i \cdot x(n-i)\right)^2\right] = 0$

for $j = 1, \dots, K$.

➡ $\sum_{i=1}^{K} h_i \cdot E[x(n-i)x(n-j)] = E[x(n)x(n-j)]$.

  ➡ These are K equalities for K unknowns h_i.

- The values $E[x(n-i)x(n-j)] = R_{xx}(|i-j|)$ are the auto-correlation values of the signal x with itself.

Linear predictive coding (3)

- The auto-correlation is computed within a window of width N
$$R_{xx}(k) = \sum_{n=n_0+k+1}^{n_0+N} x(n-k)x(n)\,.$$
- These K equations can be written in matrix form as $\boldsymbol{R} \cdot \boldsymbol{h} = \boldsymbol{P}$ with $P = (R_{xx}(1), \dots, R_{xx}(K))^t$ and
$$\boldsymbol{R} = \begin{bmatrix} R_{xx}(0) & R_{xx}(1) & R_{xx}(2) & \cdots & R_{xx}(K-1) \\ R_{xx}(1) & R_{xx}(0) & R_{xx}(1) & \cdots & R_{xx}(K-2) \\ R_{xx}(2) & R_{xx}(1) & R_{xx}(0) & \cdots & R_{xx}(K-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{xx}(K-1) & R_{xx}(K-2) & R_{xx}(K-3) & \cdots & R_{xx}(0) \end{bmatrix}$$
- The matrix $\boldsymbol{R}$ is a Toepliz-matriz, i.e. $\boldsymbol{R}^{-1}$ can be computed using the Levinson-Durbin-algorithm in $O(n^2)$.

Adaptation of the order of the predictor

- A larger order K yields a smaller bit rate for the residual error signal (i.e. better prediction), but a larger bit rate for its coefficients.
- The order is optimal, if the predictor minimizes the total bit rate.
- Adaption within the Levinson-Durbin-Algorithm is possible.

Quantization of the LPC-coefficients

- **Problem:** Direct quantization of the LPC-coefficients is in-efficient, because small errors can lead to large spectral distortions.

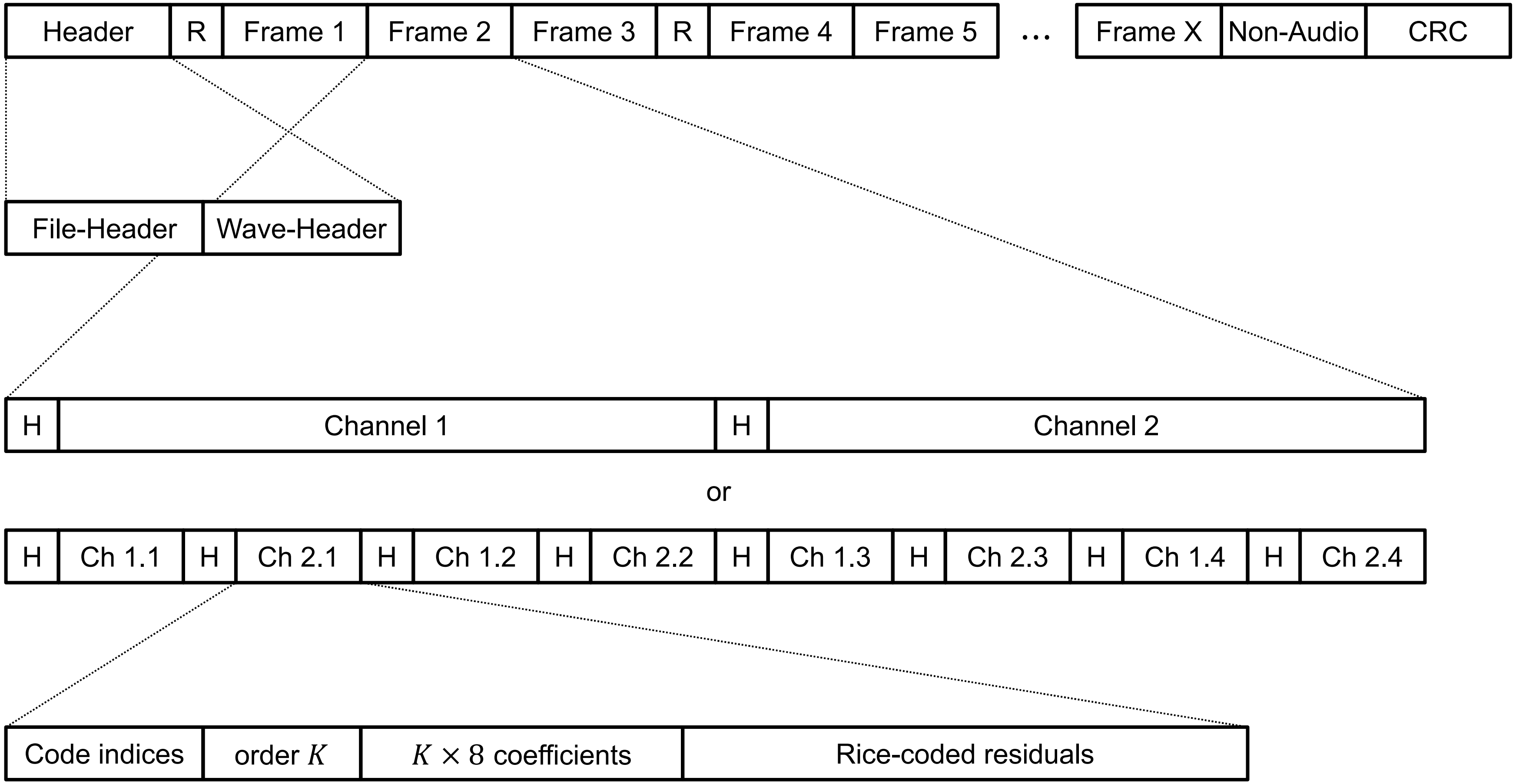
➡ **Arcsine-coefficients**

- Non-linear mapping $\alpha_k = \arcsin(h_k)$ expands the range by one.
- Linear quantization using 8 bit/coefficient for transmission.
- Predictor filter uses LPC-coefficients, computed from the quantized arcsine-coefficients (in fix-point-arithmetic).

Further properties

- **Block length switching**
- **Joint Stereo Coding** (German: Verbundcodierung)
- **Random Access:** Direct access to arbitrary segments of the coded audio signal, without decoding of preceding segments.
  - Minimal temporal resolution of 0.5 seconds
  - **Realization**
    - Generate random access (RA) frames every 0.5 seconds.
    - A RA frame contains the distance (in Bytes) to the next RA frame.
    - Submit the first K (= predictor order) original samples for prediction independent to the previous frame
    - Uses additionally 0.5 –2kbit/s, depending on the predictor order and resolution of the audio signal.

Bit-stream-format



Rate of compression

- Rate of compression for maximal CPU-load

Sampling rate/ sampling depth	MPEG-4 ALS
48 kHz/ 16 bit	46,5
48 kHz/ 24 bit	64,0
92 kHz/ 16 bit	31,1
96 kHz/ 24 bit	47,1
192 kHz/ 16 bit	21,9
192 kHz/ 24 bit	38,2

Goals

- What is masking?
- What is the principle of directional hearing?
- What are the basic building blocks of MP3?
- What is block length switching?
- How is stereo information coded?
- What are the basic building blocks of MPEG-4-ALS?
- What is LPC and how does it work?