Compact Lecture

Multimedia Coding –
Methods & Applications

Part 2: Coding of Speech and Audio

Dr. Klaus Illgner

Overview

- **Motivation**
- **Introduction to speech coding**
- **Applications**
  - AMR (speech)
- **Introduction to audio coding**
  - Psycho acoustics
- **Some specific algorithms**
  - MP3 (Audio)
  - AAC (Audio) AACPlus
- **What next?**
  - Stereo
  - Dummy head
  - Suround sound 5.1 and more
  - Wave field synthesis

Motivation

In principle „sound signals“ should be simple to encode

- (D)PCM + entropy coding
- Transform coding

… or the other way round: „Why is audio coding a challenging problem?“

- Telephony in fixed networks → 32 …. 64 kbps
- Mobile voice communications → 2.4 kbps …. 9.6 kbps
- CD / SACD → 1.4 Mbps … 4.6 Mbps (stereo)

→ transmission capacity not sufficient (GSM)

→ very high quality demand (SACD, 5.1 surround sound)

The “sound world” splits for encoding into 2 quite different environments:

- Speech coding
- Audio coding

Fundamentals of Speech Coding

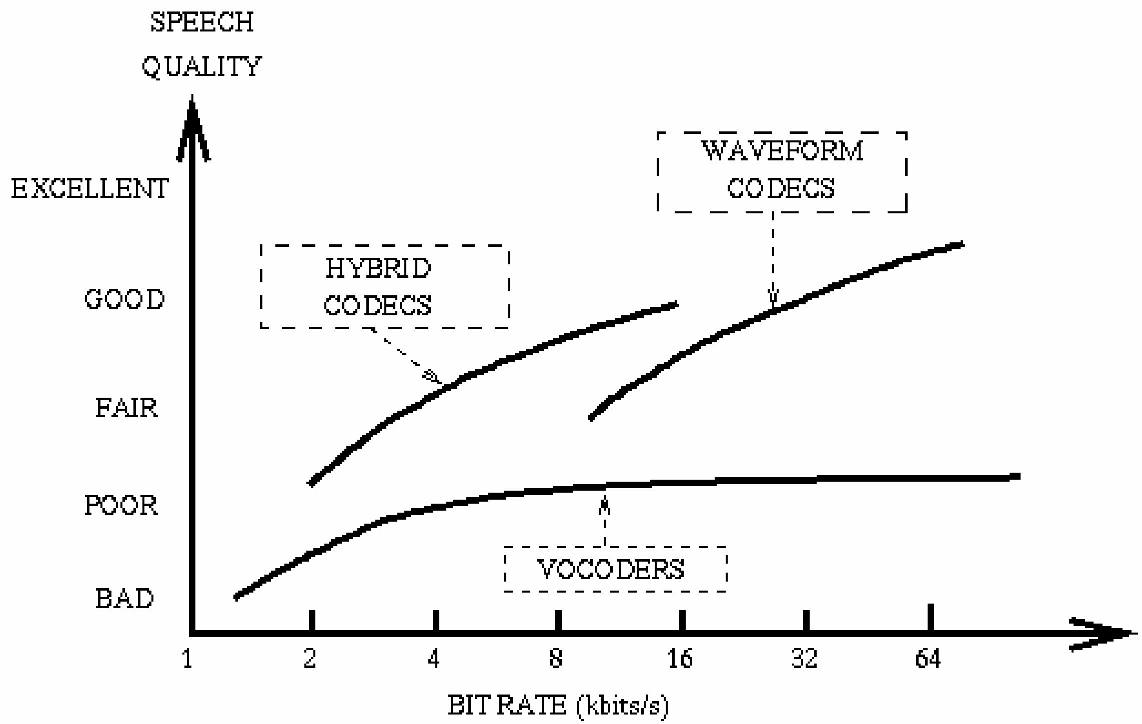
For coding speech is not audio

- **Traditional approaches regard speech as one dimensional non-stationary signal**
  - Waveform Coder
- **More recent approaches take the generation of human voice into account**
  - → Modeling the voice generation (transmitter oriented approach)
  - → Vocoder / LPC
- **Current codecs mostly follow an hybrid concept**
  - → CELP
- **Quality evaluation take place always on the basis of <u>subjective voice recognition tests</u>**

Overview of Standards for Speech Encoding

G.711	PCM for speech	8 kHz, sample	64 kbps	1972
G.721	ADPCM for speech → G.726	8 kHz, sample	32 kbps	1984
G.722	Wideband speech (SB-ADPCM)	16 kHz, sample	64 (56, 48) kbps	1988
G.722.1	Wideband speech	16 kHz, 20msec	24, 32 kbps	1999
G.723	Extension of G.721 → G.,726	8 kHz, sample	24, 40 kbps	1991
G.723.1	Dual rate speech codec (ACELP)	8 kHz, 30 msec	5.6 / 6.3 kbps	1995
G.726	ADPCM speech coding	8 kHz, sample	16, 24, 32, 40 kbps	1984
G.727	ADPCM mit 5, 4, 3, 2, bit	, sample	16, 24, 32, 40 kbps	
G.728	Low delay CELP	8 kHz,	16 kbps	1992
G.729	Conjugate structure CELP	8 kHz, 10 msec	8 kbps, 6.4 / 11.8 kbps	1995
GSM 06.10	Full rate (FR) speech coding (RPE-LTP)	8 kHz, 22.5 msec	13 kbps	1987
GSM 06.20	Half rate (HR) based on VSELP – vector sum excited prediction	8 kHz	5.6 kbps	1992
GSM 06.60	EFR – Algebraic CELP (ACELP)		12.2. kbps	1996
3GPP 126071	AMR (mandatory for UMTS) – based on ACELP		4.75 – 12.2 kbps (8 stufen)	1999
3GPP	AMR-WB → (G.722.2??); based on ACELP	16 kHz, 20 msec	6.6 – 23.85kbps (9 Stufen)	2001

Qualitative Comparison



Evaluation employs

- **Speech intelligibility test and**
- **Subjective listening tests under standardized conditions**

Speech Coding based on <u>Waveform Codecs</u>

For sufficient speech intelligibility

- Spectrum can be limited to 300- 3400 Hz (compare analogue telephone)

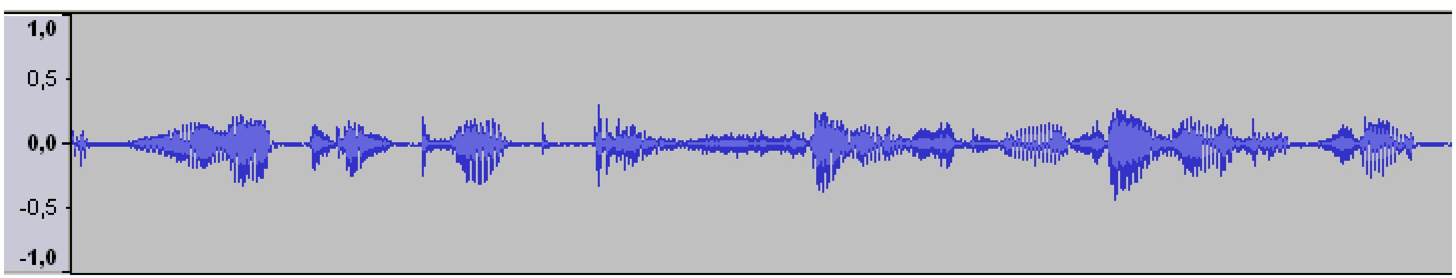
  → sampling rate of 8kHz

- Dynamic range of approximately 40 dB sufficient (theoretically)

  → 8 bit linear quantization (PCM)

  → data rate of 64 kbps (enables transmission of speech over a single 64 kbps ISDN-channel)

Approach: **PCM**: just sample and quantize the signal (8kHz, 8bit)



Challenge: signal characteristics is strongly non-stationary

PCM Approach (G.711 – 1972)

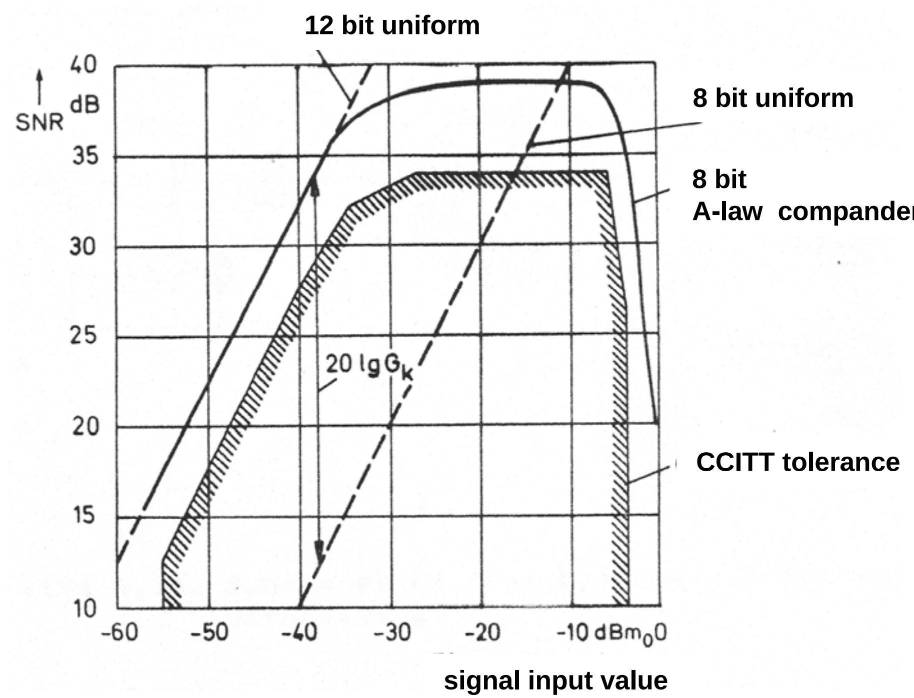
But: listening trials revealed an insufficient speech intelligibility

→ High dynamics of signal power

→ sampling (linear quantization)
requires 12bit

→ level dependent SNR
(→ has implications on intelligibility)

Solution:
non-linear quantization, implemented
via a compander (A-lax / u-law in US)



Bit assignment
1 bit -- sign
3 bit -- exponent
4 bit – amplitude

Resulting 8bit signal comparable to 12bit linear PCM signal

Specific Compander Curves

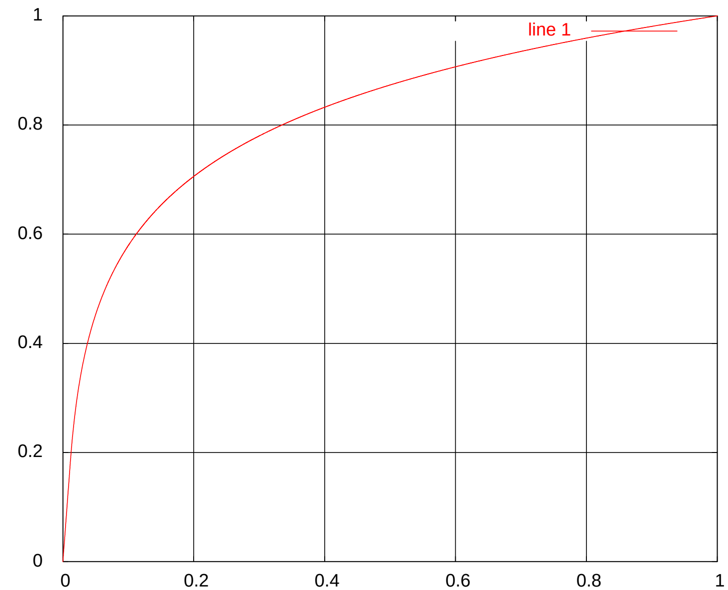
$$y = \frac{Ax}{1+\ln(A)}, \quad 0 \leq x < 1/A$$

$$y = \frac{1+\ln(Ax)}{1+\ln(A)}, \quad 1/A \leq x \leq 1$$

$$x = s/S_{max}, \quad A = 87,56$$

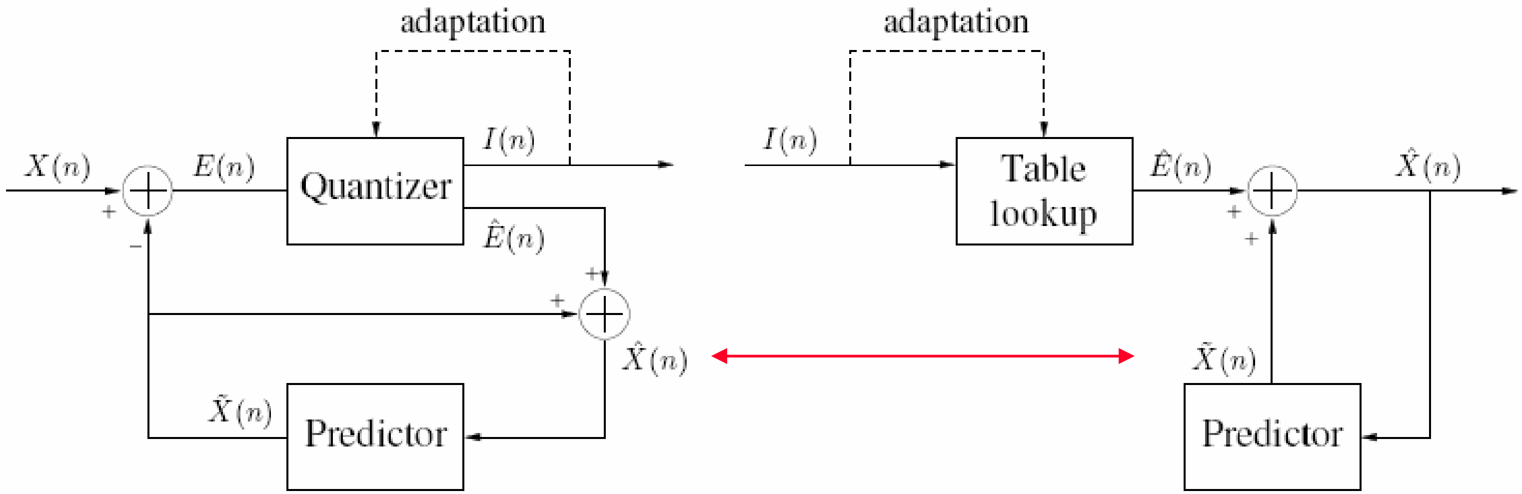
$$y = \mathrm{sgn}(x)\frac{\log(1+\mu|x|)}{\log(1+\mu)}, \quad x \leq 1$$

$$\mu = 100, 255$$

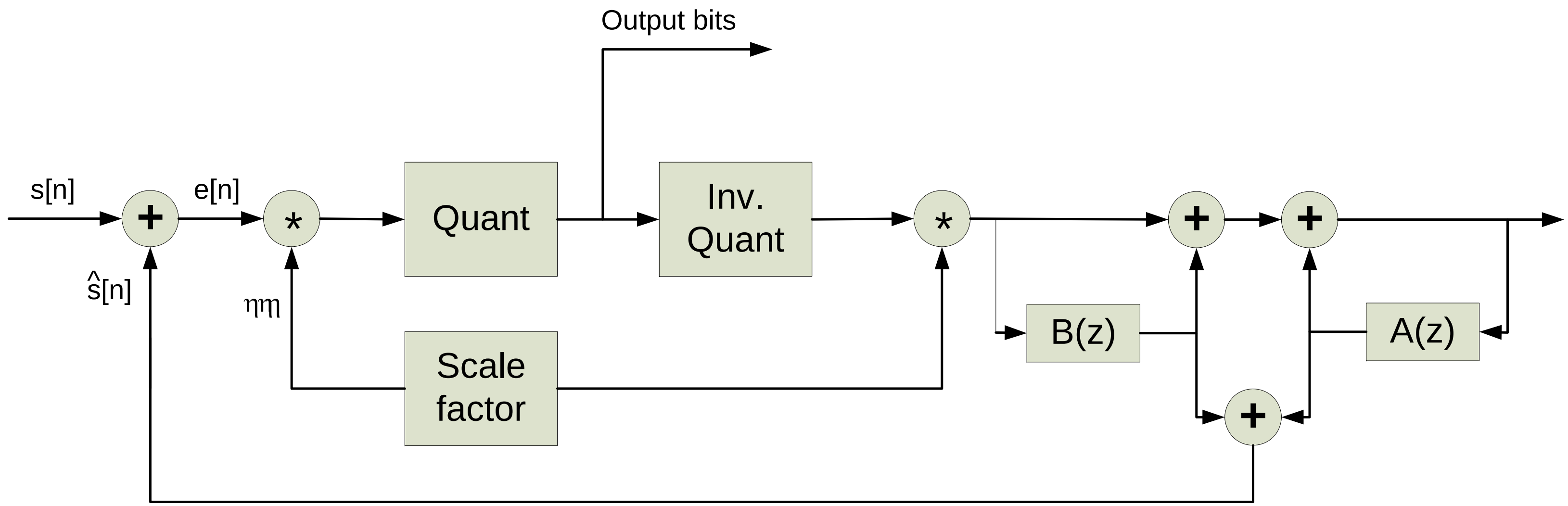


ADPCM → G.726 (G.721)

- **Input: u-law 16 bit PCM**
- **Processing of input signal in blocks**
  - Length: 20 ms → 160 samples at 8kHz sampling frequency
  - 4bit non-uniform quantizer (15 steps); adaptive
- **ADPCM (adaptive differential)**
  - G.721 → 32kbps (original), later extension for 26, 24, 40 kbps → now termed G.726
  - at 32kbps almost no difference to original
  - Based on backward prediction (no side information, prediction utilizes coded signal values)



ADPCM cont.



$$e[n] = (s[n] - \hat{s}[n])/\alpha$$

Adapting the fixed quantizer to varying input amplitudes

$$\alpha_n > \alpha_{n-1} \quad for \quad e[n] > e[n-1]$$

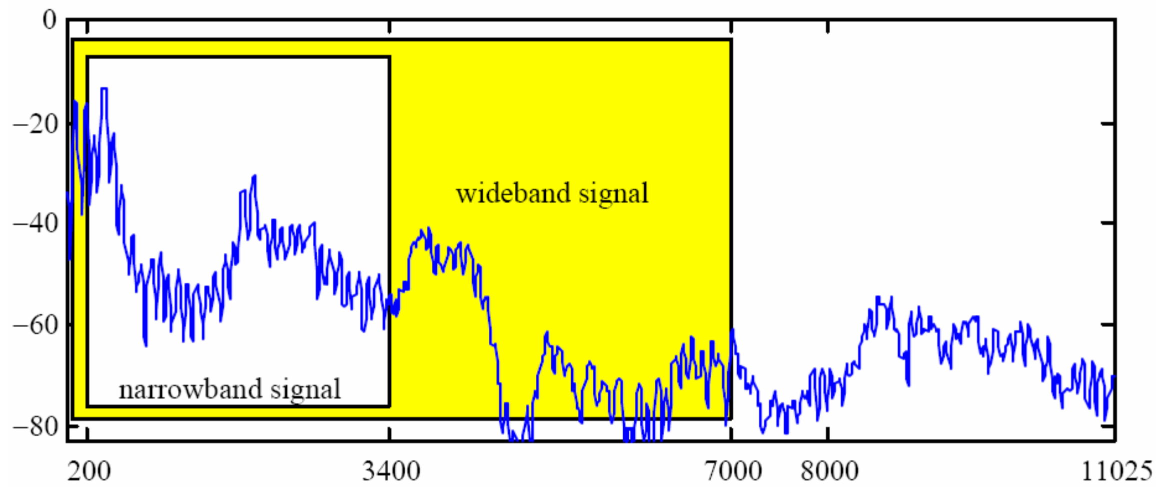
applied e.g. in the DECT-Standard

Filter:
A(z) → 2 tap FIR
B(z) → 6 tap FIR

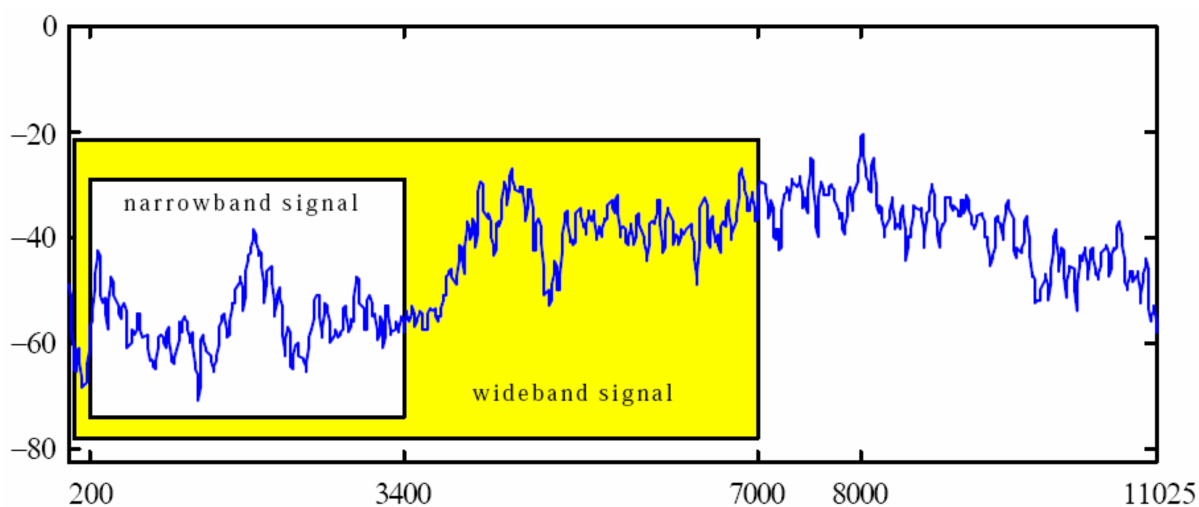
Variants:
IAM ADOCM, MS ADOCM

Are 8kHz Sampling Frequency Sufficient?

<u>voiced</u>
speech segment



<u>Un-voiced</u>
Speech segment



Source: VoiceAge

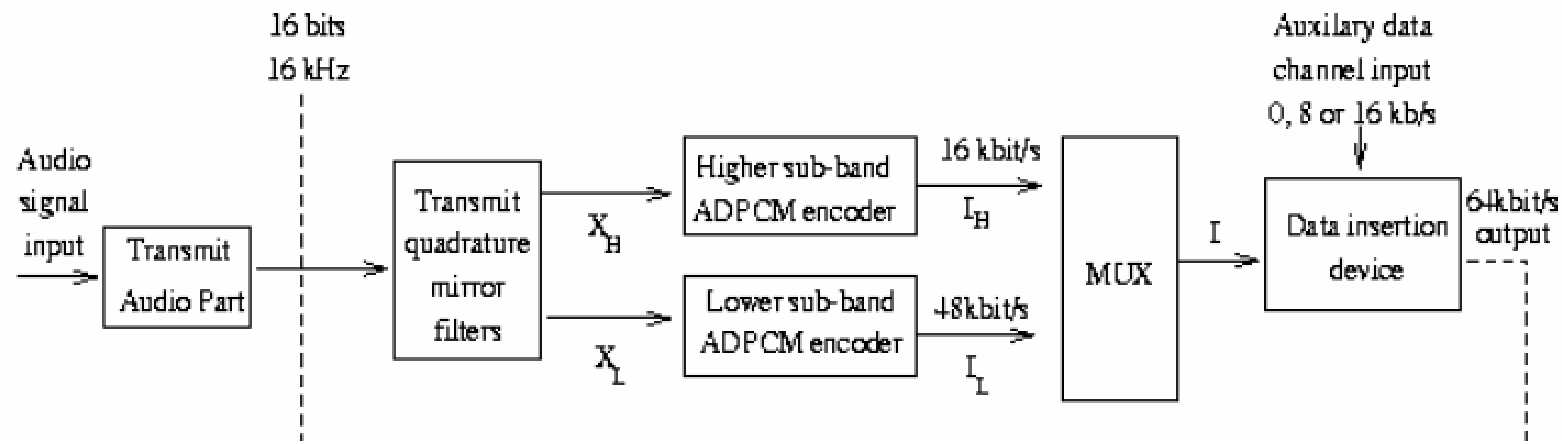
Wideband Speech – G.722

Improving the intelligibility by:

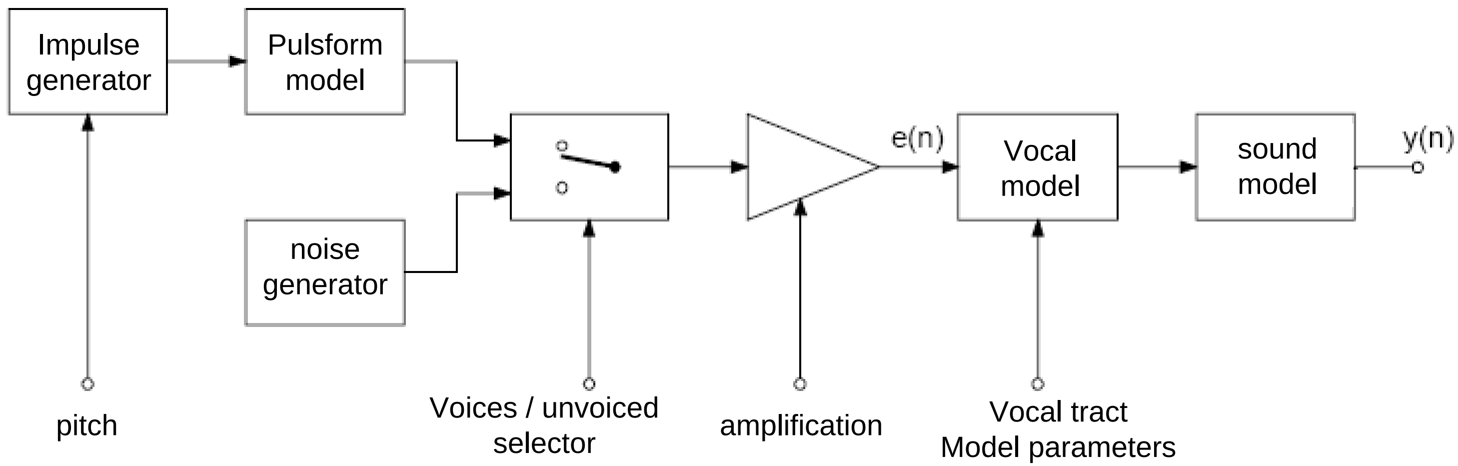
- Extending the transmitted signal spectrum to 50 Hz -- 7 kHz
- Increasing correspondingly the sampling rate to 16 kHz
- Required data rate results in 48, 56, 64 kbps (LB: 4 bit/sample, 5 bit, 6 bit)

Processing approach:

- Decomposition with QMF filter bank (24 coefficients → 3 ms delay)
  - channel 1 (LB): 0-4 kHz   channel 2 (HB): 4-7 kHz
- Separately encoded with ADPCM algorithm (LB: 6bit/sample; HB: 2 bit/sample)



Model of Human Speech Generation

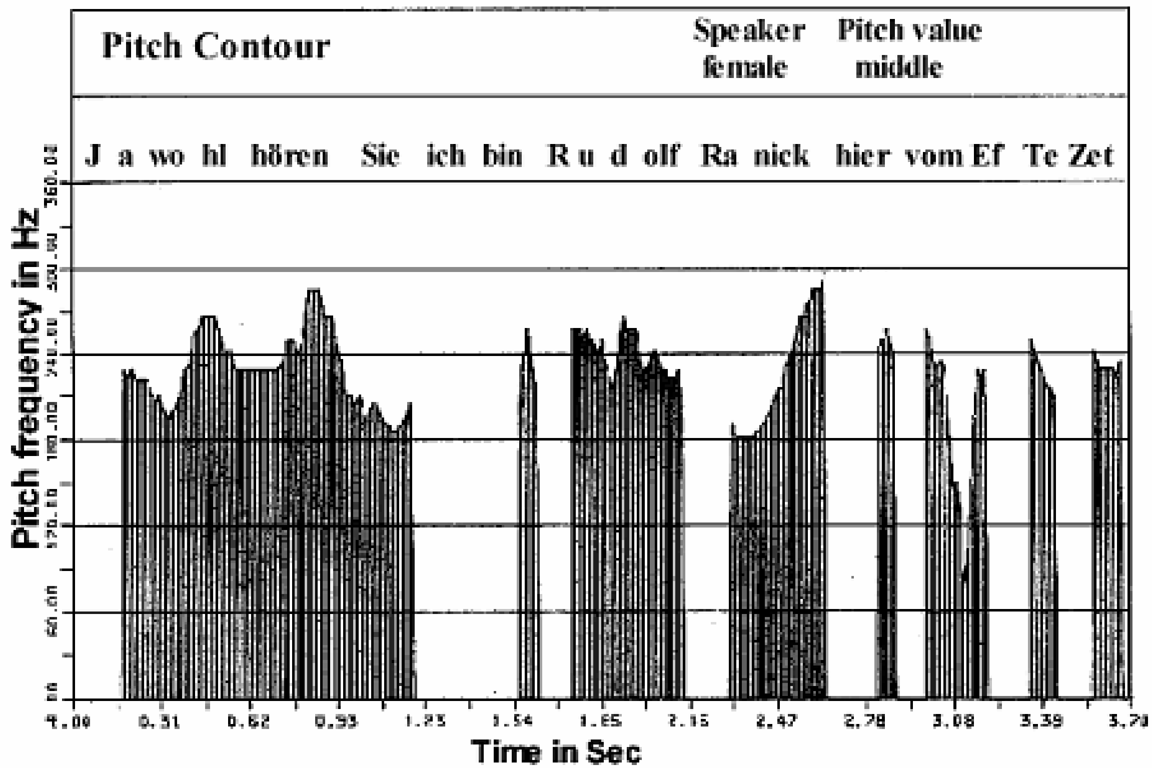


Model of the vocal tract:

→ tube model (tubes of different diameter connected to each other)

→ Filter models resonance characteristics of mouth, throat, and nose cavity

Pitch



Source: http://www.kt.tu-cottbus.de/speech-analysis/

Different approaches for estimating the pitch frequency
e.g. based on the auto-correlation function

Model-based Coding

Takes the human vocal tract as a reference → creating a model

Estimating and coding the model parameters

- Voiced sounds
  - Characterized as „pitch“ → excitation of the vocal cords with the fundamental frequency
  - men: 50 – 250 Hz, women 120 – 500 Hz
- Unvoiced sounds:
  - Excitation with noise
- Formant (sound model):
  - Maximum of the envelop of the power density spectrum
  - 3 lowest formant are always in the range of 300- 3400 Hz
  - <u>Sufficient for characterizing speech</u>

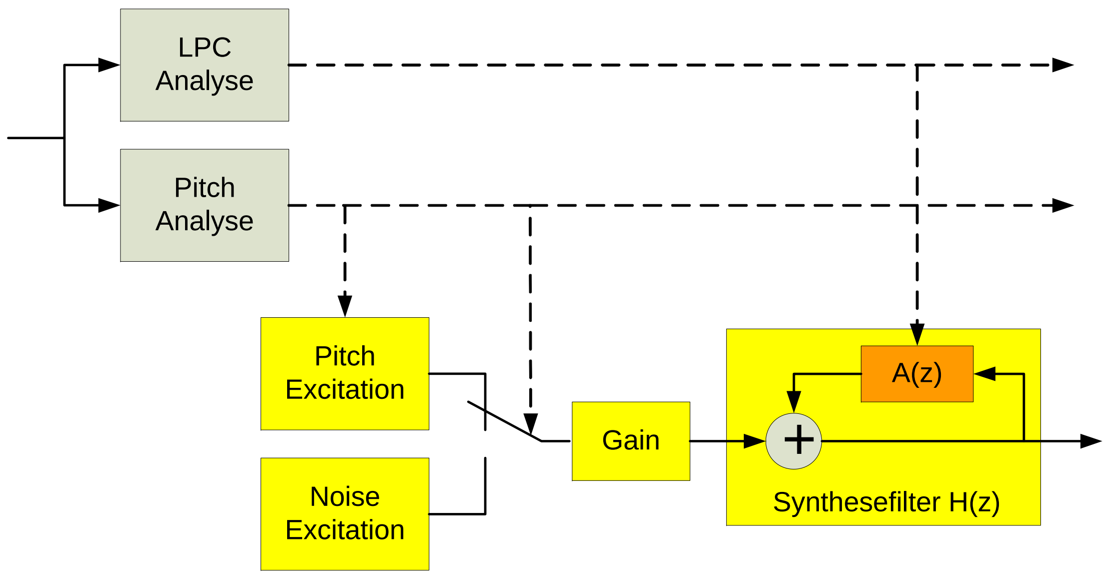
Realization example: Vocoder

- Time varying filter, excited with pulse trains (pitch) or with noise
- Updating the pitch every 10-20 ms
- Works down to about 2.4 kbps (applied primarily for very low rate coding of <0.5 bit /sample)

- **Advantage: speech intelligibility even at very low bit rates**
- **drawback: speech does not sound naturally, higher rates improve quality only marginally**

LPC-Vocoder

LPC -- Linear predictive coding
→ Estimating the similarity of neighbored samples (Short term prediction)



Crucial is the LPC analysis and the design of the synthesis filter

LPC

Excitation parameters are typically <u>not constant</u>

→ block-based processing assuming constant parameters within a block

→ block length typically 10—30 ms (standards: 20 msec)

Calculating the parameters of the synthesis filter H(z) via the prediction filter A(z)

$$A(z) = 1 - \sum_i a_i z^{-i}$$

$$\hat{x}[n] = \sum_{i=1}^{p} a_i x[n-i]$$

Prediction error filter → linear FIR filter, typically of order p=10
Calculating the coefficients, e.g. using the MSE criteria

Synthesis filter is the inverse prediction filter (Careful with roots)
Outputs the synthesized speech signal
when excited with white noise and pitch

$$H(z) = \frac{1}{A(z)}$$

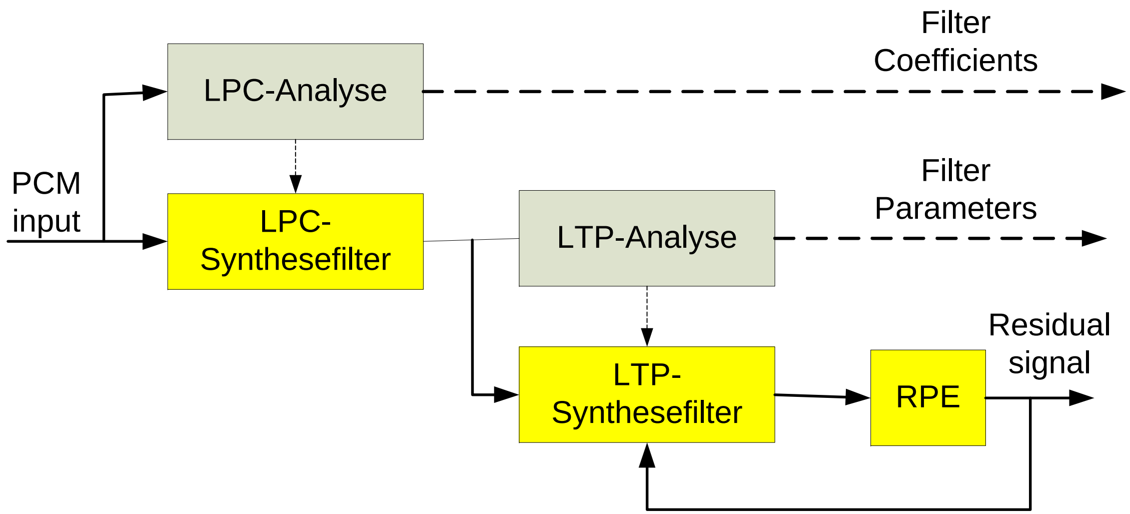
Speech Coding in GSM → RELP

Characteristics:

- Sampling frequency 8kHz
- 13 bit linear
- Partitioning in frames 20 ms → 160 samples per frame

Algorithmic features:

- RPE (residual pulse excitation) / LTP – LPC coding → 13 kbps (260bit / frame)
- Detecting speech pause and fill it with comfort noise
- Adding forward error correction → adds up to 22,8 kbps



System Diagram GSM Speech Coder

- **LPC: linear prediction, 8 Tap filter in GSM → 1.8 kpbs**
  - Modeling the vocal and nasal tract
  - Levinson-Durban algorithm for calculating the coefficients
  - Coefficients have the meaning of reflection coefficients (pipe model)
  - Excitation by model signals voiced / unvoiced
  - 36 bit / 160 samples; logarithmic quantizer (6, 6, 5, 5, 4, 4, 3, 3) bit
- **LTP : Long term prediction → 1.8 kpbs**
  - Excitation by RPE – residual pulse excitation
  - Output: model signals voiced / unvoiced
  - 4 x 40sample frames; calculating time shift N0 and amplification b
  - calculate RPE(block n-1[n-N0}) * b and calculate difference; transmit N0, b with 2+7 bit
- **RPE: Residual Pulse Excitation**
  - Linear low pass filter (FIR) of order 10
  - Decomposition in 3 polyphase bands
  - Encode subband with most significant energy
  - Select coefficient with maximal energy for normalization
  - Linear quantization of 13 coefficients with 3 bit per coefficient

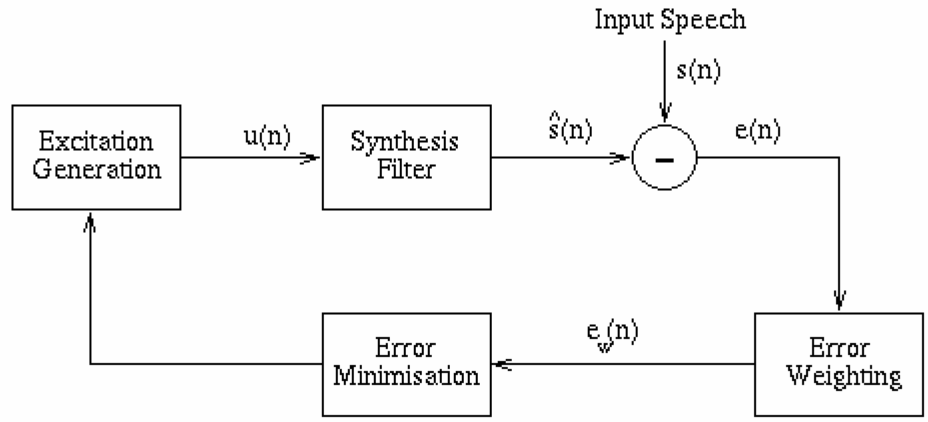
Hybrid Coder

combination of vocoder and waveform coder

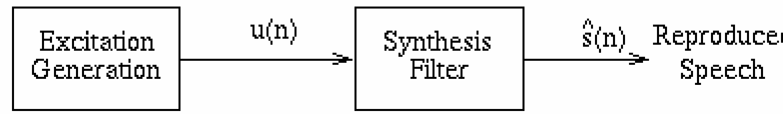
- **System commonly referred to: „Analysis by Synthesis“-Coder**
  - Synthesis = LPC Vocoder
  - Analysis filter

- **CELP (Code Excited linear Prediction)**
  - 4 – 24 kbps
  - 15 – 45 ms delay

- **Advanced developments**
  - Regular Pulse Excited CELP



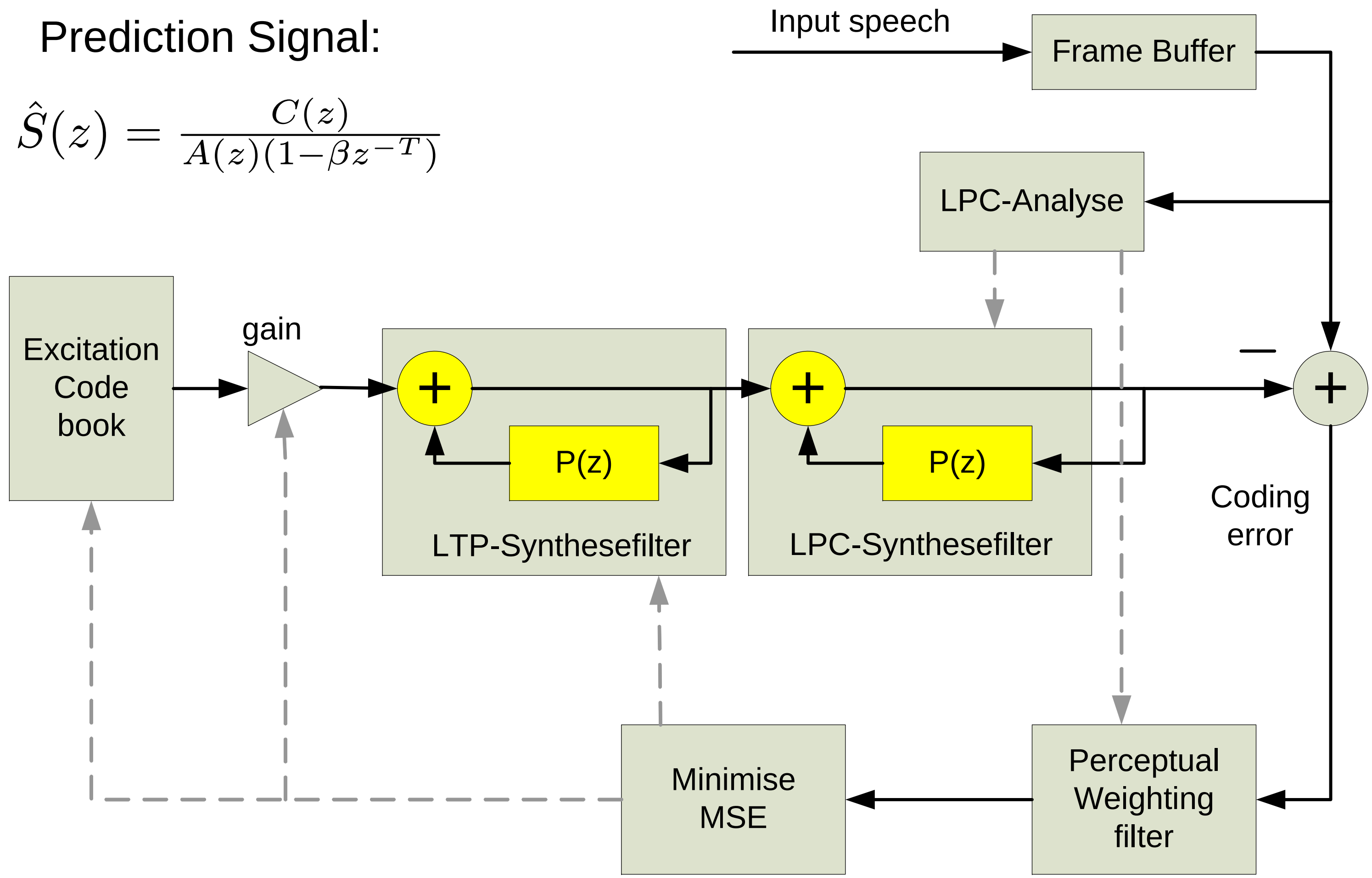
Encoder



CELP System Diagram

Prediction Signal:

$$\hat{S}(z) = \frac{C(z)}{A(z)(1-\beta z^{-T})}$$



Excitation Code Book

- **Modeling the error signal (Excitation Signal)**
  - MPE (multi-pulse excitation):
    exciting the filter with pulse train of variable amplitude and frequency
  - RPE (regular pulse excitation)
    fixed distance between pulses – variable amplitude and position of first pulse
  - CELP (code excited linear predictive)
    vector code book with excitation signals
    e.g. 10 bit for index + 5 bit for amplitude → 15 bit compared to 47 bit for GSM RPE

CELP

Each entry consists of 60 samples (→ 7.5 msec) [30 msec frames]

- **Adaptive Codebook (long term prediction)**
  - Delayed versions of earlier excitation signals, multiplied by an amplification value
- **Statistical code book**
  - 1092 random values {-1, 0, 1}
  - Start reading out from this “vector” from the position indicated by the (transmitted) index k
  - distance is 2 x k → 9bit

CELP: Additional Components

- **LTP: Long term prediction**
  - Estimating the pitch frequency
  - Exploiting longer correlations

$$e[n] = \beta e[n - T]$$

- **„Noise Shaping“ in the filter „perceptual error weighting“**
  - Moving the error to formants of higher energy

$$W(z) = \frac{A(z/\gamma_1)}{A(z - \gamma_2}$$

$$\gamma_1 \approx 0.9, \quad \gamma_2 \approx 0.6$$

G.728 Low Delay CELP (1992)

Characteristics:

5 msec delay – results from small block sizes (5 samples)

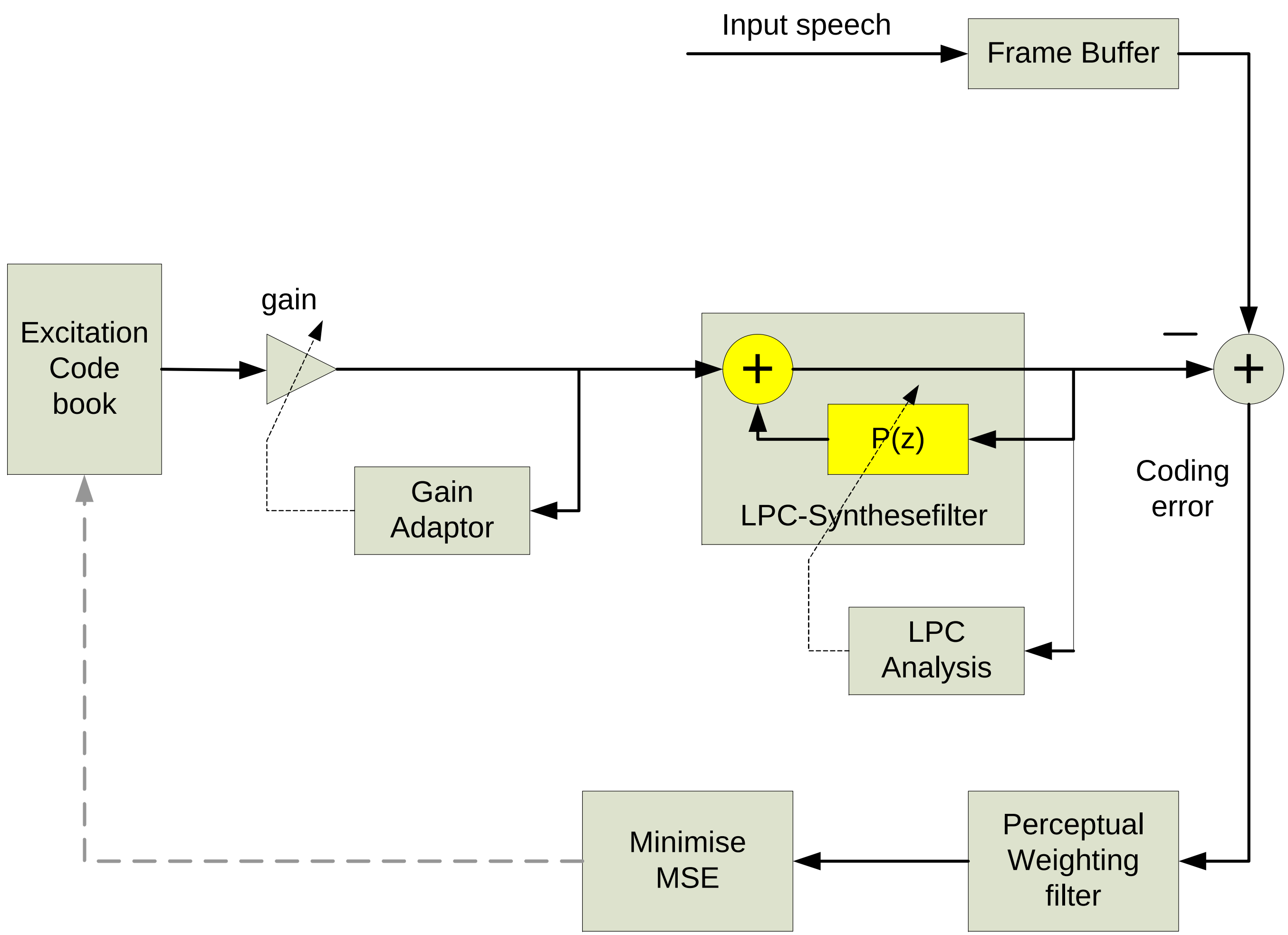
8 kHz sampling rate (16 bit linear PCM)

speech quality comparable to 32 kbps ADPCM

Algorithm: Backward adaptive CELP

- Backward LPC analysis with LPC filter of order 50; updating coefficients every 2,5 msec → short term predictor only
- Backward adaptive linear prediction → low delay
- Backward controlled gain-scaled vector quantization for excitation signal
- AbS- code book search of CELP coder
- Adaptive post-filter
- Excitation vector composed of 5 samples
- Decoder contains “post”-filter to improve quality

LD-CELP System Diagram



GSM AMR (Adaptive Multirate)

Motivation:

Adapting the code rate according to the channel conditions (error rate)

Characteristics:

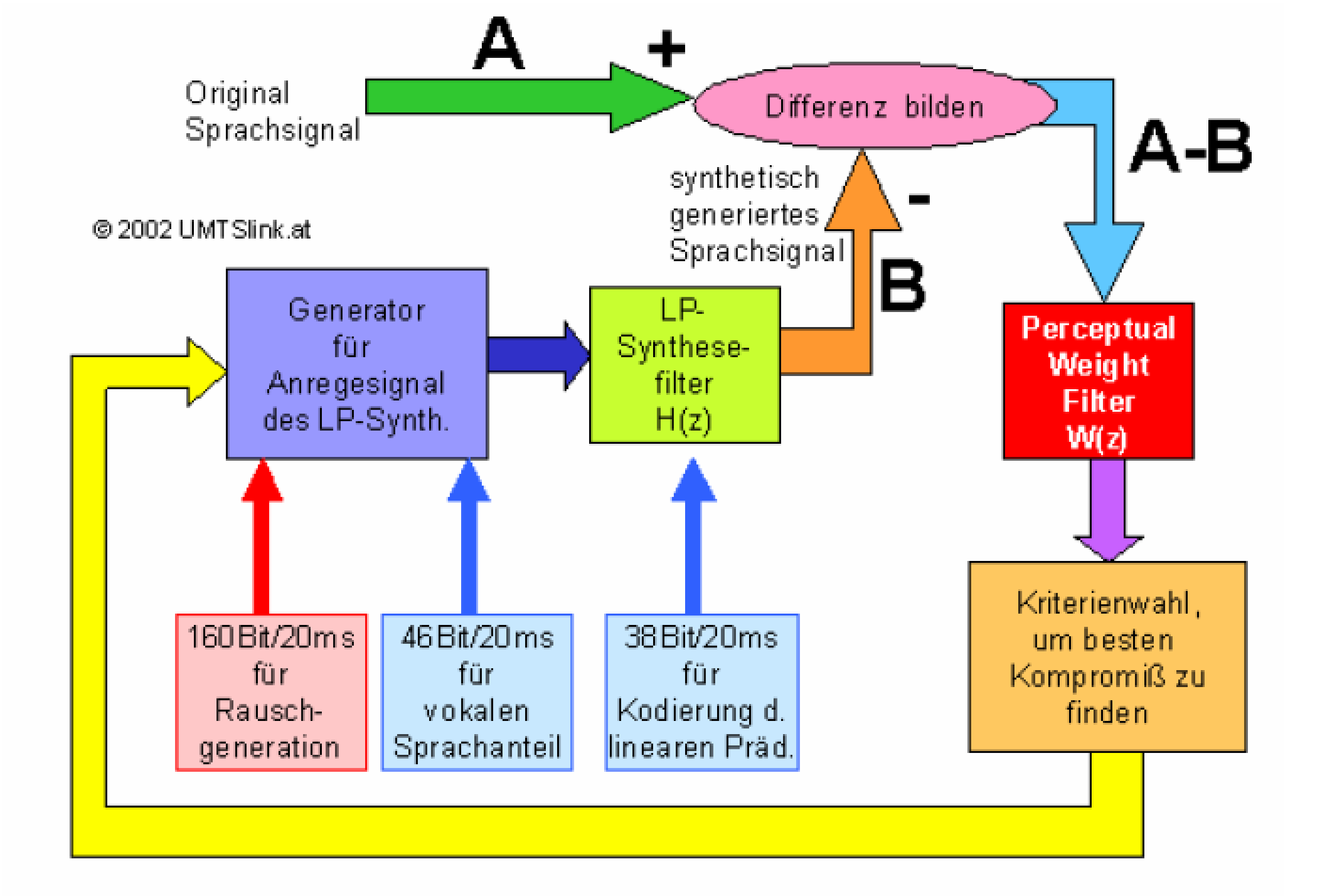
- Grounds on the ACELP (10tap LPC-Filter)
- 20 ms frames
- Open-loop pitch estimation every 10 msec
- 2 code books → every 5 msec recalculation and updating the code parameters
  - Fixed code book (FCB)
  - Adaptive code book (ACB)
- Code rates: 4,75 / 5,15 / 5.90 / 6,70 / 7,40 / 7,95 / 10,20 / 12,20 kbps

Exploits also

- Voice activity detection (VAD)
- Discontinuous transmission (DTX)

→ stopping the transmission in speech breaks

AMR System Diagram



Structure identical to CELP

3GPP AMR-WB → ITU-T G.722.2

Codecs takes into account higher naturalness of high bandwidth speech

Characteristics:

- 6.6 – 23.85 kbps
- 20 msec frames → 320 samples / frame
- Calculating the inner excitation parameters every 5 msec (4x per speech frame)
- ACELP (Algebraic Code Excitation Linear prediction) → similar to G.729, GSM EFR
- 2 frequency bands: 50- 6400 / 6400 – 7000
- LPC; LTP, samples at 12,8 kHz
- LB: ACELP
- HB: coding based on low band LB → calculation of a gain factors (in the encoder)
  decoder utilizes a 16kHz random excitation signal, applying a synthesis filter with parameters derived from the LB-signal

Components

- Discontinuous transmission (DTX)
- Voice activity detection (VAD)
- Comfort noise generation (CNG)

Applications

- Utilization of the same codec for wireline and wireless communications (no transcoding at transition gateway)
- Very robust against transmission errors (multirate → adaptive bit assignment for source and channel coding)
- Not suitable for music

3GPP AMR-WB+

Suitable for speech <u>and</u> audio

(competition to AAC+ primarily in the low bit rate range

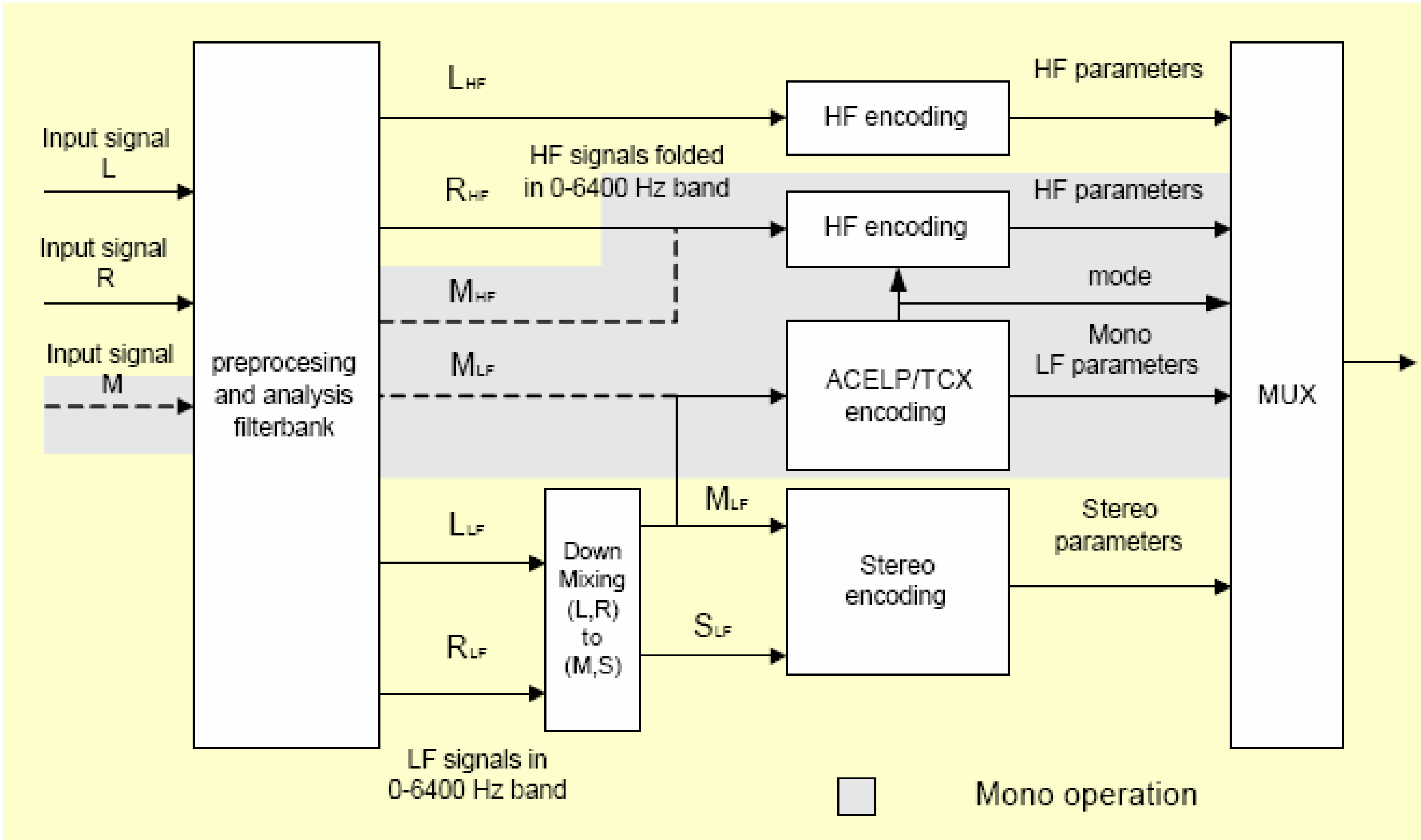
features:

- Sampling rates: 16, 24, 48 kHz
- Code rates: 9.6 – 23.2 kbps

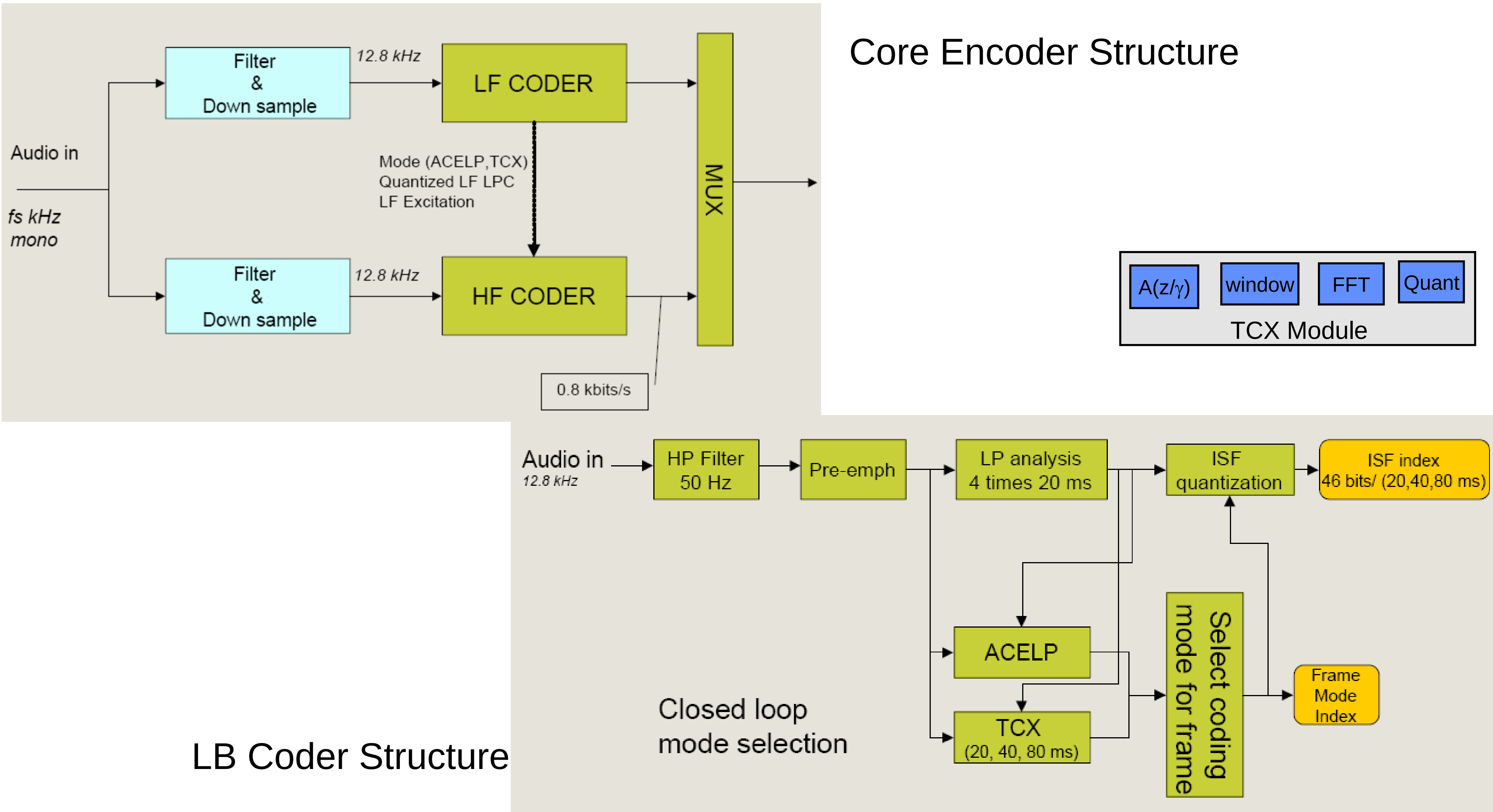
Tools:

- Multimode ACELP (AMR-WB),
- TCX (Transform Coded Excitation) → Coding of audio signals
- Coding additional band with higher spectral components (HF-band extension)
- Over-clocking

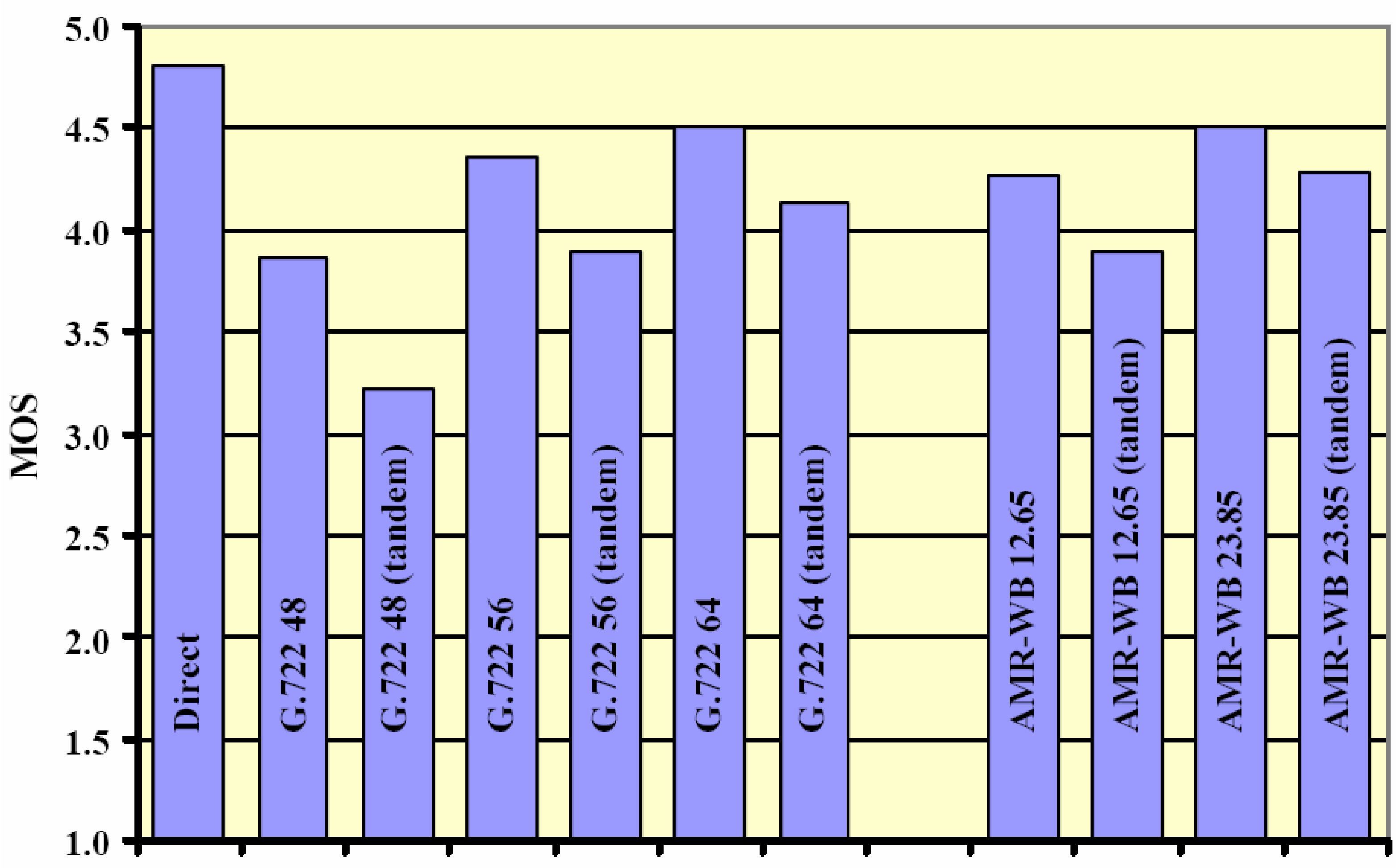
Encoder Architecture



Encoder Details



Comparison of Speech Coding Schemes



Source: VoiceAge

Fundamentals of Audio Coding

Demanding an unaltered and original listening experience

- Relevant spectral range: 20-20kHz
- Quantization with 16bit / sample

  → increased dynamic range (96 dB) (assuming a uniform PDF)

CD „coding: 44.1 kHz, stereo, 16bit / sample

- reference since its introduction 1988
- extensions: → SACD: 24bit, 96kHz sampling frequency

When talking about audio coding, one needs to differentiate between

- Coding scheme
- File format
- player software for playback

Relevant Audio Coding Schemes

Common Audio Coding standards:

- PCM as used for CD / SACD
- MPEG-1 – ISO/ IEC 11172-3 (Nov 1992)
- MPEG-2 – ISO/IEC 13818-3 (Nov 1994, 1997)
- MPEG-4 AAC
- AACPlus, HE-AAC
- Dolby Digital (formerly known as AC3)
- DTS (Digital Theatre System)
- AC97
- ATRAC (Sony)

Related Terms:

- 5.1 Multichannel
- DOLBY – noise reduction versus Dolby Coding
- Dolby E – production format for surround sound signals

Psycho-Acoustics

To improve coding efficiency without sacrificing quality

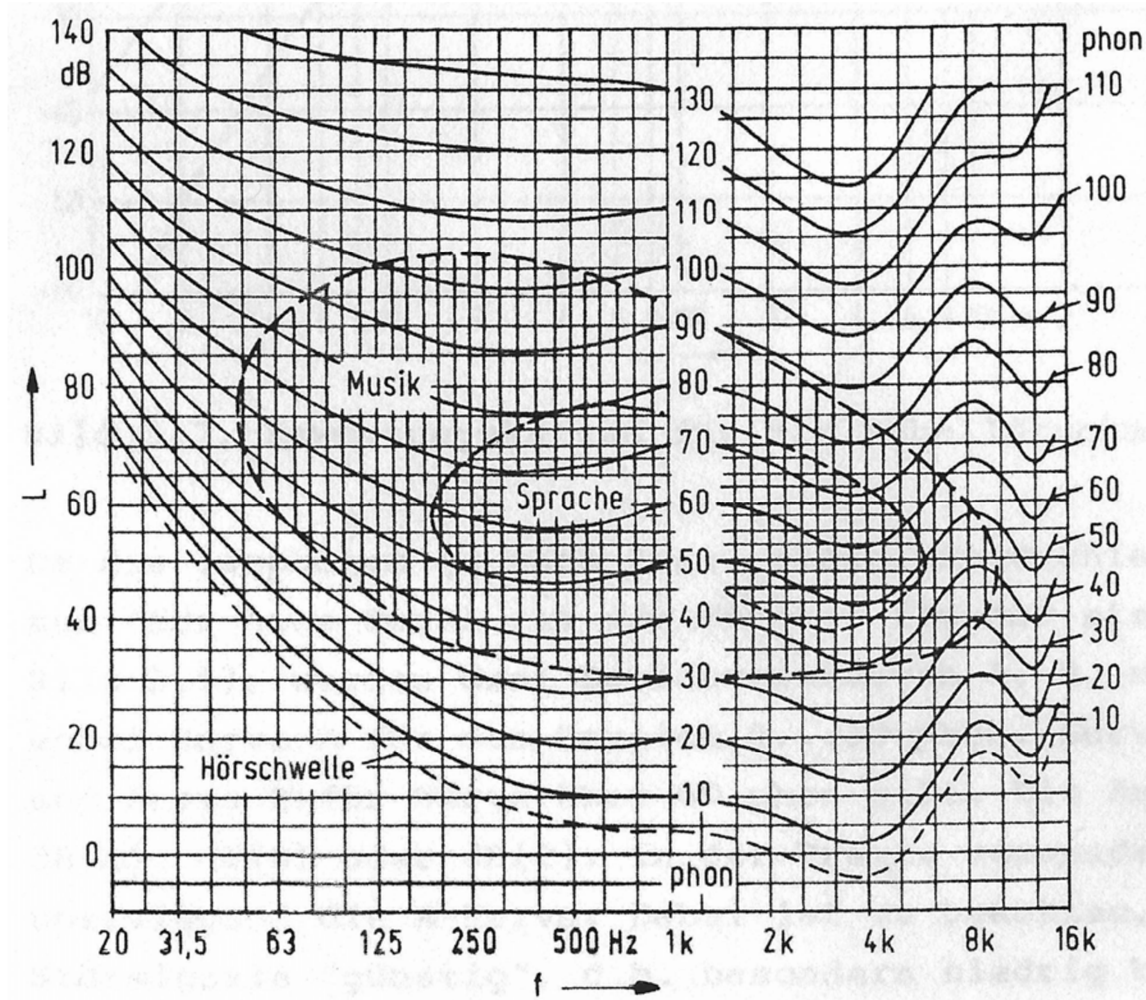
→ Take into account psycho-acoustics characteristics of the humans ear

→ results in receiver oriented coding approach

→ no further characterization of the transmitter

Loudness ~

Signal power (sum of squared amplitudes)



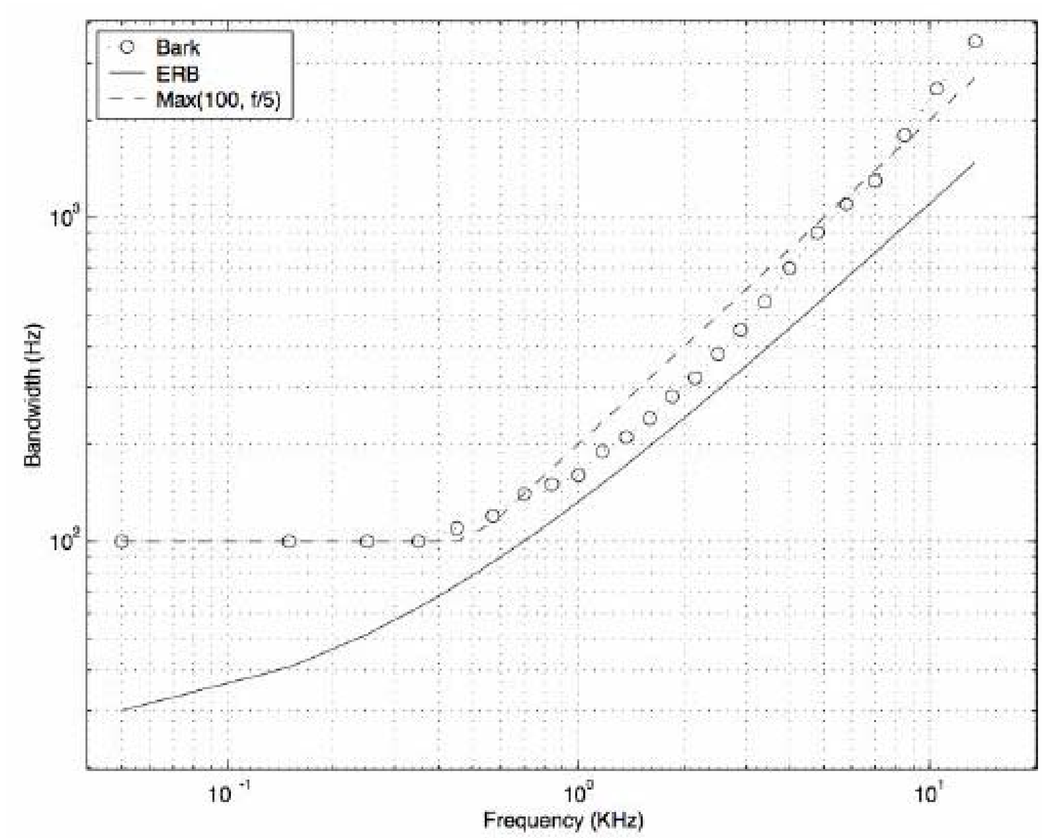
Critical Bands

Definition:

- Minimum frequency difference to differentiate to sinusoids
- Difference depends on the pitch

Ear Characteristics:

- Partitioning of audible spectral range into 25 groups
- Ear works as a filter bank (third octave band filters)

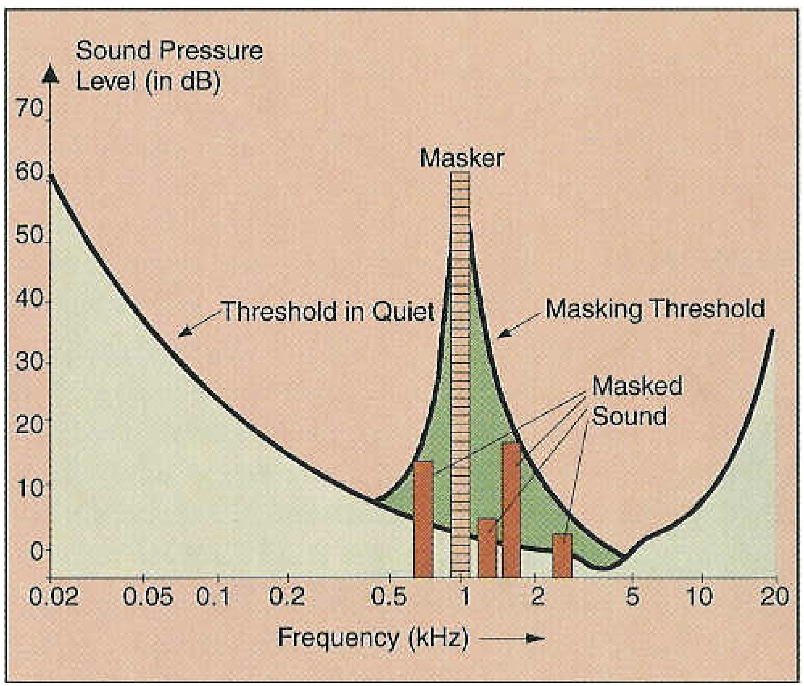


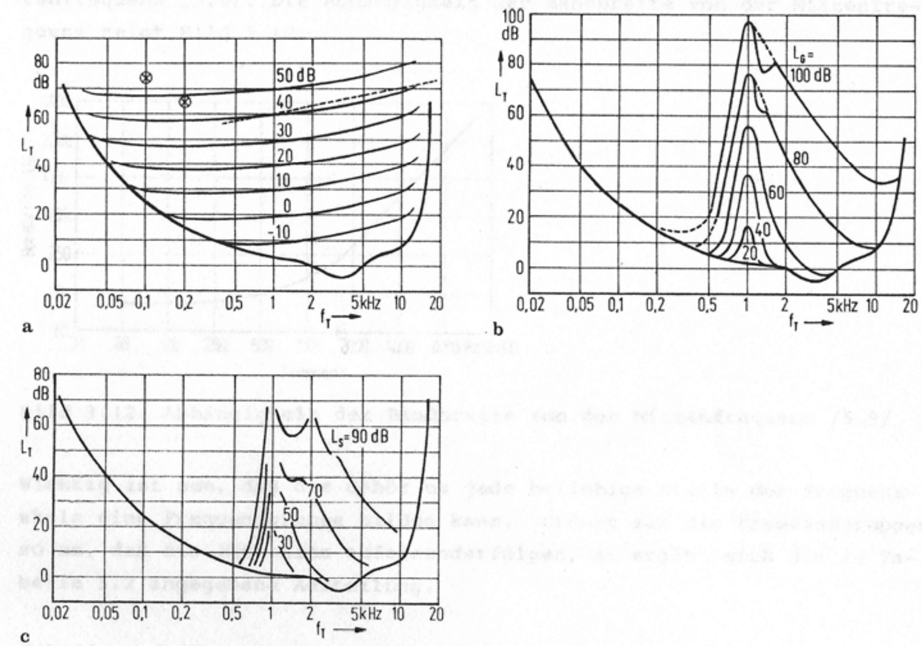
ERB: Equivalent rectangular bandwidth

Masked Threshold and Masking

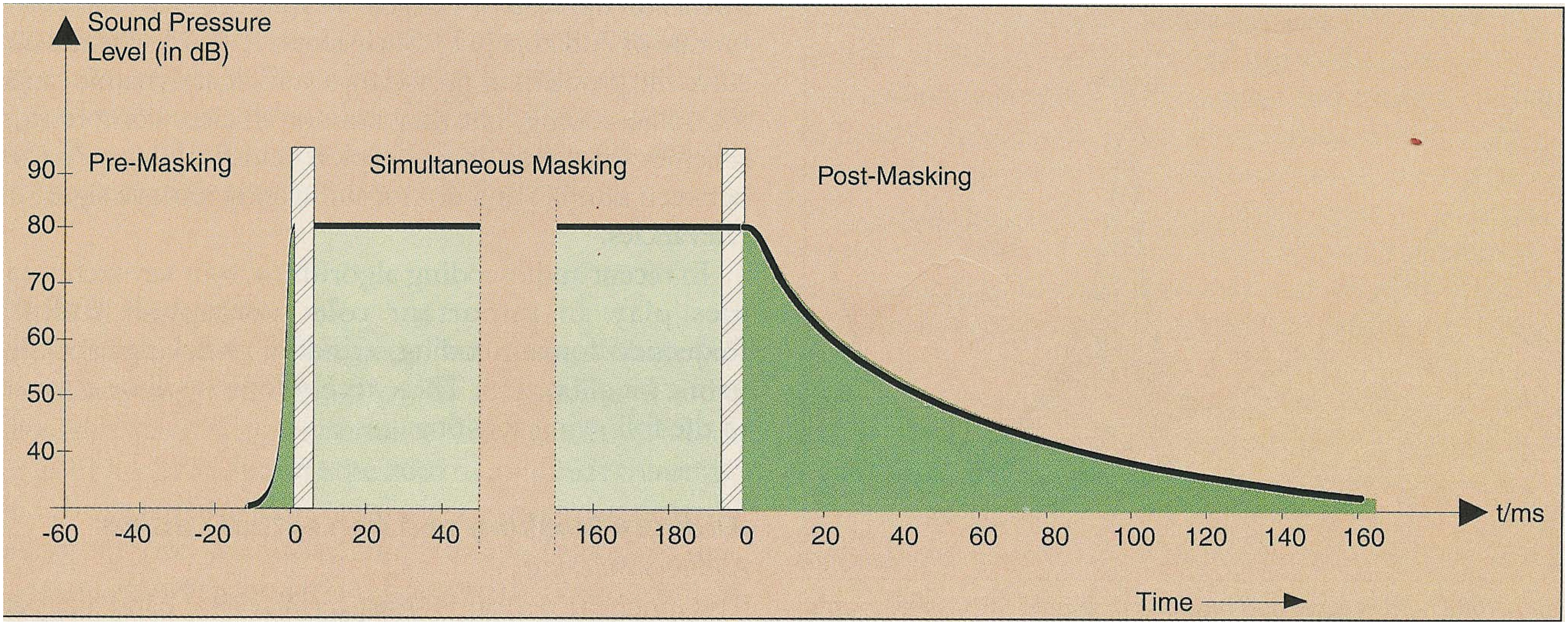
Masking

- amplitude: loud ton masks muted ones
- time: a muted ton immediately after a loud tone can not be heard for some time
- frequency: a loud sinusoid of a specific frequency masks sinusoids of smaller amplitude with a frequency in the neighborhood of the masking tone





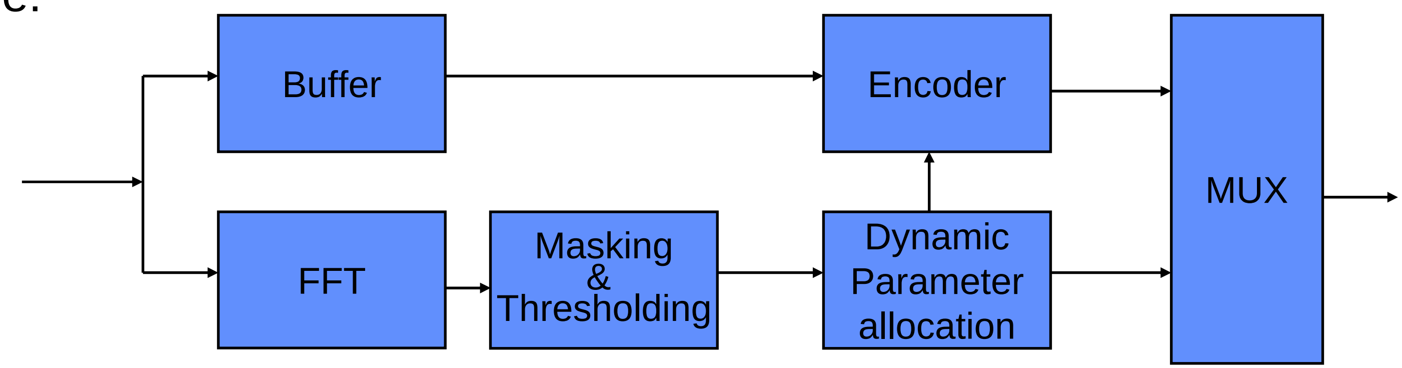
Masking in Time



MPEG-1 Audio

- First high quality audio compression standard 1992 (ISO/IEC 11172-3)
- employs a psycho-acoustical model of the human hearing
  - MUSICAM-Approach (*Masking pattern adapted Universal Subband Integrated Coding And Multiplexing*)
  - and ASPEC (*Adaptive Spectral Perceptual Entropy Coding)*
- Standard specifies the decoder → leaving room for individual optimization at the encoder
- Supports sampling frequencies 32 kHz, 44.1 kHz, 48 kHz
- Operation modes: mono, two channels (e.g. multi lingual), stereo (ind. channels), joint stereo
- Average bit rate for transparent audio between 128 and 384 kbps

Principle:



MPEG-1 Audio Layer

3 Layer (hierarchical):

Increasing complexity, delay and quality (all are rated as perceptually lossless)

Layer 1: CD quality @ 384 kbps stereo

- 32 subbands, 511tap-Filter (PQMF)
- 8 msec frame (12 x 32 = 384 samples / frame)
- DCC (Digital Compact Cassette)

Layer 2: CD quality @192 – 256 kbps / stereo

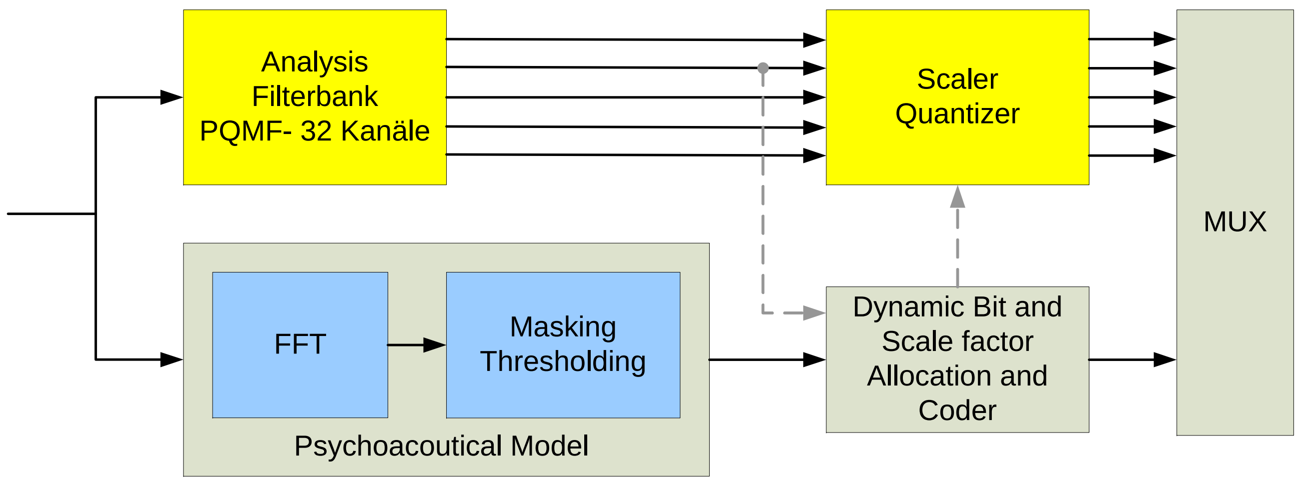
- 32 subbands, 511Tap Filter (PQMF)
- 24 msec frame, (3 x 12 x 32 = 1152 samples/ frame)
- Digital Radio (DAB)

Layer 3: CD @ 112 – 128 kbps / stereo

- Layer II + ASPEC – components
- MP3-Player / Podcast etc.

Encoder Characteristics (Layer 1 and Layer 2)

- Decomposition into 32 subbands (polyphase filterbank)
  - Layer 1: processing 12 samples per subband as one unit
  - Layer 2: processing 3 x 12 samples per subband as one unit
- Quantization and coding of each spectral coefficient such that coding error is below the masking threshold
  - Individually determining scaling value and number of quantizer bits per block



MPEG-1 Subband Decomposition (Normative)

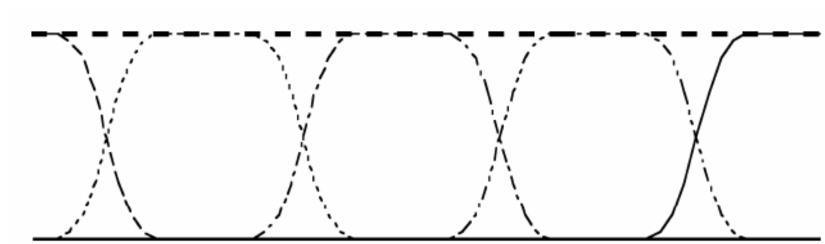
Filter bank implementation based on prototype-filter

$$h_i[n] = h[n] \cos\left(\frac{2k-1}{2 \cdot 32} + \varphi[k]\right)$$

Decomposition into 32 equally spaced subbands

Prototype-Filter (48kHz sampling rate):
3dB band width → 375 Hz, center frequency: (2n+1) * 375

Problem: decomposition does not match the critical band partition
48 kHz sampling frequency → band width = 750Hz

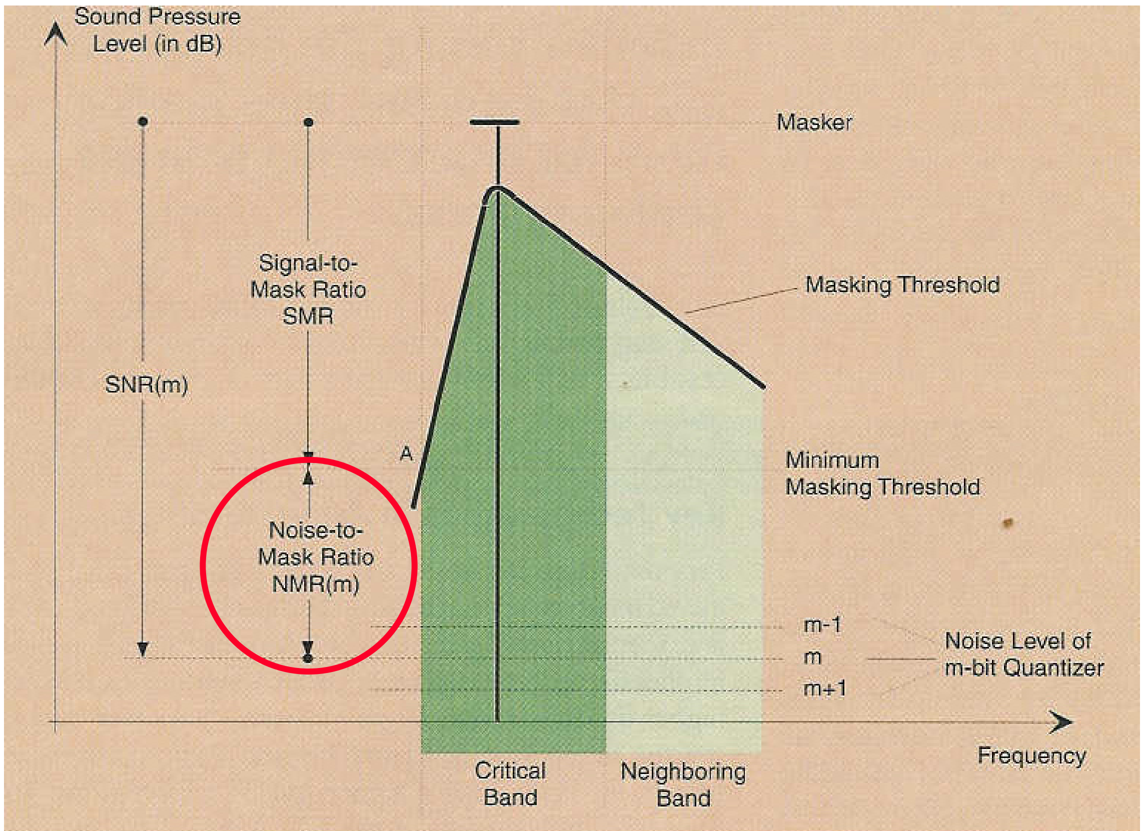


→ Frequency bands overlap but still a perfect reconstruction is possible

Bit Allocation

- Set of linear mid-thread quantizer, variable number per subband
- Iterative selection of “optimal” quantizer considering SNR and masking

$$NMR = SNR - SMR$$

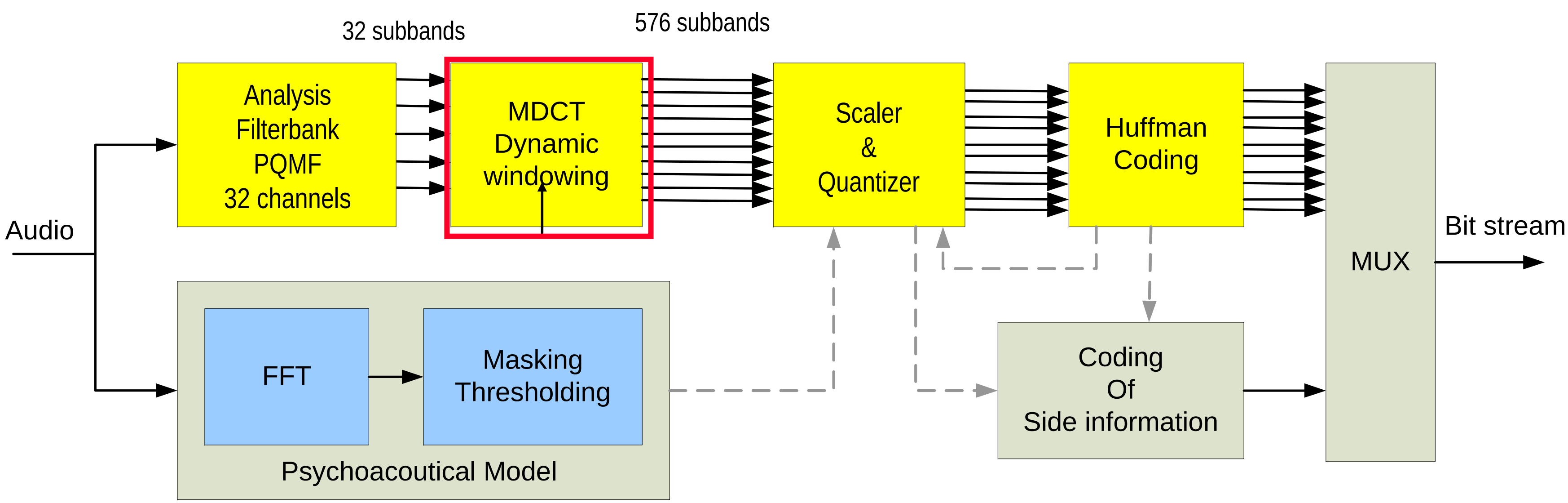


MPEG-1 Layer 3

Essential Features:

- switched hybrid filter bank

  PQMF followed by 6tap / 18tap MDCT (modified DCT) with 50% overlap
- improved pre-echo controller (caused by spreading quantization errors over a frame)
- non-linear quantization
- entropy coding utilizes run-length and Huffman codes
- iterative „analysis-by-synthesis“ optimization of bit allocation
- bit buffer („bit reservoir“)

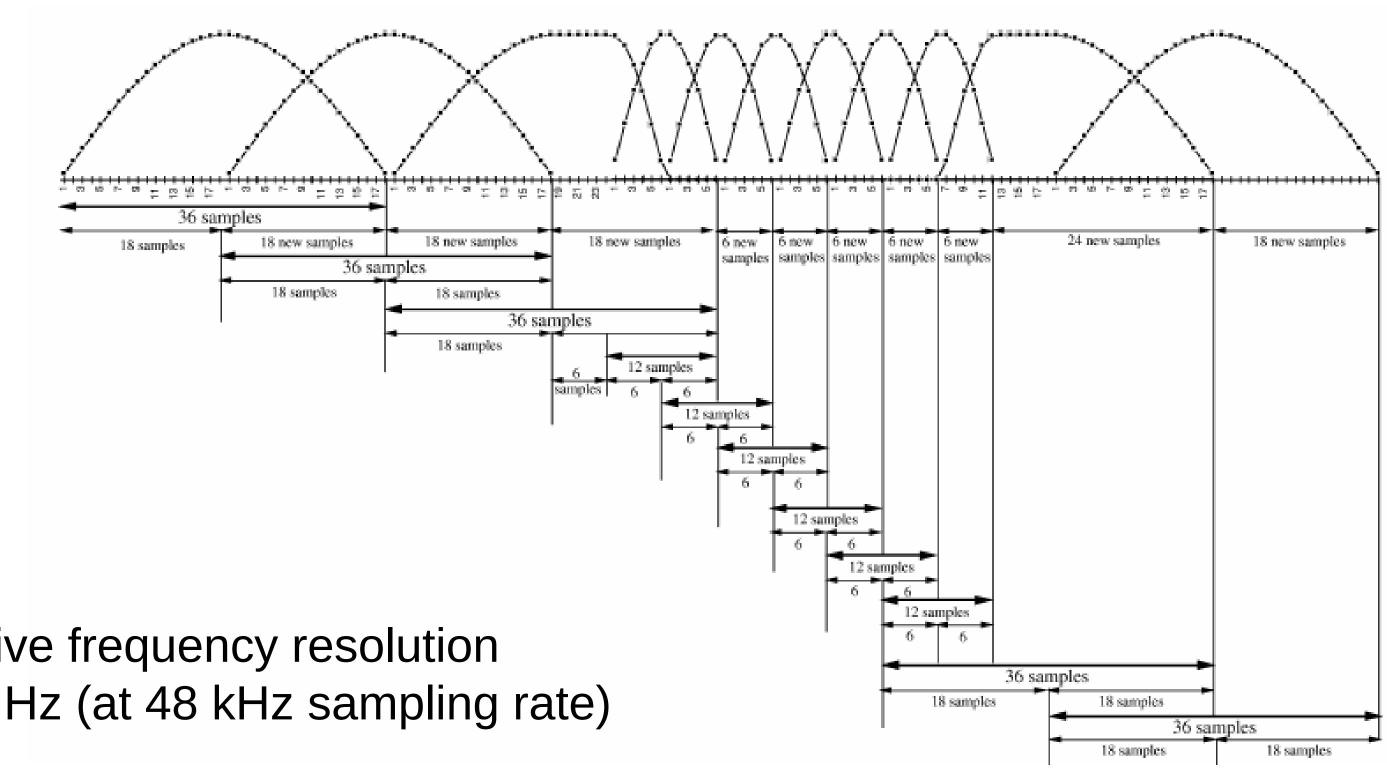


Dynamic Windows and MDCT

MDCT

$$X_i[m] = \sum_{k=0}^{n-1} w[k] x_i[k] \cos\left(\frac{\pi}{2n}(2k+1+n/2)(2m+1)\right)$$

i. Suband



Effective frequency resolution
41,67 Hz (at 48 kHz sampling rate)

MPEG-2 Audio

2 driving motivations

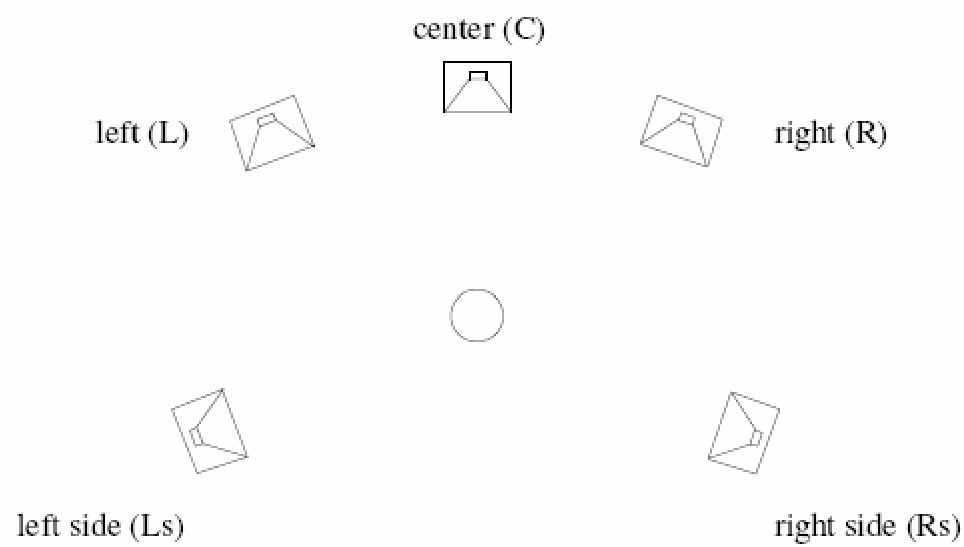
- Improve coding efficiency
- Add functionality for multi-channel

- **BC: Backward compatible multi channel extensions (Nov 1994)**
  - designed for DVD (lost in the market against Dolby Digital)
  - Backward compatible as MPEG-1 decoder can generate 2.0 Signals from 5.1 multi-channel signals
  - Forward compatible as MPEG.2 decoders can decode MPEG-1 mono and stereo signals
- **LSF: low sampling frequency Extension**
  - 16, 22.5, 24 kHz sampling rates
  - Extends MPEG-1 down to 8 kbps
- **AAC: advanced audio coding scheme (1997)**
  - Not backward compatible
  - Reaches at ~96kpbs the quality of MP3 at 128 kpbs

Multi-Channel Improves Room Experience

channel	configuration	description
1	1/0 (+1)	Center (Mono)
2	2/0 (+1)	L,R (Stereo)
3	3/0 (+1)	L,R,Center
4	3/1 (+1)	L,R,Center, mono
5	3/2 (+1)	L,R,C,Surround L, Surround R

Some multi-channel configurations



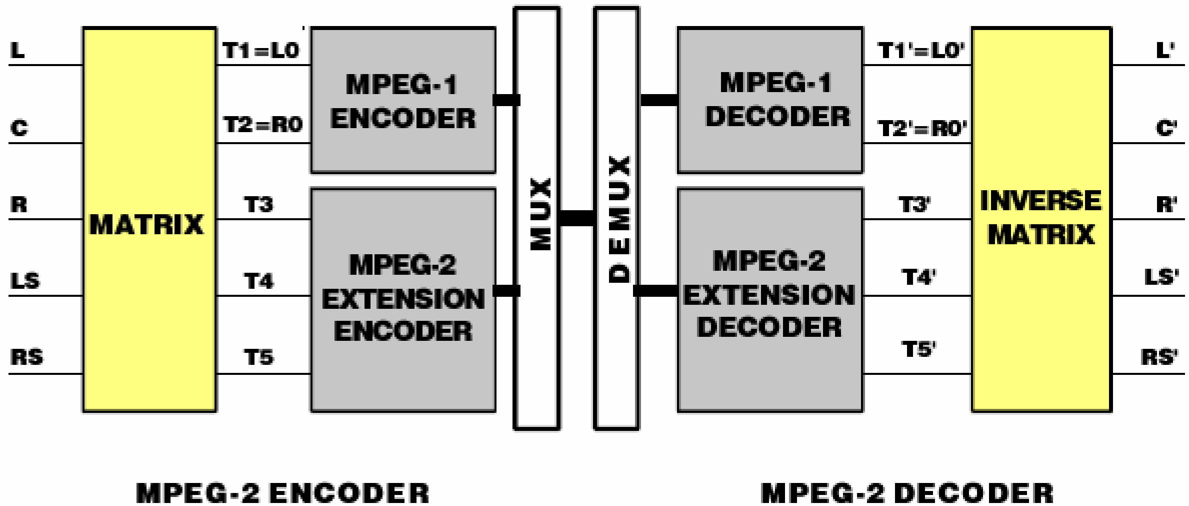
Subwoofer (low frequency enhancement) can be added to each configuration (up to 120 Hz)
3/2 + 1 is commonly referred to as 5.1 multi-channel

Aspects for processing:

- Down mix for backward compatible reduction of output signals (loudspeakers)
- efficient coding by utilizing correlation between channels

MPEG-2 Approach to Multi-channel Support

Down mix matrix is the key feature of MPEG-2 multi channel audio



Backward compatibility

$$L0 = \alpha\left(L + \beta \cdot C + \delta \cdot LS\right)$$
$$R0 = \alpha\left(R + \beta \cdot C + \delta \cdot RS\right)$$

MPEG-2 supports many different matrices, including time dependent ones

$$\alpha = \frac{1}{1+\sqrt{2}}; \beta = \delta = \sqrt{2}$$

Coding of Stereo / Multichannel Signals

- **Intensity stereo coding**
  - Transmission of a combination of left and right signal
  - Directivity information is part of the scaling factor
- **Middle / Side stereo coding**
  - Transmitting normalized sum and differential signals

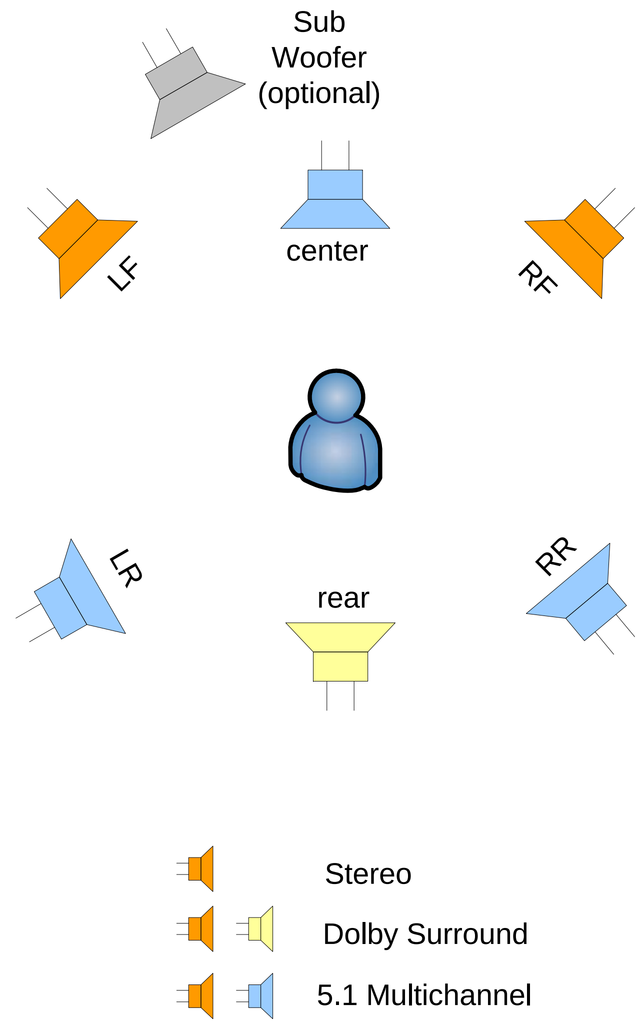
- MPEG-2 5.1 allows bit rates between 384 and 640 kpbs

Market environment:

- Currently employed for digital radio over DVB-S (in Germany)
- multi-channel radio in DAB possible (trials)

Other „Suround-Sound“ Algorithms

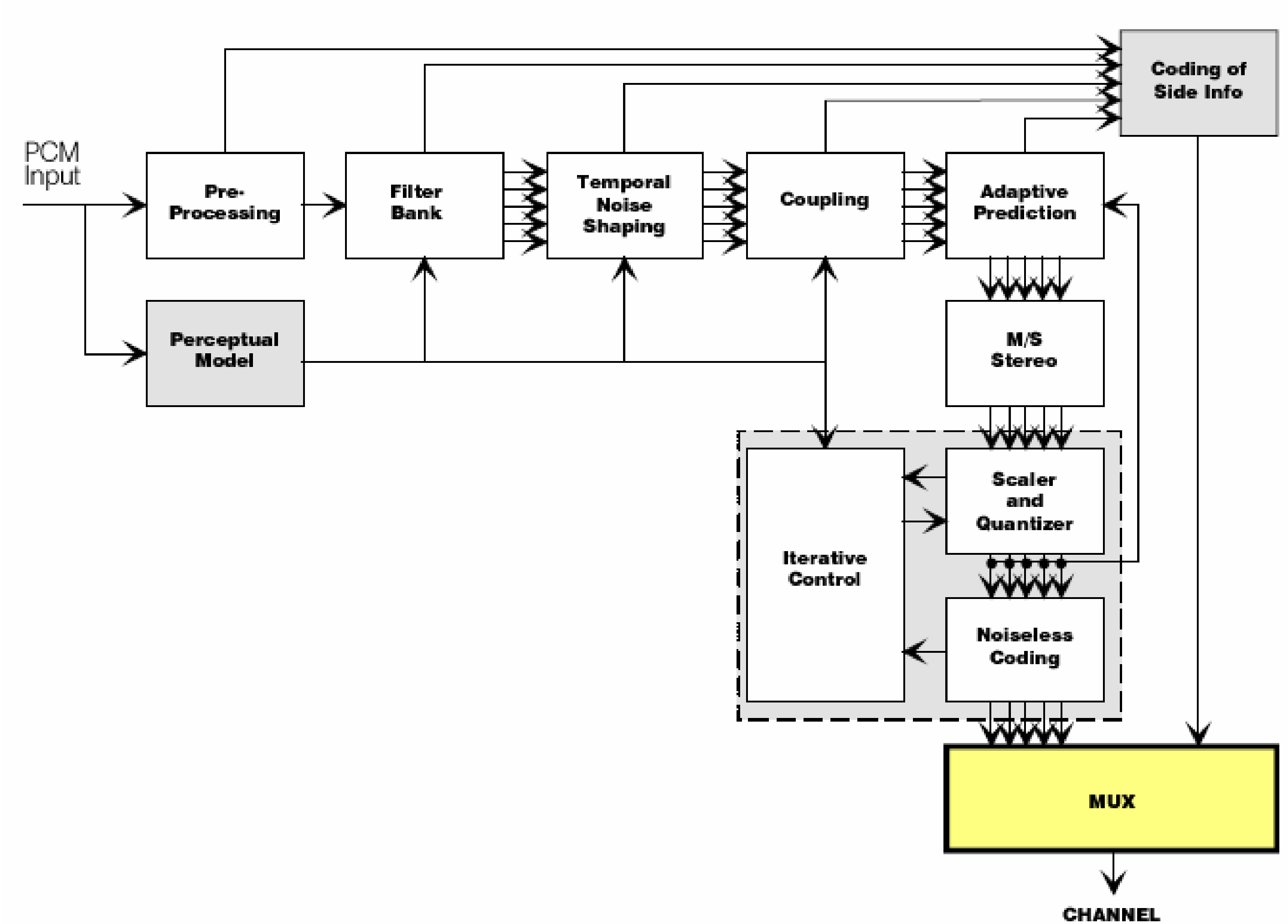
- **Dolby Surround:**
  - No complete reconstruction of multi-channel possible
    - Complete information only in 2-channel stereo signal
    - as matrix signal available
  - can also be synthesized from stereo signal
- **5.1 Multi-channel Solutions**
  - Dolby Digital (AC3)
    - Joint coding of up to 5 channels in a single bitstream
    - Not backwards compatible to MPEGr
    - Very high market penetration
  - DTS (Digital Theatre System)
    - Proprietary coding scheme (5 independent channels)
    - Devices only for DVD in the market (no broadcast)
  - MPEG Surround (matrix approach)
    - Backward compatible extension of MPEG-1 audio
  - Other schemes (matrix based)
    - Dolby Pro Logic II / SRS Circle Surround / Logic 7



MPEG-2 AAC (→ MPEG-2 NBC)

- **MPEG-4, completed 1997**
  - ITU-R indistinguishable audio quality per stereo channel at 128kbps
- **Profile:**
  - Main: all tools except „gain control module“ → maximum quality
  - Low complexity: no „prediction tool“, reduced complexity of noise shaping tools
  - Sample rate scalable profile: low complexity profile + gain control tool → encoder of lowest complexity
- **Extends MP3 with tools**
  - Temporal noise shaping
  - Backward adaptive linear prediction
  - Enhanced joint stereo coding
- **Sampling: 8 – 96 kHz,**
- **Bit rates: 16 – 576 kbps**
- **1 – 48 audio channels**
- **Delay: 24kHz@ 24 kbps → 110 ms + 210 ms bit buffer**

MPEG-2 AAC Struktur



AAC Features

- **Filterbank:**
  - MDCT as perfect reconstructing filter bank (approacht: Time Domain Aliasing Cancellation)
  - length: 2048 samples, 50% overlap → 1024 „new“ samples → 23.4 Hz frequency resolution
  - NO preceding PQMF
- **Window function**
  - Adaptive length ranging from 1024 to 128 (frequency resolution versus PRE-Echo cancellation)
  - Sinusoid or Kaiser-Bessel window
- **Non-linear quantiser**
- **Noise Shaping („hiding“ of quantizer noise)**
  - Adapting the quantizer step sizes by means of a scaling factor
- **Noiseless coding**
- **Temporal Noise Shaping**
  - Shifting quantizer errors in time
  - Forward prediction in the spectral domain
- **Prediction**
  - Predicting spectral coefficients from previous frame

MPEG-4 Audio Coding

- **Part 1: (10/1997)**
  - Audio and speech coding @ 2 .. 64 kbps
  - Analysis stage based on <u>source model</u> (extracting parameters)
  - Coding based on a <u>perception model</u>
  - speech
    - HVXC (parametric speech coding at very low bit rates ranging from 2 .. 4 kbps)
    - CELP (NB + WB, 4 … 24 kbps, 8 / 16 kHz sampling frequency)
  - Audio – Twin VQ (6 …16 kbps / ch) / AAC (+scalable) (16 … 64+ kpbs / ch)
  - System: composition of <u>audio objects</u> to <u>audio scenes</u> (mixing, effects → structured audio)
- **Part 2: (12/1999)**
  - Extended functionality
  - Backward compatible, includes v1
  - Extension tools
    - Error robustness
    - Low delay audio coding
    - Small step scalability
    - Parametric audio coding
    - CELP / HVXC Silence Suppression
  - System: environmental spatialization / File format

MPEG-4 Audio Profile

- **Speech Audio Profile**
  - Parametric speech (HVXC)
  - CELP
- **Synthesis Audio profile**
  - Generate speech and audio
- **Scalable Audio Profile**
  - Includes speech audio profile
  - AAC
  - TwinVQ tools
  - Scalable coding of speech and music
- **Main audio profile**
  - Contains all other profiles

MPEG-4 Audio Tool Categories

- **Natural speech**
  - CELP Coder: MPEG specifies DECODER and syntax of bit stream
  - HVXC Coder: acceptable quality at 2 / 4 kbps
    - → Harmonic Vector Excitation coding
- **natural audio**
  - Based to a large extend on MPEG-2 AAC
  - Coding efficiency: (PNS, LTP, TwinVQ)
  - Functionality: AAC-LD, AAC-ER, MPEG-4 Lossless, scalability
- **synthetic speech**
  - Text-to-Speech (TTS)
- **synthetic audio**
  - Parametric audio coding
  - Structured audio

MPEG-4 AAC Tools

MPEG4-AAC ==
MPEG2-AAC + PNS tool + AAC Long Term Prediction profile (AAC LTP)

PNS: Perceptual Noise Coding

- Vocoder principle (reproduces the sound not the exact waveform)
- for noise signals only the power value is transmitted – the decoder generates artificial noise

LTP: Long-Term Prediction

- reconstructing the coded signal and calculating the difference to the original signal
- adjusting the parameters “pitch lag” and “gain” over time

PS: Parametric Stereo

- generate and encode a mono signal from a stereo signal
- transmit control parameters which allow to synthesize the stereo signal (inter-channel intensity difference, inter-channel cross correlation, inter-channel phase-correlation)
- works primarily at very low bit rates

MPEG-4 AAC Functionality

TwinVW:

- Transform domain weighted interleave vector quantization
- Good quality at very low bit rates (~ 6 kbps)

MPEG-4 Low Delay

- Application: 2 way communication
- 20 msec algorithmic delay
- Approach
  - Reduced window length (512 Samples))
  - No switching of “windows”
  - No utilization of the „bit reservoirs“

MPEG-4 Error Robustness (Phase 2)

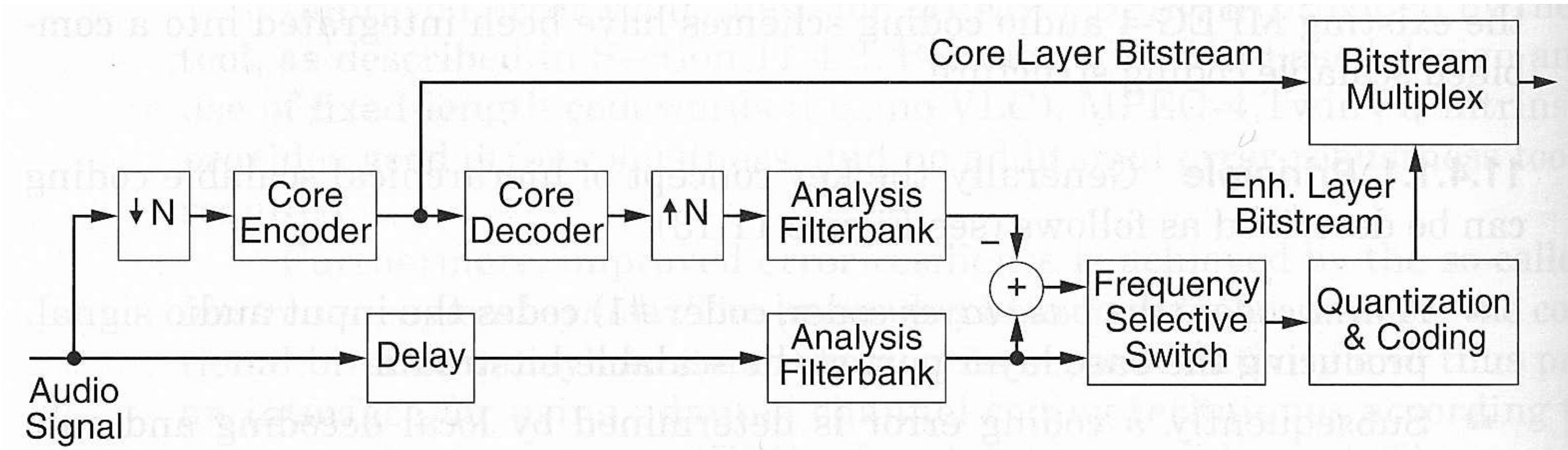
- Virtually extend the code books for large coefficients
- RVLC (reversible VLC) instead of Huffman.Code

MPEG-4 ALS (Lossless)

- Is based on a forward - predictor
- Employs Rice codes

Scalable Audio Coding

Large step size (refinement layer with > 8 kbps)



Small step size (refinement layer with ~1 kbps)

→ Fine granular Scalability

→ MPEG-4 Tools BSAC (bit sliced arithmetic Coding) → Bit plane coding replaces the processing block „Quantizer and Coding“

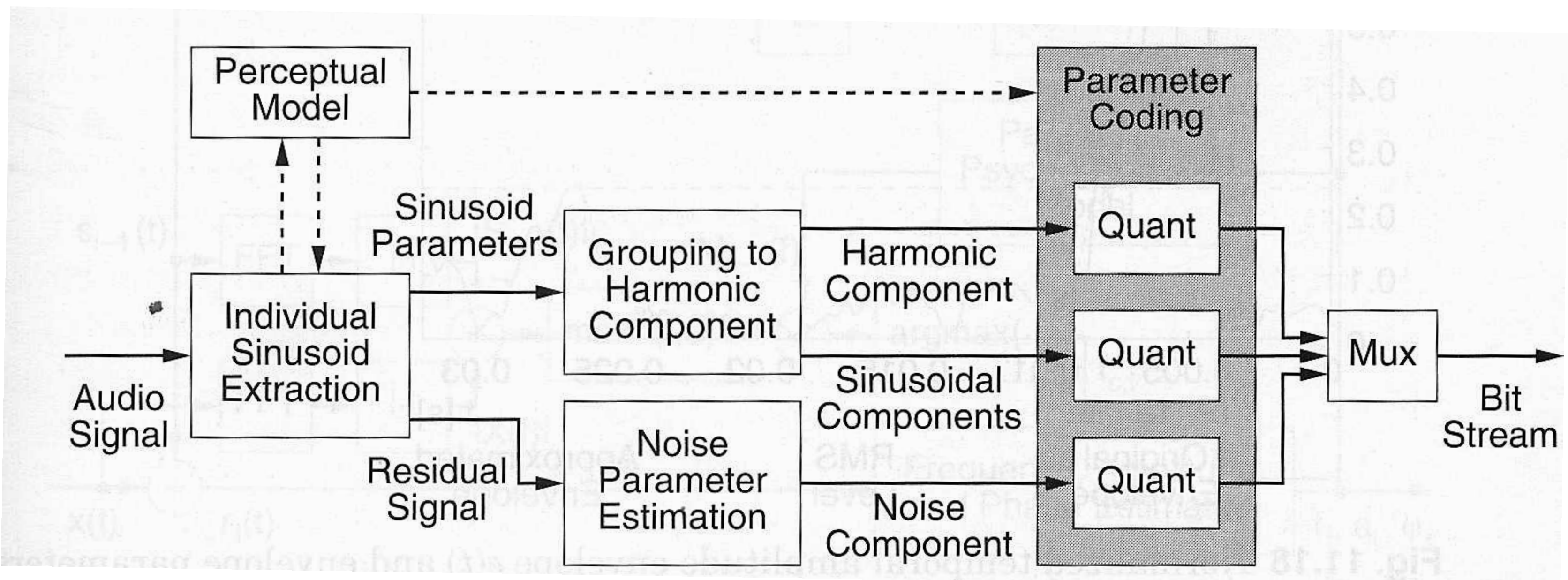
Parametric Audio Coding

Goal: Coding of audio at very low bit rates (4 … 16 kbps)

Approach: Audio „Vocoder“

Signal model is a composition of

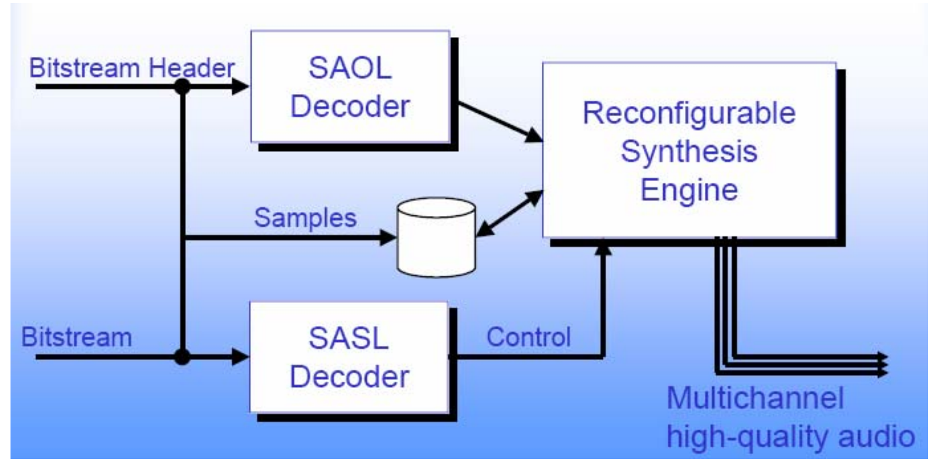
- Transients (highly dynamic sounds, such as percussion)
- Noise
- Harmonic sounds



Structured Audio

Structured Audio

- sound specification by means of a structured description
- Decoder synthesizes the sound employing the description



SAOL: Sound Synthesis Language „Structured Audio Orchestra Language“

→ describing the synthesis method

SASBF: Structured Audio Sample Bank Format

→ describing the waveform tables

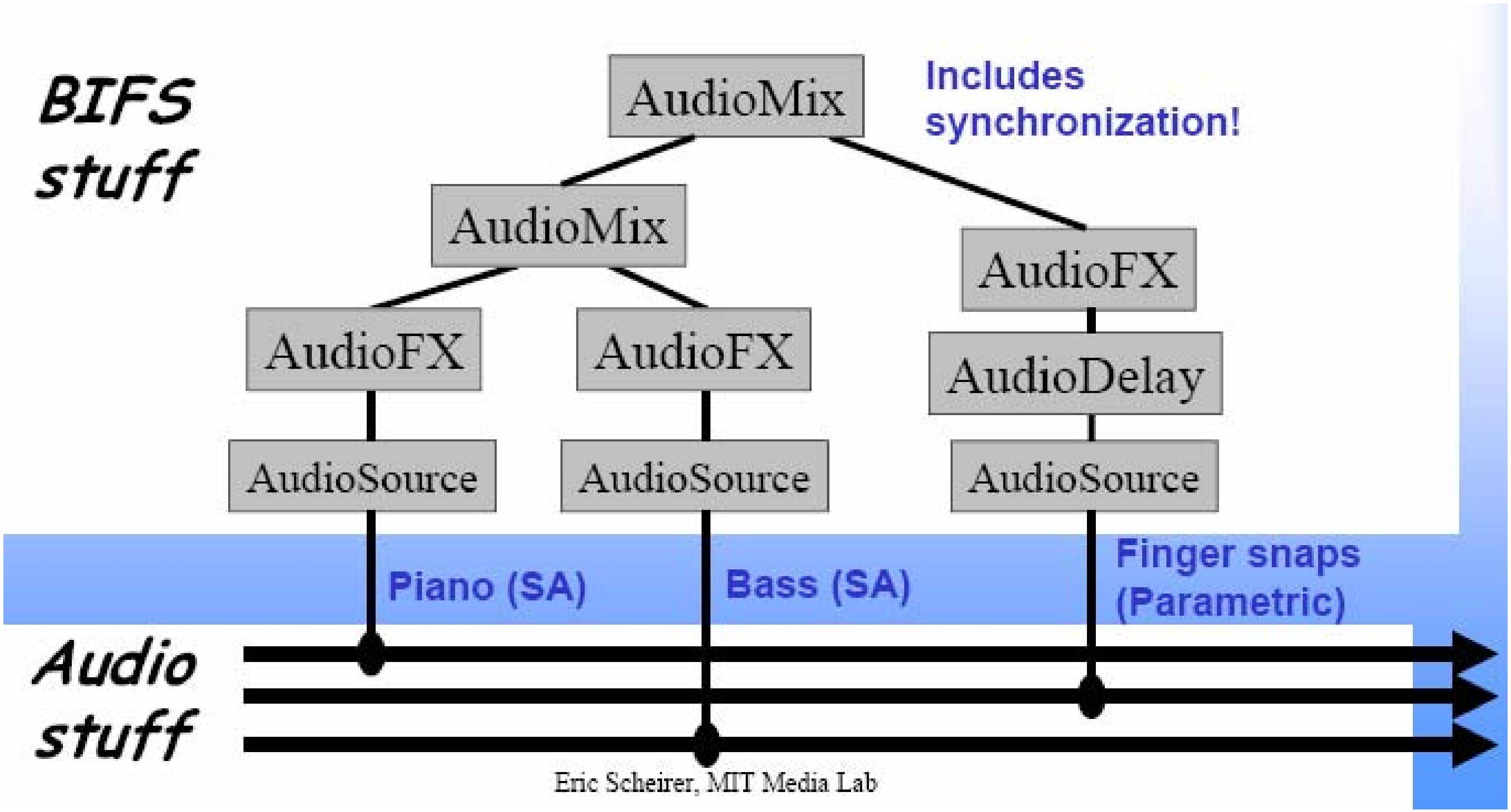
SASL: Structured Audio Score Language

→ describing the control parameters

MIDI: Musical Instrument Digital Interface

→ simplified format for describing control functionalities

Structured Audio Scene

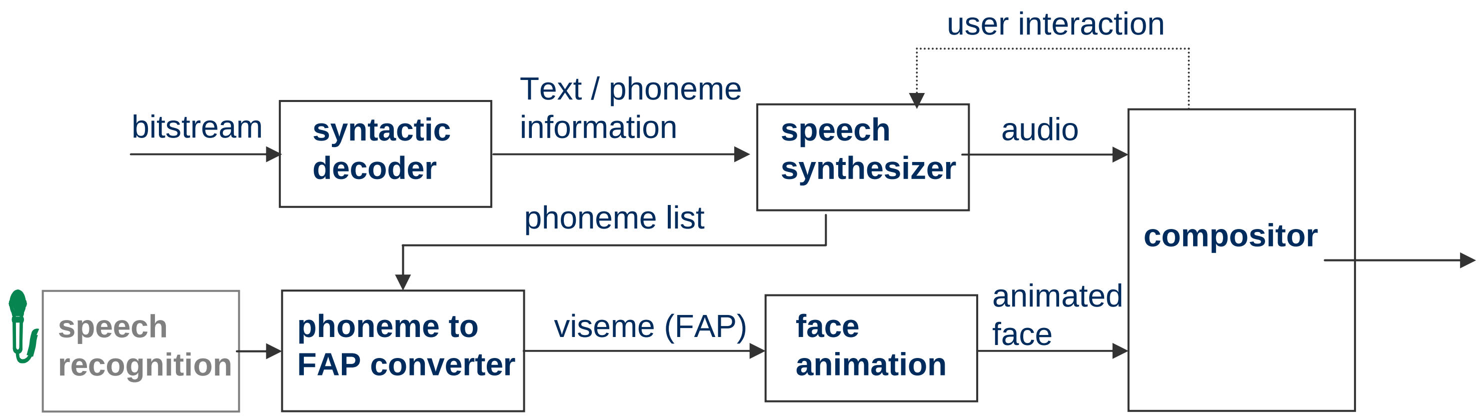


Text – to – Speech

principle: transmitting of sentences

complemented by

- Speaker related information
- Prosody
- „Lip shape“ information
- Speech code (ID)
- Emoticons (Smilie)
- Parameters controlling the face animation



Further Improvements: HE-AAC / AACPlus v1

Observation:

- Significant correlation between higher and lower frequency components

Approach:

→ Spectral Band Replication (SBR)

- Reconstruct higher frequency components from the coded lower frequency signal components in combination with some control information
- applicable for speech and audio codecs

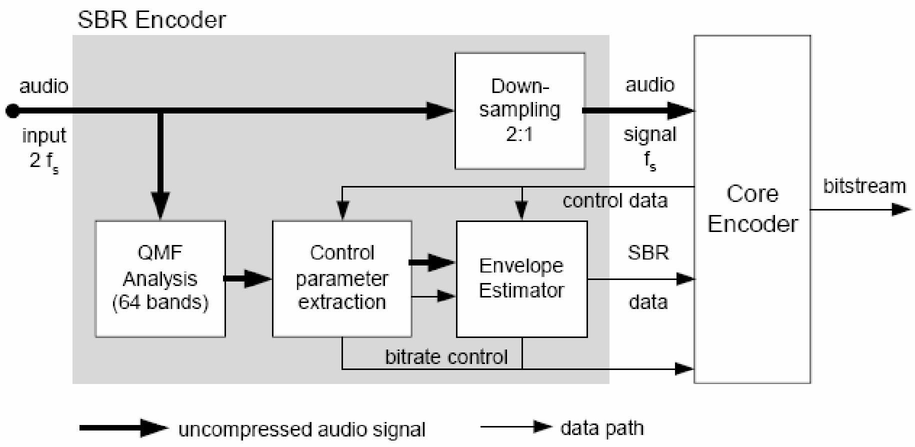
AAC + SBC = aacPlus / MPEG-4 HE-AAC (2003) / mp3PRO (2001)

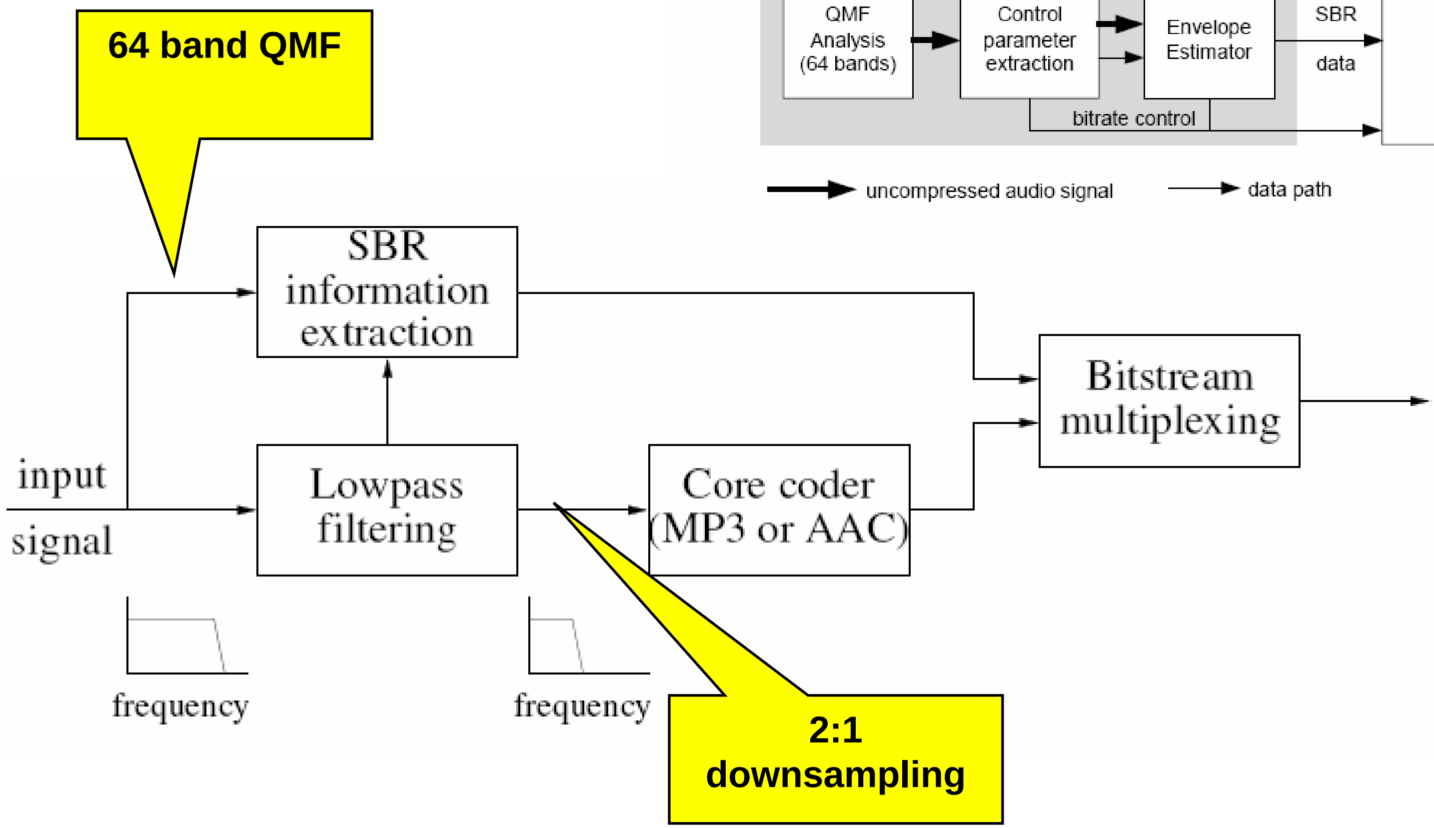
→ improves coding efficiency by about 30% compared to AAC

SBR side information (ca. < 10% of the entire bit rate)

- Frequencies to be reconstructed
- Spectral envelop of the higher frequencies
- Tonality of higher frequencies
- Time frequency resolution of envelop and tonalityt

AAC-SBR


SBR Encoder
audio
input
2 f_s
Down-
sampling
2:1
audio
signal
f_s
control data
Core
Encoder
bitstream
QMF
Analysis
(64 bands)
Control
parameter
extraction
Envelope
Estimator
SBR
data
bitrate control
uncompressed audio signal
data path


64 band QMF
SBR
information
extraction
Bitstream
multiplexing
input
signal
Lowpass
filtering
Core coder
(MP3 or AAC)
frequency
frequency
2:1
downsampling

Quality Comparison

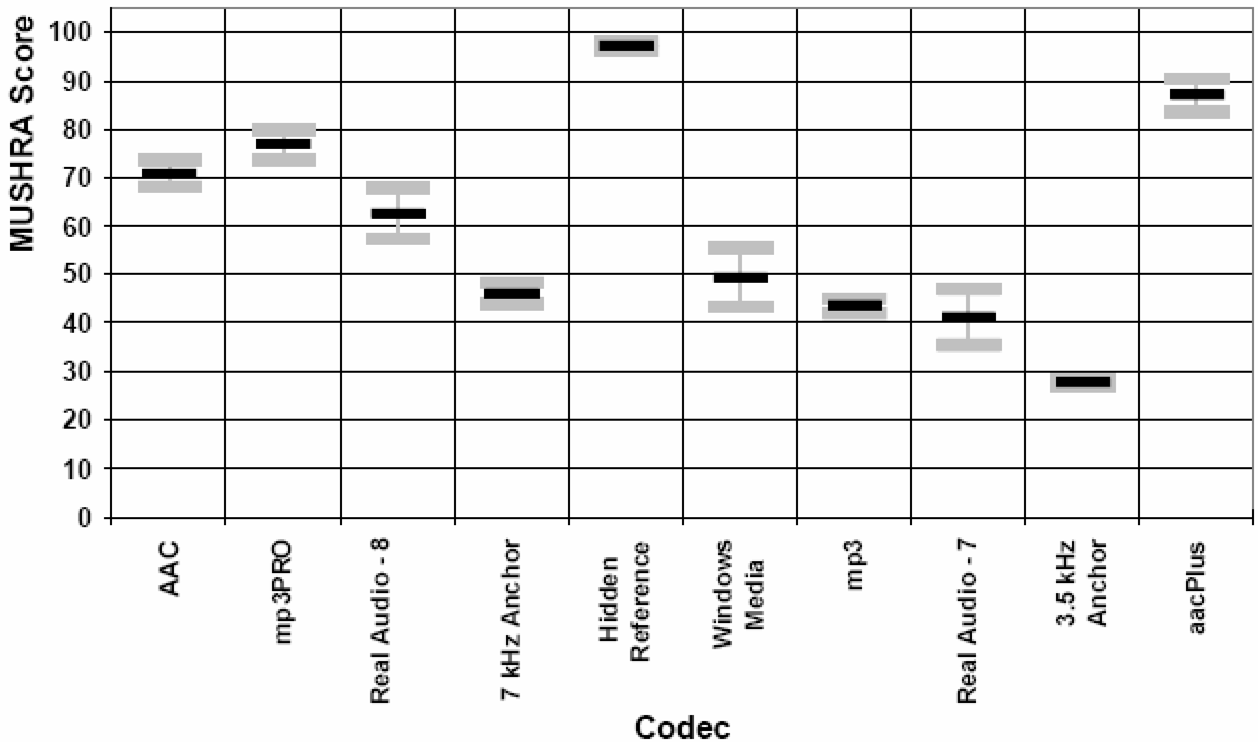
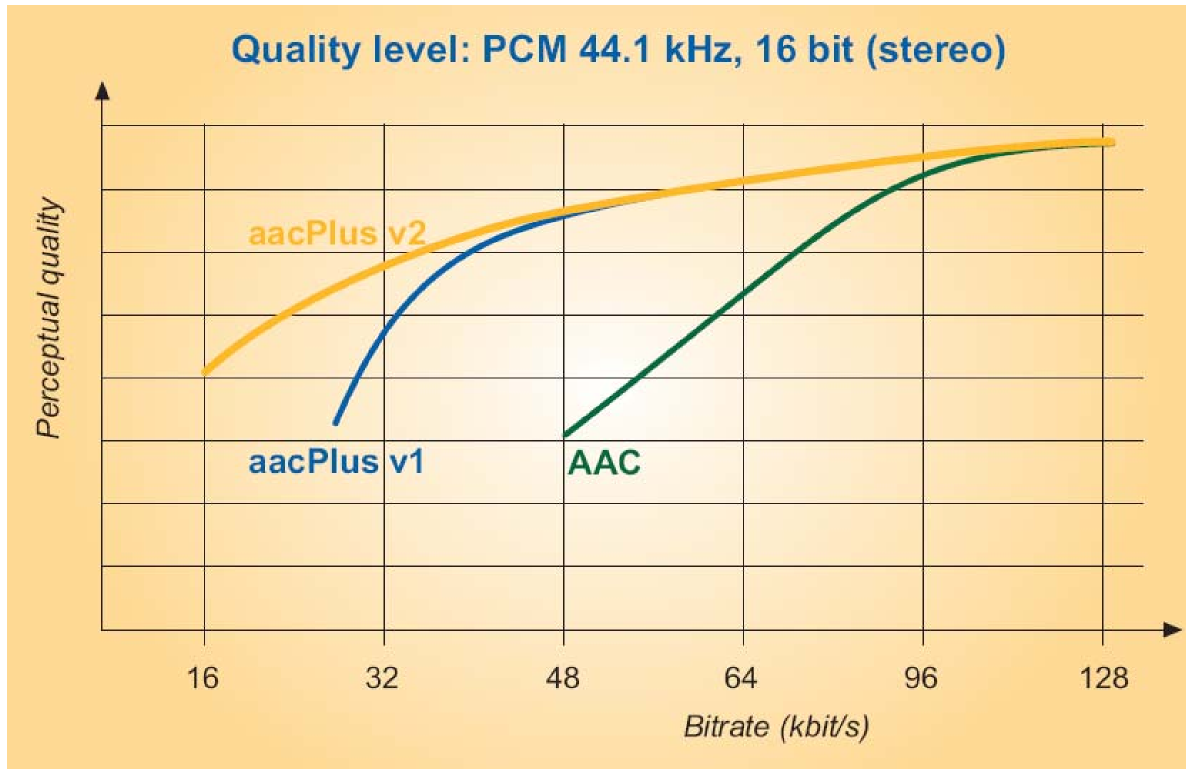


Figure 3: EBU test results for 48 kbps stereo items. Mean values and 95 % confidence intervals. IRT test site.

HE-AAC v2 / AACPlus v2

- **AACPlus v2 = AACPlus v1 + PS (parametric stereo)**
- **Most efficient audio codec known today**
- **Employed in a variety of systems**
  - Digital Radio Mondeale
  - XM Satellite Radio
  - S-DMB
  - 3GPP



What Next?

Natural reproduction of sound fields at the location of the listener:

- Simulating acoustical fields
- Synthesizing sound fields at different locations
- Wave field synthesis
- *Binaural sky*™ (virtual headphone)

Commercial Aspects of Digital Audio

- **Enables entirely new markets**
  - MP3-player, “iPOD”
- **Digital media distribution independent of networks**
  - DAB, DVB, DRM, Internet Radio, Music on Demand, „file sharing“ (Napster)
- **Significant impact on markets and business models**
  - Reason: Perfect digital copy in combination with easy and fast copy turns over markets
  - Digital millenium rights act
  - RIAA is sueing private persons on a large scale
  - Enforcement of content and copy protection
- **Establishing new alliances**
  - Apple and music industry → iTunes
  - Music groups publish in the Internet
  - EMI announces to relax rights enforcement for iTunes and similar platforms