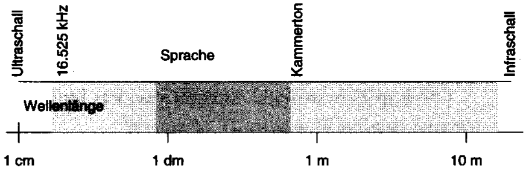
Audio Coding and MP3

Wolfgang Leister
contributions by:
Torbjørn Ekman

Norsk Regnesentral

What is Sound?

- Sound waves: 20Hz - 20kHz
- Speed: 331.3 m/s (air)
- Wavelength: 165 cm - 1.65 cm



Analogue audio

- frequencies: 20Hz - 20kHz
- mono: x(t) scalar
- stereo: $x(t) = \begin{bmatrix} x_r(t) \\ x_l(t) \end{bmatrix}$

Audio Compression

- small files, low data rate at transmission
- reconstruction must be (as much as possible) similar to original signal
- redundancy (lossless coding)
- irrelevancy (do not code what you cannot hear)

Data rates

Quality	Sample Rate	Bit/Sample	Channels	Data Rate kb/s	Frequency
Telephone	8.000	8	Mono	64,00	200-3400
MW	11.025	8	Mono	88,00	
UKW	22.050	16	Stereo	705,60	
CD	44.100	16	Stereo	1411,00	20-20000
DAT	48.000	16	Stereo	1536,00	20-20000

Dynamics compression

- A-Law

$$S' = \begin{cases} sign(S) \cdot \dfrac{A \cdot abs(S)}{1 + \ln A} & \text{for} \quad abs(S) \le \dfrac{1}{A} \\ sign(S) \cdot \dfrac{1 + \ln(A \cdot abs(S))}{1 + \ln A} & \text{else} \end{cases}$$

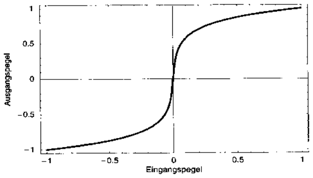
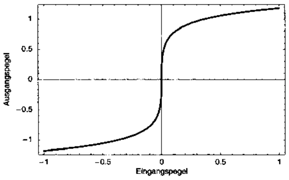


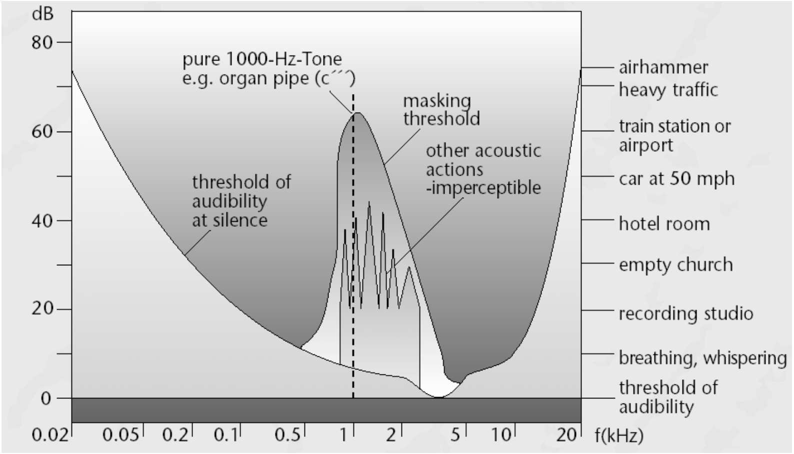
Bild 5.2: A-Law-Kompressionskennlinie für normierte Signalpegel

- μ-Law

$$S' = sign(S) \cdot \frac{1 + \ln(1 + \mu \cdot abs(S))}{\ln(1 + \mu)}, \mu = 255$$

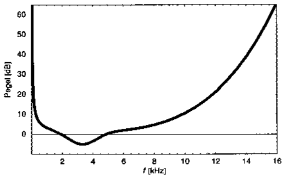


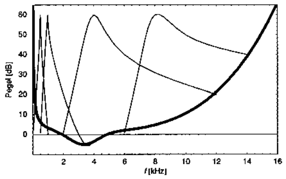
Masking

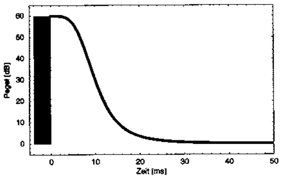


Masking

- Threshold for human ear
- Threshold changes:
  - neighbouring frequencies
    (Example 0.5, 1, 4, 8 kHz)
  - in time







Sampling

- When x(t) is bandwidth-limited:

$$|f| > \omega \quad \Rightarrow \quad x(f) = 0$$

- then

$$x(t) = \sum_{n=-\infty}^{\infty} x[n] g(t - n \cdot \Delta t)$$

- with $\Delta t = \frac{1}{f_s} < \frac{1}{2\omega}$ $\quad x[n] = x(n \cdot \Delta t)$ $\quad g(t) = \frac{\sin(2\pi\omega t)}{2\pi\omega t}$

Quantisation

- $x \to Q(x)$
- k bits $\Rightarrow$ $\mathrm{L} = 2^{\mathrm{k}}$ representations
- $\{y_1, \dots, y_n\}$
- $|x - y_i| \le |x - y_j| \quad \Rightarrow \quad Q(x) = y_i$

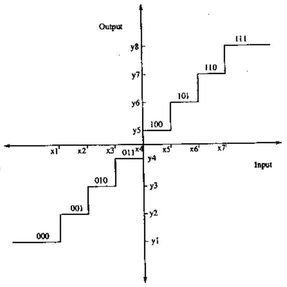


Figure 2.2 The input output characteristics of a uniform quantiser

Scalar Quantisation

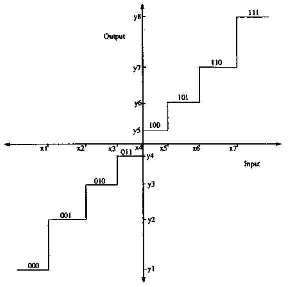
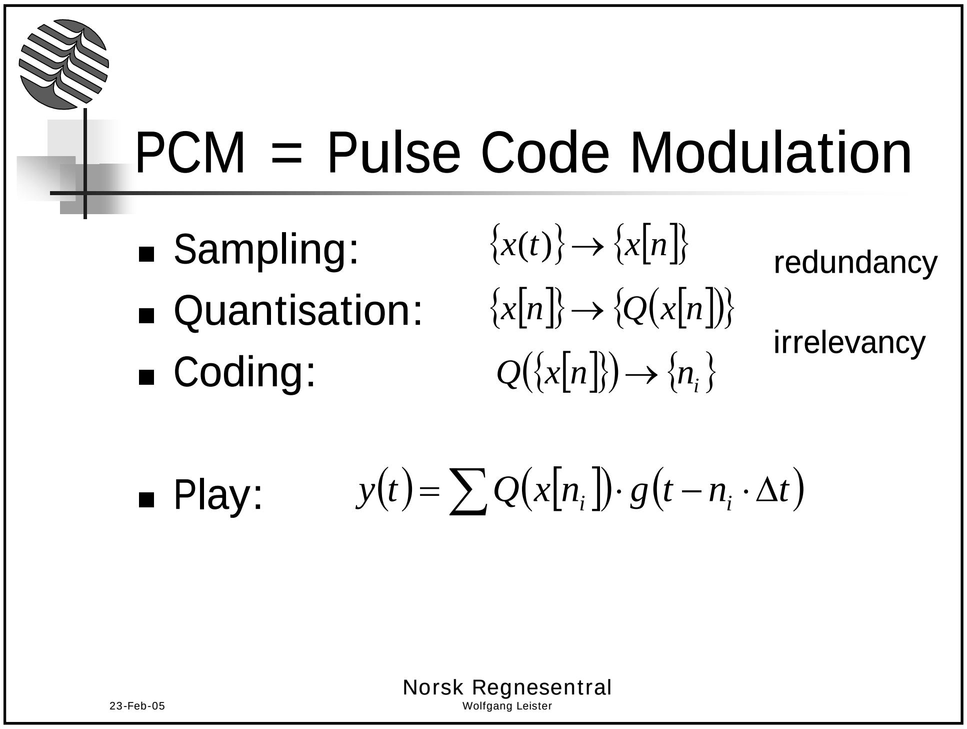


Figure 2.3 The input output characteristics of a non-uniform quantiser



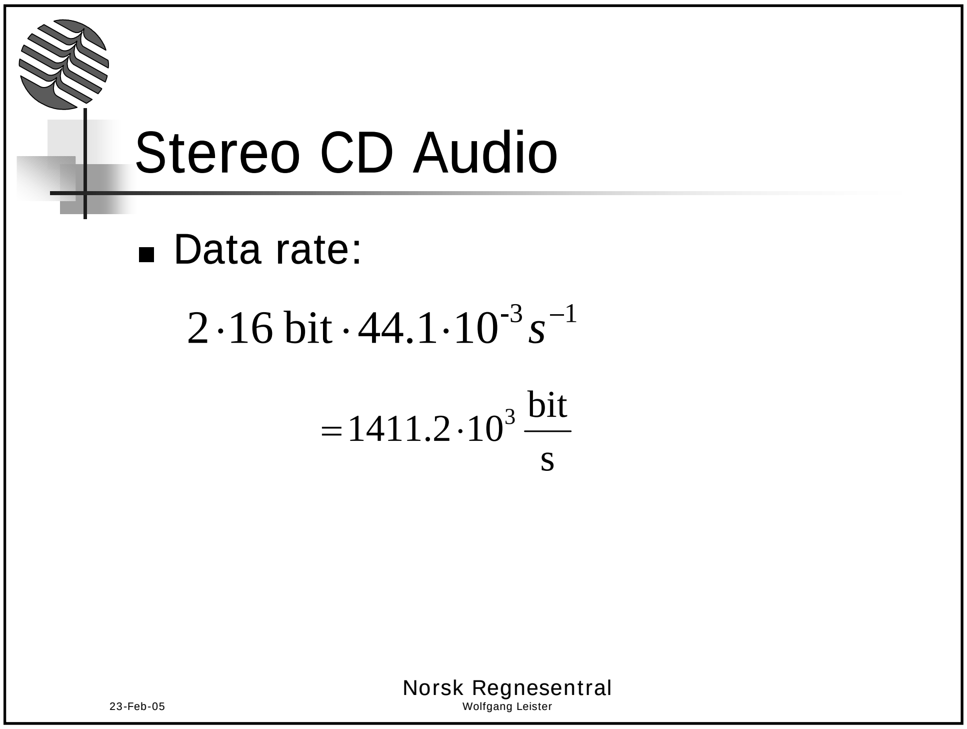
PCM = Pulse Code Modulation

- Sampling: $\{x(t)\} \rightarrow \{x[n]\}$
- Quantisation: $\{x[n]\} \rightarrow \{Q(x[n])\}$
- Coding: $Q(\{x[n]\}) \rightarrow \{n_i\}$

redundancy

irrelevancy

- Play: $y(t) = \sum Q(x[n_i]) \cdot g(t - n_i \cdot \Delta t)$

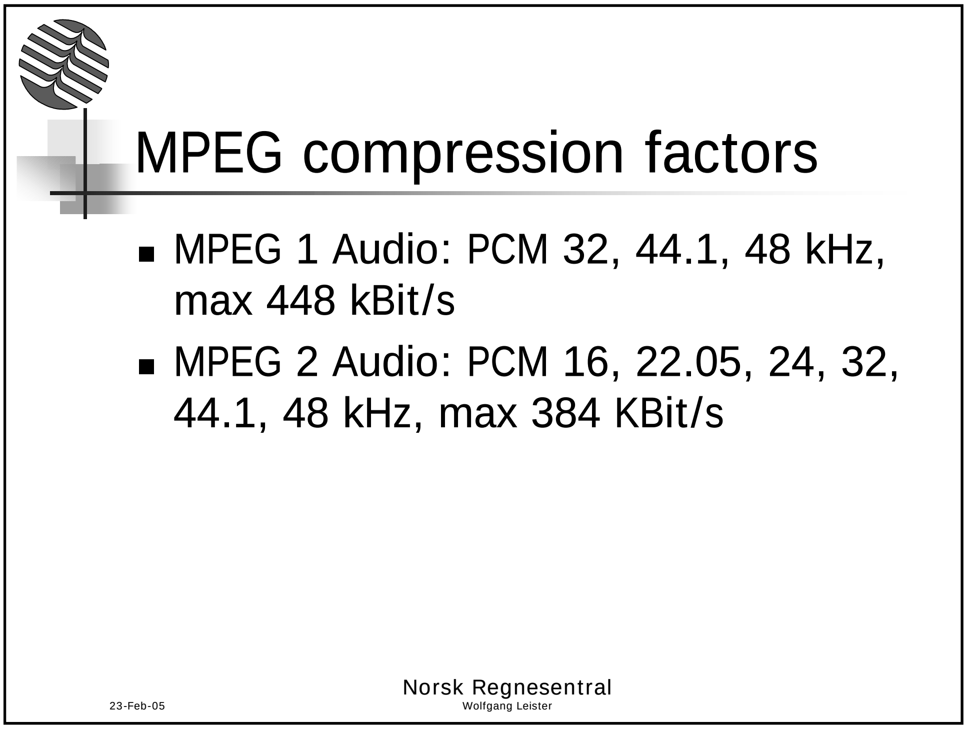


Stereo CD Audio

- Data rate:

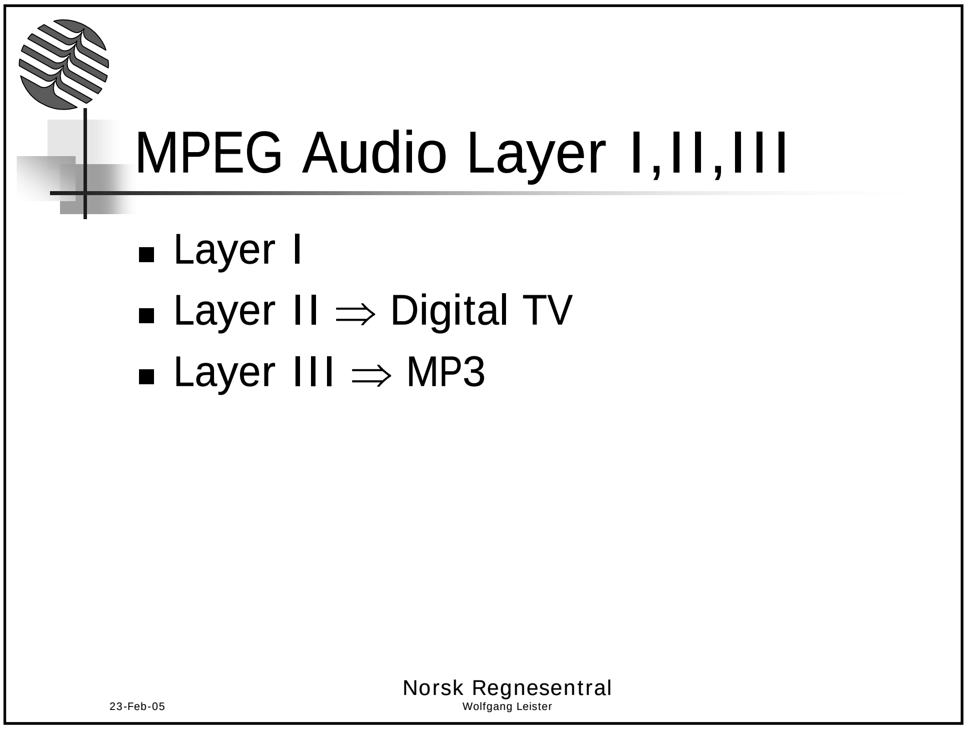
$$2 \cdot 16\,\mathrm{bit} \cdot 44.1 \cdot 10^{-3} s^{-1}$$

$$= 1411.2 \cdot 10^{3} \frac{\mathrm{bit}}{\mathrm{s}}$$



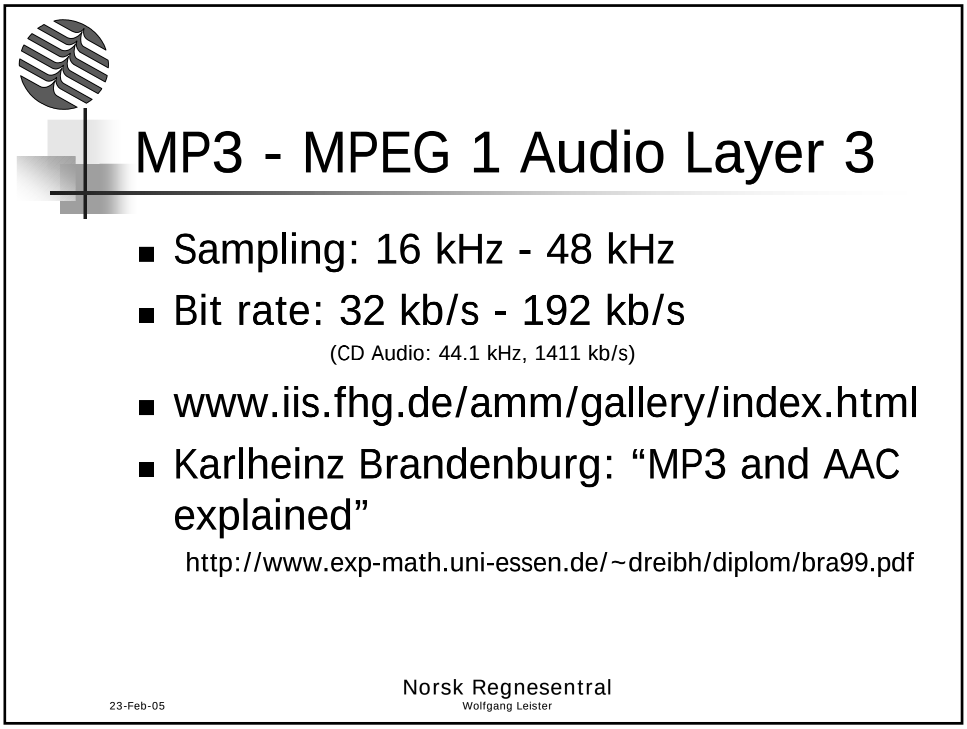
MPEG compression factors

- MPEG 1 Audio: PCM 32, 44.1, 48 kHz, max 448 kBit/s
- MPEG 2 Audio: PCM 16, 22.05, 24, 32, 44.1, 48 kHz, max 384 KBit/s



MPEG Audio Layer I,II,III

- Layer I
- Layer II ⇒ Digital TV
- Layer III ⇒ MP3



MP3 - MPEG 1 Audio Layer 3

- Sampling: 16 kHz - 48 kHz
- Bit rate: 32 kb/s - 192 kb/s

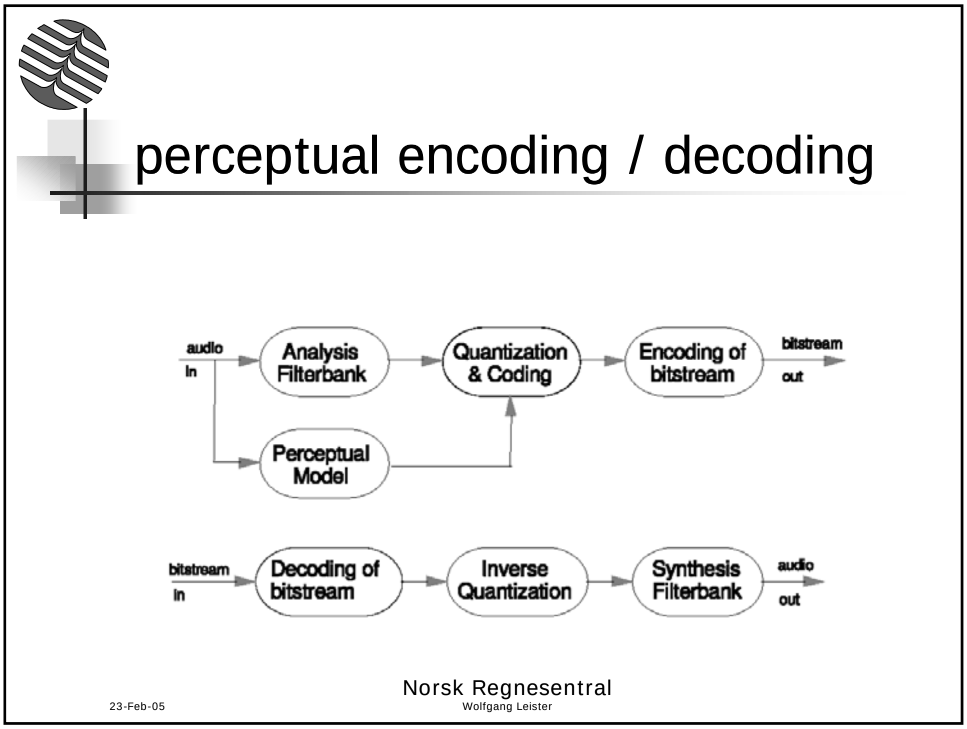
  (CD Audio: 44.1 kHz, 1411 kb/s)
- www.iis.fhg.de/amm/gallery/index.html
- Karlheinz Brandenburg: “MP3 and AAC explained”

  http://www.exp-math.uni-essen.de/~dreibh/diplom/bra99.pdf

Norsk Regnesentral

23-Feb-05

Wolfgang Leister



perceptual encoding / decoding

audio in

Analysis Filterbank

Quantization & Coding

Encoding of bitstream

bitstream out

Perceptual Model

bitstream in

Decoding of bitstream

Inverse Quantization

Synthesis Filterbank

audio out

Norsk Regnesentral

23-Feb-05

Wolfgang Leister

Filterbank

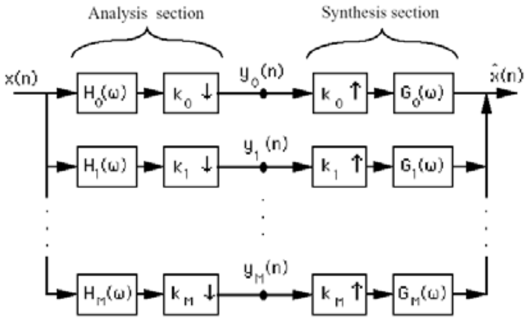


Figure 4.2: An analysis/synthesis filter bank.

$$\begin{aligned}\hat{X}(\omega) &= \frac{1}{k}\sum_{i=0}^{M-1}\left[\sum_{j=0}^{k-1} H_i\left(\omega+\frac{2\pi j}{k}\right) X\left(\omega+\frac{2\pi j}{k}\right)\right] G_i(\omega) \\ &= \frac{1}{k}\sum_{i=0}^{M-1} H_i(\omega) G_i(\omega) X(\omega) \\ &\quad + \frac{1}{k}\sum_{j=1}^{k-1} X\left(\omega+\frac{2\pi j}{k}\right)\sum_{i=0}^{M-1} H_i\left(\omega+\frac{2\pi j}{k}\right) G_i(\omega) \qquad (4.2)\end{aligned}$$

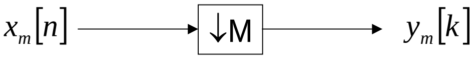
Ideal sub-band coder

- impossible: ideal sub-band coder
- downsampling ⇒ aliasing
- possible: “nearly perfect”

$$H_m(f) = \begin{cases} 1 & \text{for } |\text{f}| \in D_m, m = 1, \dots, M \\ 0 & \text{else} \end{cases}$$

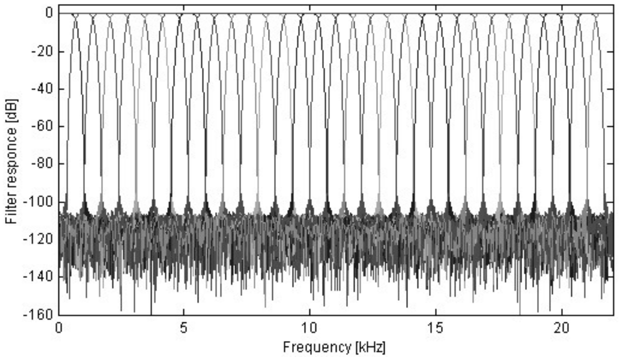
Downsampling

- from $M \cdot f_s$ back to f_s
- sub-bandwidth B, upper frequency is multiple of B
- can sample at $f_s = 2B$
  (instead of $f_s = 2M \cdot B$)



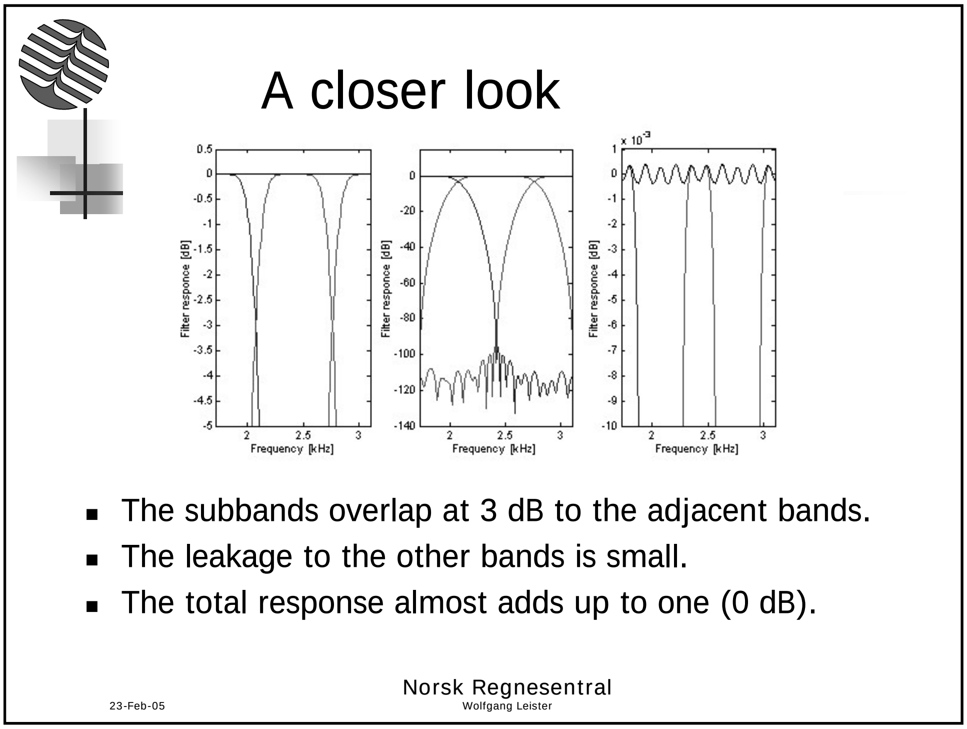
$$y_m[k] = x_m[k \cdot M]$$

Filterbank in MPEG-1 audio layer 1-3



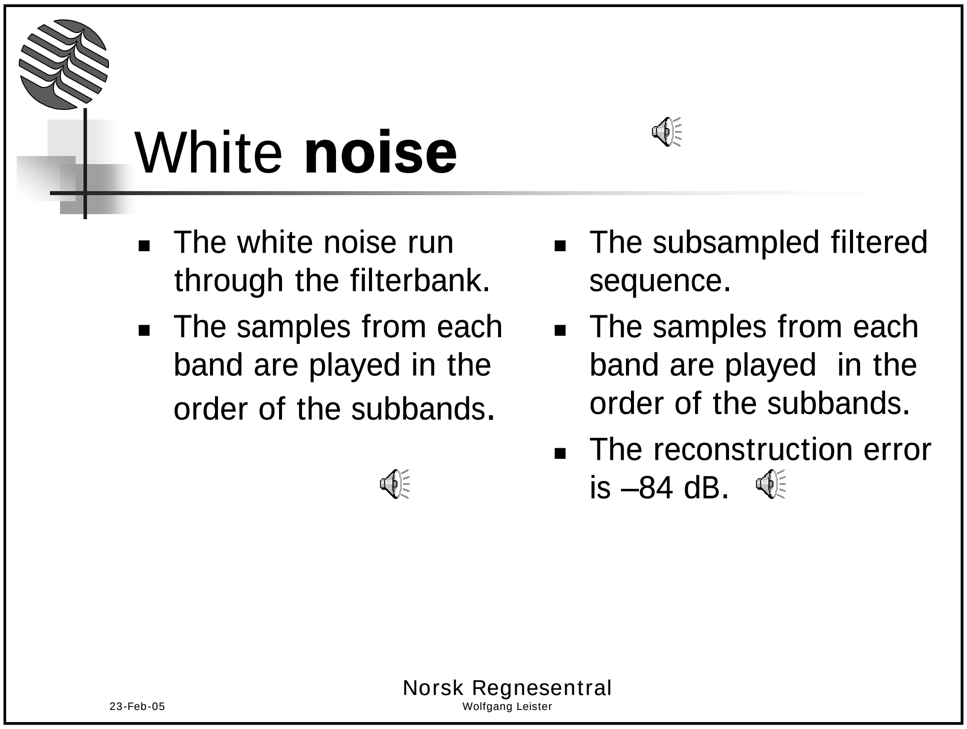
- Polyphase filterbank
- 32 subbands
- 512 tap FIR-filters
- 80 + and * per output

- Equal width
- Not perfect reconstruction
- Frequency overlap



A closer look

- The subbands overlap at 3 dB to the adjacent bands.
- The leakage to the other bands is small.
- The total response almost adds up to one (0 dB).



White **noise**

- The white noise run through the filterbank.
- The samples from each band are played in the order of the subbands.

- The subsampled filtered sequence.
- The samples from each band are played in the order of the subbands.
- The reconstruction error is –84 dB.

Nonideal filterbanks

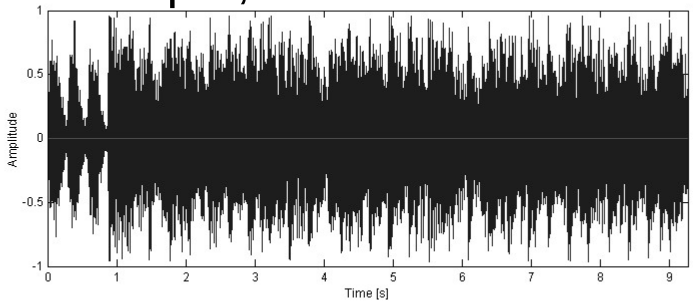
$$Y(e^{j\omega}) = X(e^{j\omega})\underbrace{\frac{1}{M}\sum_{k=0}^{M-1} H_k^R(e^{j\omega})H_k^A(e^{j\omega})}_{\approx 1} +$$

$$\sum_{n=1}^{M-1} X\left(e^{j\left(\omega - \frac{2\pi n}{M}\right)}\right)\underbrace{\frac{1}{M}\sum_{k=0}^{M-1} H_k^R(e^{j\omega})H_k^A\left(e^{j\left(\omega - \frac{2\pi n}{M}\right)}\right)}_{\approx 0}$$

- In a perfect filterbank the first part is the only part.

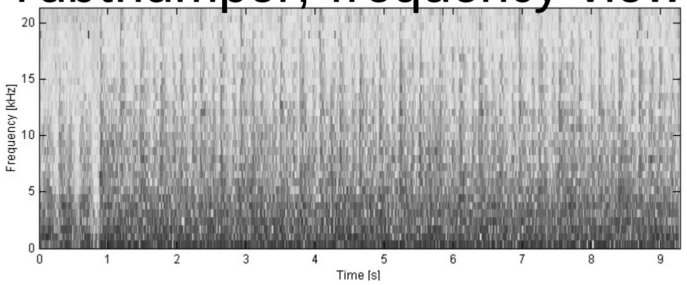
- The second part consists of the aliasing terms.
- The filterbank is designed so that the aliasing is small.

Tubthumper, a time domain view



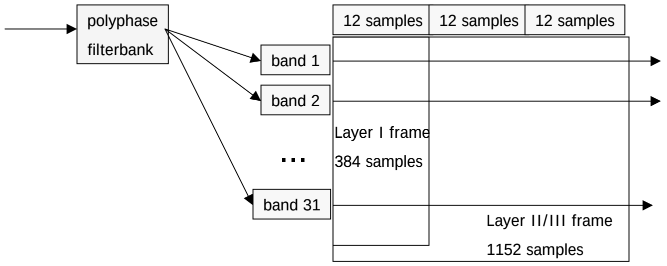
The red line is the reconstruction error after splitting the signal in subbands, down sampling and applying the synthesis filterbank. The reconstruction error is –84 dB and sounds like

Tubthumper, frequency view 🔈



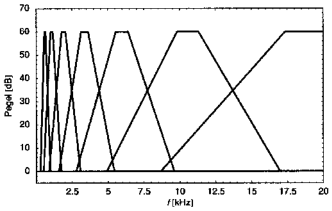
Subband	1	2	4	8	16	32
Center frequency [kHz]	0.3	1.0	2.4	5.2	10.7	21.7
No subsampling	🔈	🔈	🔈	🔈	🔈	🔈
Subsampled 32 times	🔈	🔈	🔈	🔈	🔈	🔈

Filterbank MPEG



Critical Bands

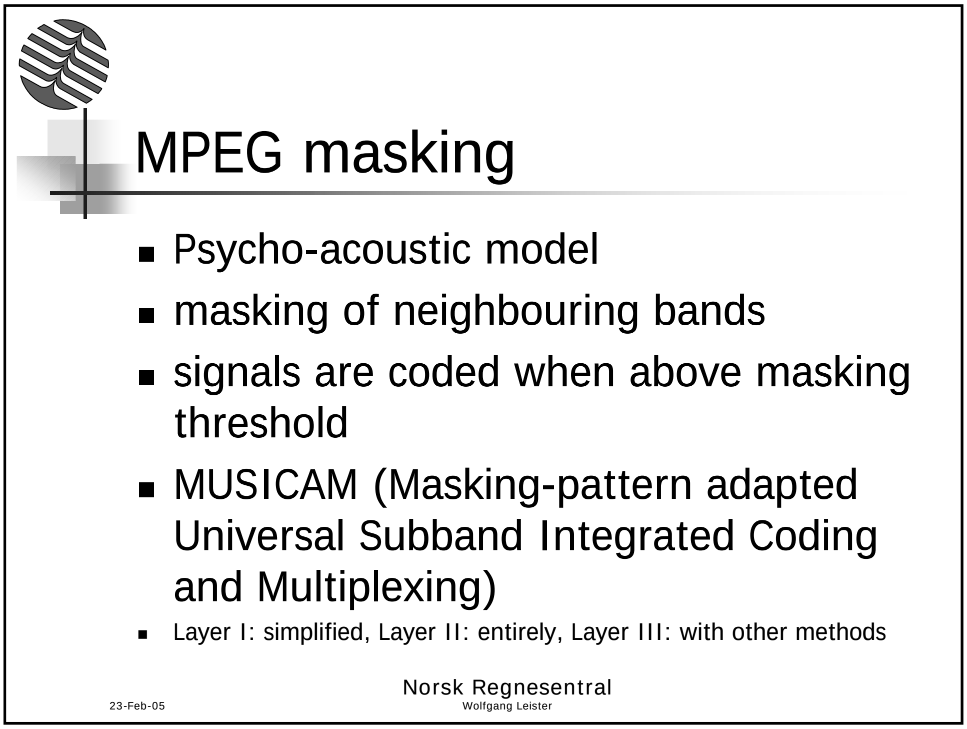
- Heinrich Barkhausen (1881-1956)
- psycho-acoustic
- width measured in bark



$$1\,bark = \begin{cases} f/100 & for\ f < 500 \\ 9 + 4 \cdot \log(f/1000) & else \end{cases}$$

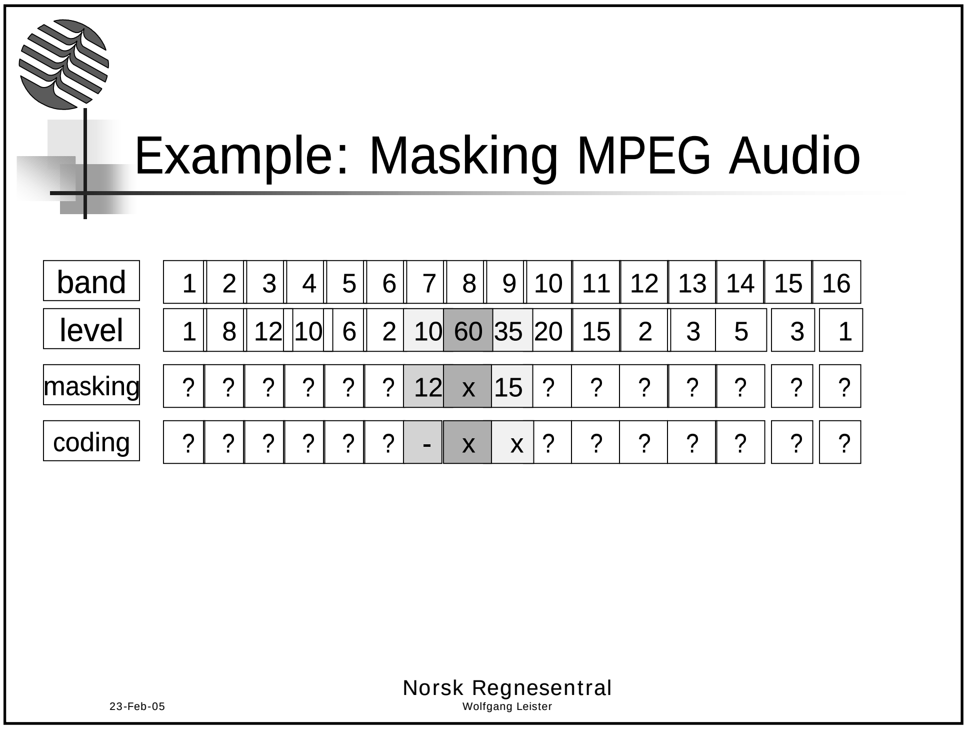
MPEG - Sub bands

- Layer I: 32 bands, 625 Hz each, Fourier transform
- Layer II: 32 bands, three frames, time masking
- Layer III: Division according to critical bands



MPEG masking

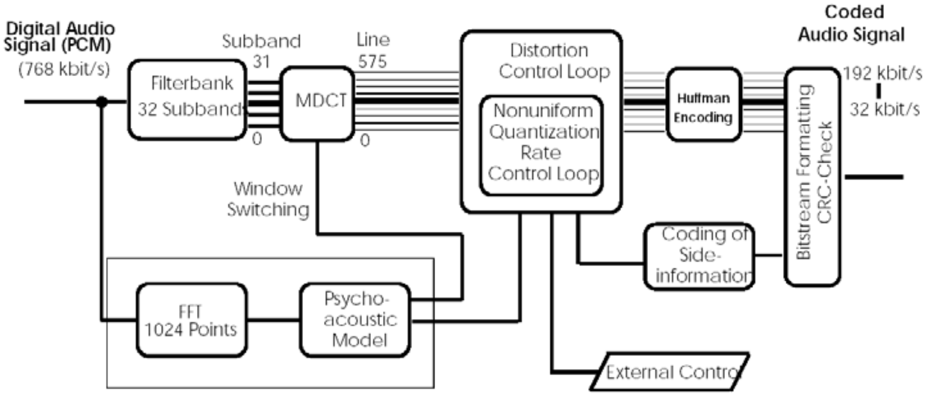
- Psycho-acoustic model
- masking of neighbouring bands
- signals are coded when above masking threshold
- MUSICAM (Masking-pattern adapted Universal Subband Integrated Coding and Multiplexing)
- Layer I: simplified, Layer II: entirely, Layer III: with other methods



Example: Masking MPEG Audio

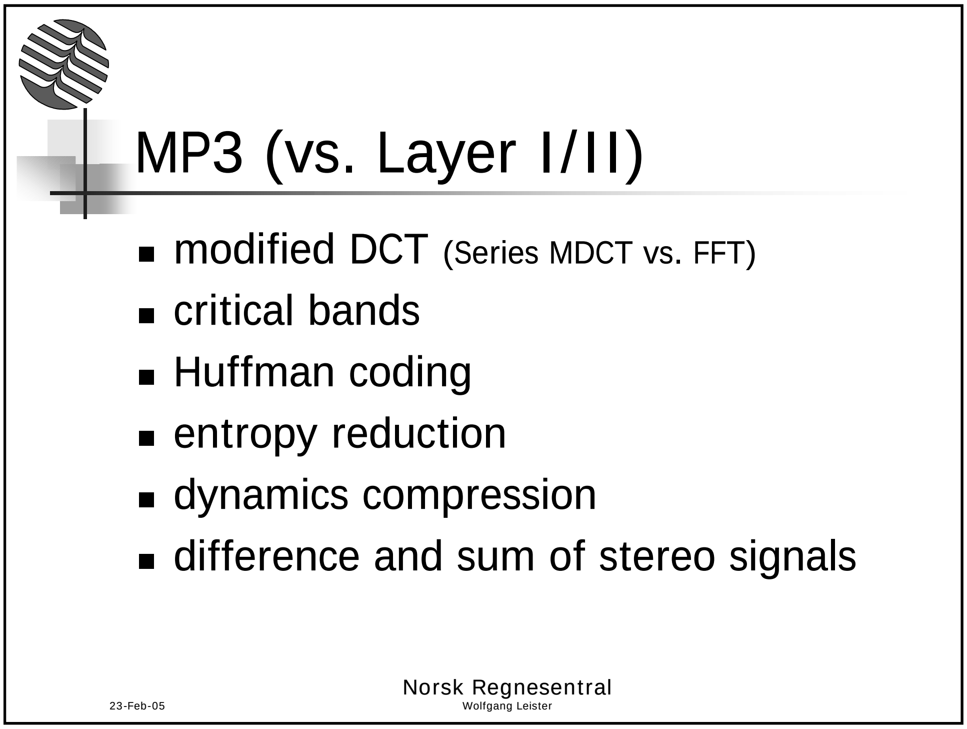
band	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
level	1	8	12	10	6	2	10	60	35	20	15	2	3	5	3	1
masking	?	?	?	?	?	?	12	x	15	?	?	?	?	?	?	?
coding	?	?	?	?	?	?	-	x	x	?	?	?	?	?	?	?

MPEG-1 Layer 3 encoder



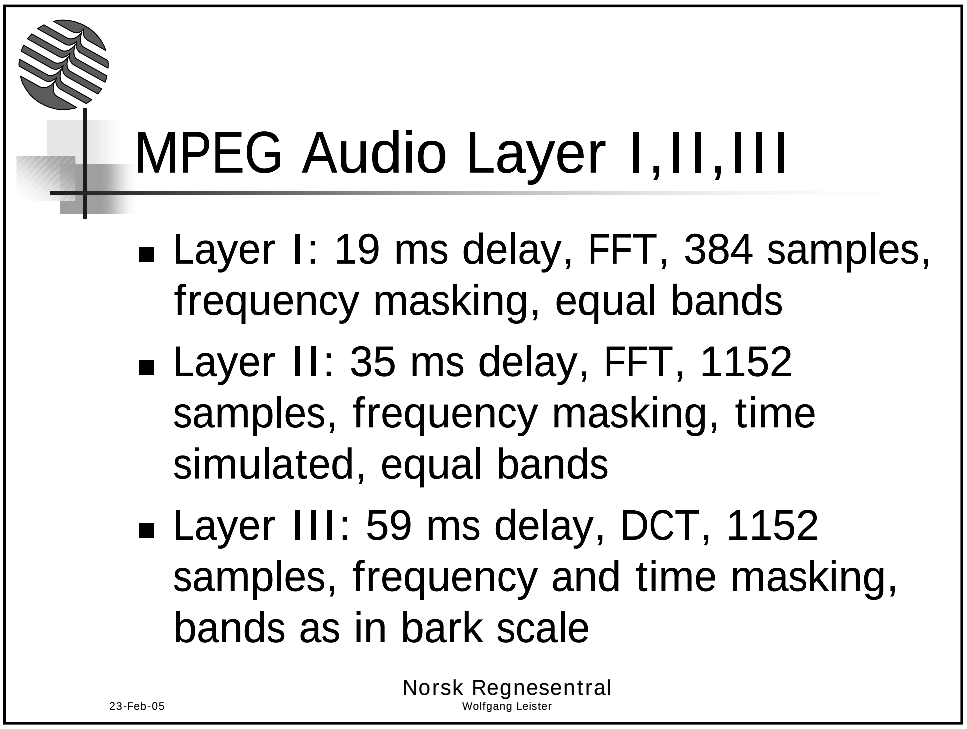
MP3

- Filter bank - sub bands
- Series MDCT
- fine grain frequency resolution
- non-uniform quantisation
- perception model
- Huffman coding



MP3 (vs. Layer I/II)

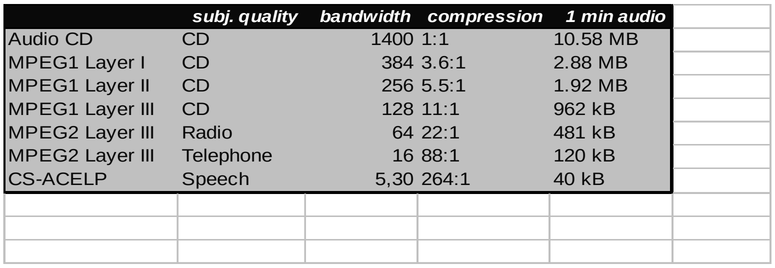
- modified DCT (Series MDCT vs. FFT)
- critical bands
- Huffman coding
- entropy reduction
- dynamics compression
- difference and sum of stereo signals



MPEG Audio Layer I,II,III

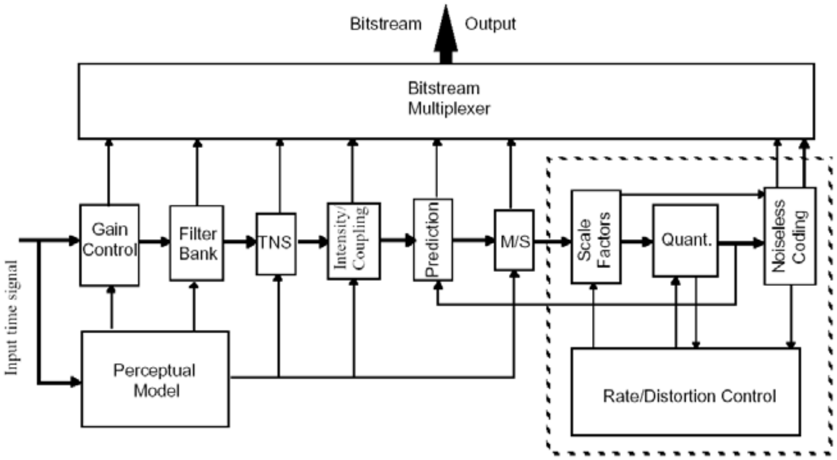
- Layer I: 19 ms delay, FFT, 384 samples, frequency masking, equal bands
- Layer II: 35 ms delay, FFT, 1152 samples, frequency masking, time simulated, equal bands
- Layer III: 59 ms delay, DCT, 1152 samples, frequency and time masking, bands as in bark scale

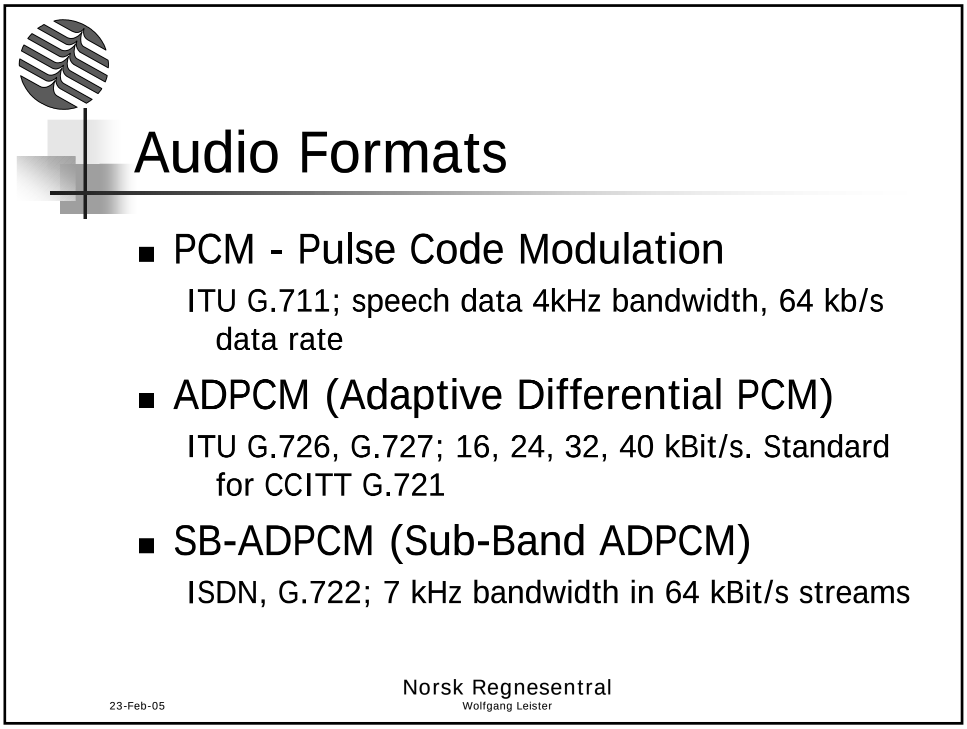
MPEG Layer I, II, III



	subj. quality	*bandwidth*	*compression*	*1 min audio*
Audio CD	CD	1400	1:1	10.58 MB
MPEG1 Layer I	CD	384	3.6:1	2.88 MB
MPEG1 Layer II	CD	256	5.5:1	1.92 MB
MPEG1 Layer III	CD	128	11:1	962 kB
MPEG2 Layer III	Radio	64	22:1	481 kB
MPEG2 Layer III	Telephone	16	88:1	120 kB
CS-ACELP	Speech	5,30	264:1	40 kB

MPEG-2 AAC





Audio Formats

- PCM - Pulse Code Modulation

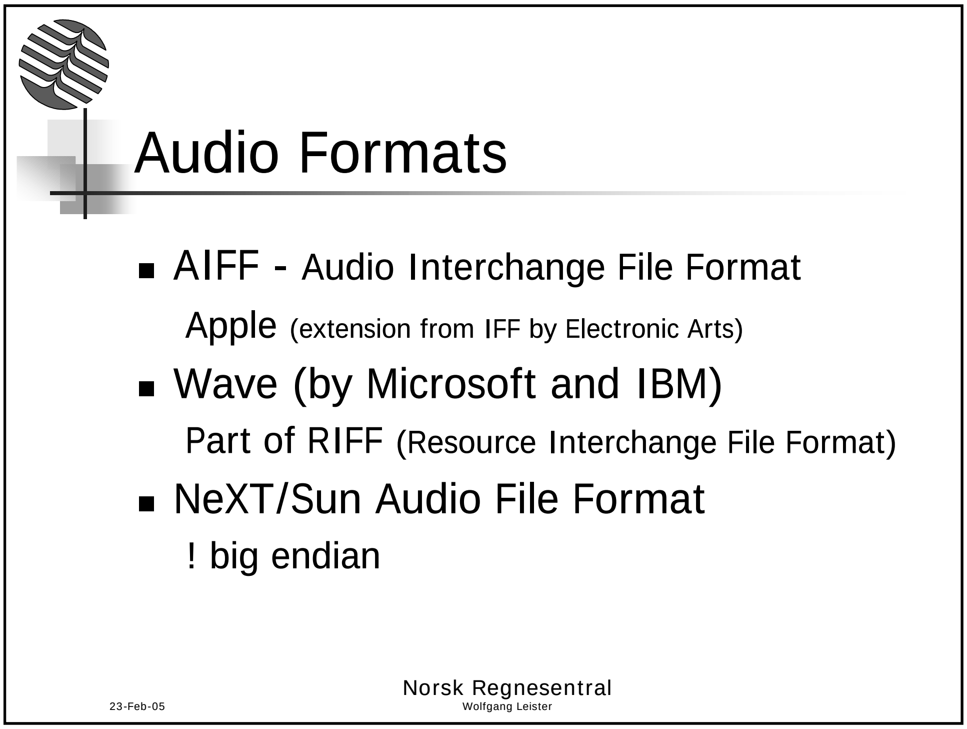
  ITU G.711; speech data 4kHz bandwidth, 64 kb/s data rate

- ADPCM (Adaptive Differential PCM)

  ITU G.726, G.727; 16, 24, 32, 40 kBit/s. Standard for CCITT G.721

- SB-ADPCM (Sub-Band ADPCM)

  ISDN, G.722; 7 kHz bandwidth in 64 kBit/s streams



Audio Formats

- AIFF - Audio Interchange File Format

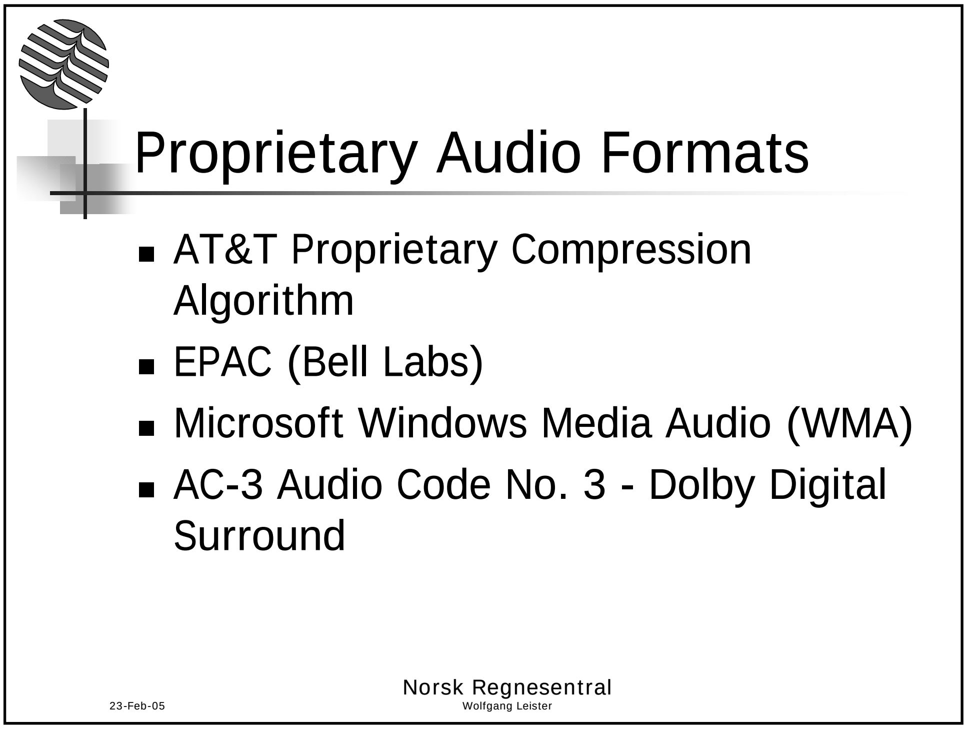
  Apple (extension from IFF by Electronic Arts)

- Wave (by Microsoft and IBM)

  Part of RIFF (Resource Interchange File Format)

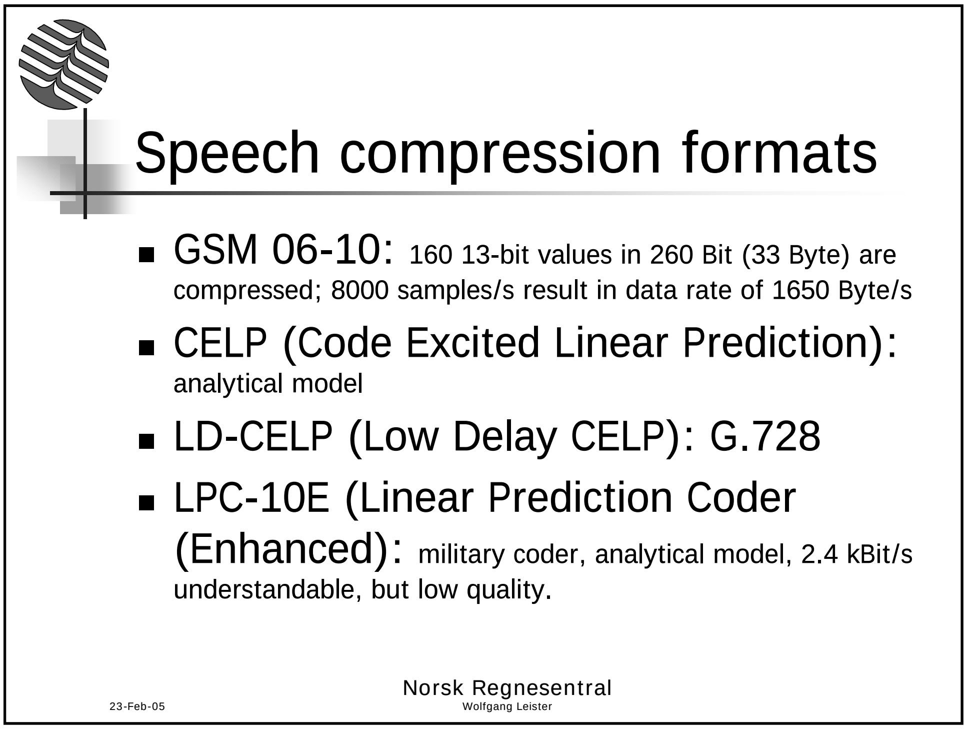
- NeXT/Sun Audio File Format

  ! big endian



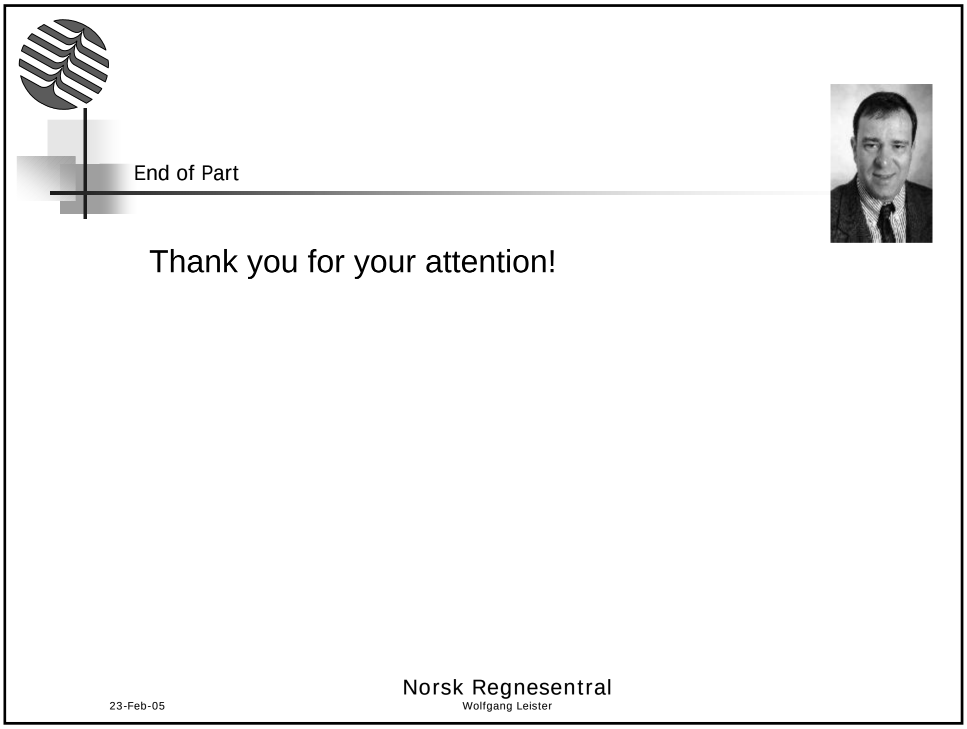
Proprietary Audio Formats

- AT&T Proprietary Compression Algorithm
- EPAC (Bell Labs)
- Microsoft Windows Media Audio (WMA)
- AC-3 Audio Code No. 3 - Dolby Digital Surround



Speech compression formats

- GSM 06-10: 160 13-bit values in 260 Bit (33 Byte) are compressed; 8000 samples/s result in data rate of 1650 Byte/s
- CELP (Code Excited Linear Prediction): analytical model
- LD-CELP (Low Delay CELP): G.728
- LPC-10E (Linear Prediction Coder (Enhanced): military coder, analytical model, 2.4 kBit/s understandable, but low quality.



End of Part

Thank you for your attention!