Lecture

Speech and Audio Signal Processing

Lecture 4: Audio coding, Part II

Content of the lecture

- Audio coding

  Part I:
  - Motivation and Principle
  - **Predictive coding:**
    - Signal form coders

  Part II:
  - Two other **predictive coders**:
    - Vocoder and Hybrid coders
  - **Frequency domain / sub-band coders:**
    - MP3 and AAC coders of MPEG2 and MPEG4 standards

Transmission of samples and quantized audio data

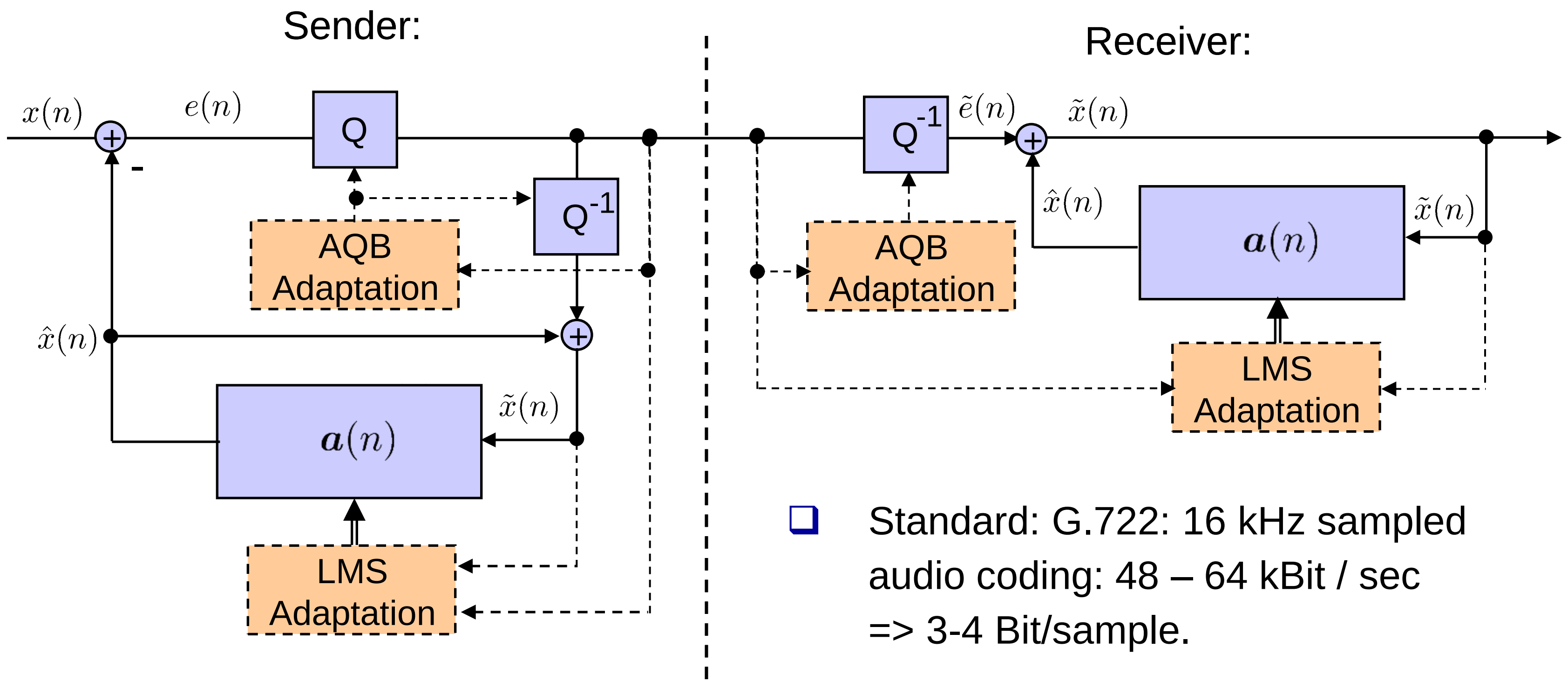
Direct PCM (pulse code modulation) coding:

- Quantize each sample by 8 - 16 Bit
- Telephone speech => 8 kHz sampling with 8 Bit/sample=> 64 kBit / sec (ISDN coding)
- Wideband speech => 16 kHz sampling with 8 Bit/sample=> 128 kBit / sec
- Audio data 16 bit / sample
  (SNR = 6*16 = 96 dB SNR, i.e., signal to quantization noise) :

  1) 16 kHz sample rate: 256 kBit / sec
  2) 22.05 kHz sample rate: 352.8 kBit / sec
  3) 44.1 kHz sample rate (CD): 705.6 kBit / sec
  4) 48 kHz sample rate: 768 kBit / sec

Repetition of Signal form coders

- High quality coding with Bit rates > 1.5 Bit / sample.
- Typically ADPCM based
- Adaptive quantization and prediction calculation
- Transmission of prediction error filter output signal, no transmission of predictor coefficients.



- Standard: G.722: 16 kHz sampled audio coding: 48 – 64 kBit / sec => 3-4 Bit/sample.

Vocoder

Vocoder

QUEE? isso é mt baixo

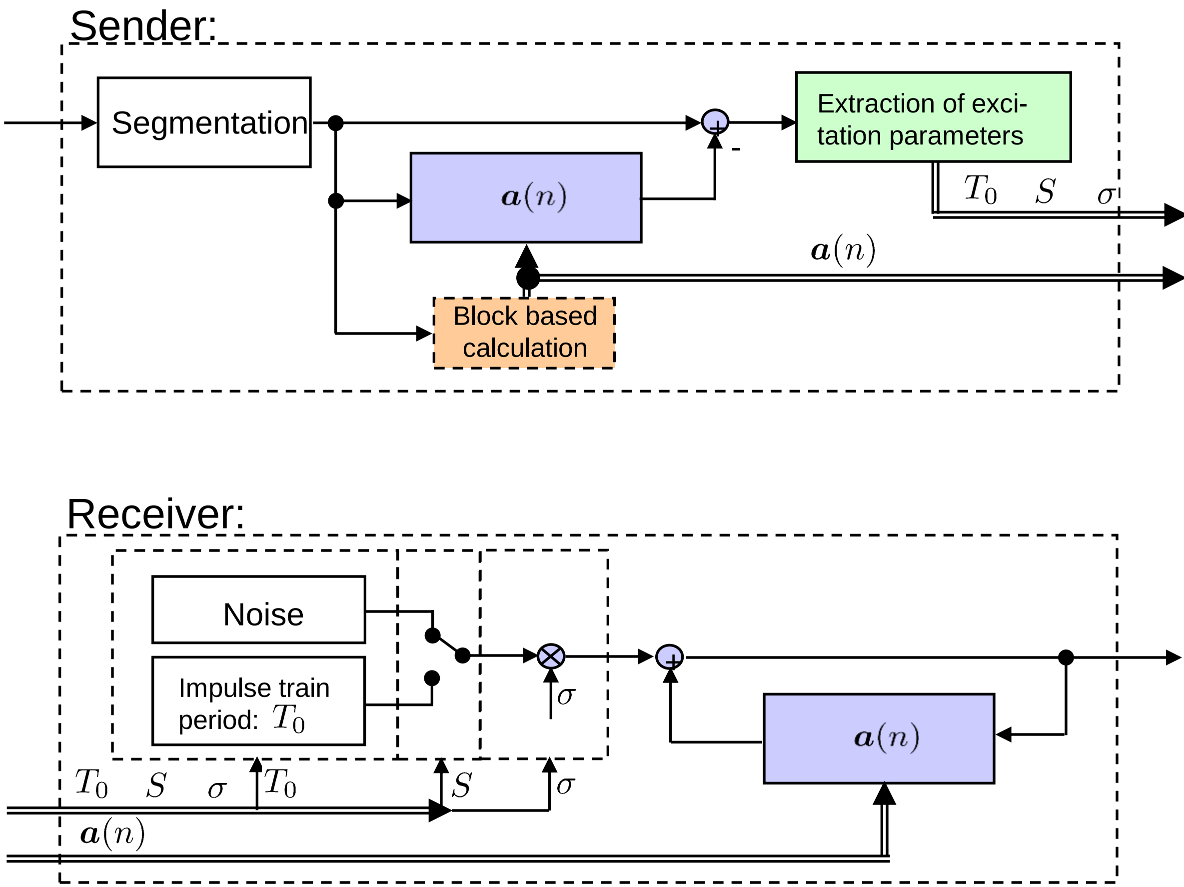
- **Vocoder, i.e., Voice coder**, developed for a low data rate coding of speech (0.1 – 0.5 Bit / sample).
- Concept based on speech generation model (combination of noise or period train excitation with spectral forming)
- Typically, the decoded speech signals show a low-natural speech quality, however, a rather good intelligibility.

- **Concept:**
  - Extraction of the short-term spectral envelope and
  - information about excitation signal.

- **Transmitted data:**
  - Spectral envelope => Predictor reflection coefficients
  - Info of excitation signal => voiced / unvoiced flag, pitch period, excitation power

Vocoder

- **LPC-Vocoder (LPC: linear predictive coding):**



- **Example:**

Parameter:
- fs = 8 kHz
- 22.5 ms frame = 180 samples

- **Predictor:**
10 prediction coefficients (voiced)
4 prediction coefficients (unvoiced)
voiced: 41 Bit (2x5 Bit+ 8x [2..5 Bit])
unvoiced: 20 Bit (4x5 Bit)

- **Signal excitation (13 Bit):**
6 Bit: Pitch period: T_0
1 Bit: Voiced / unvoiced: S
5 Bit: gain: σ
1 Bit: synchronization

Data rates:

Voiced (0.3 Bit / sample):
54Bit / 22.5ms = 2.4 kBit/s

Unvoiced (0.18 Bit / sample):
33Bit / 22.5ms = 1.47 kBit/s

Hybrid Coder

Hybrid Coder

- **Hybrid coder,** exhibit a mean data rate (between “signal form coder” and vocoder (0.5 – 1.5 Bit / sample).

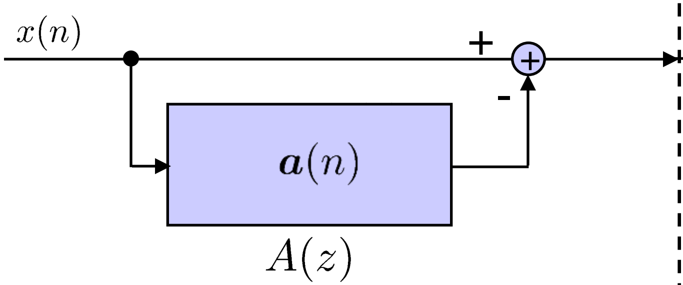
- Main application: speech (fs = 8 kHz) => data rates between 4 and 12 kBit / s. Mobile Phones, storing of speech data, audio-channel multiplexing.

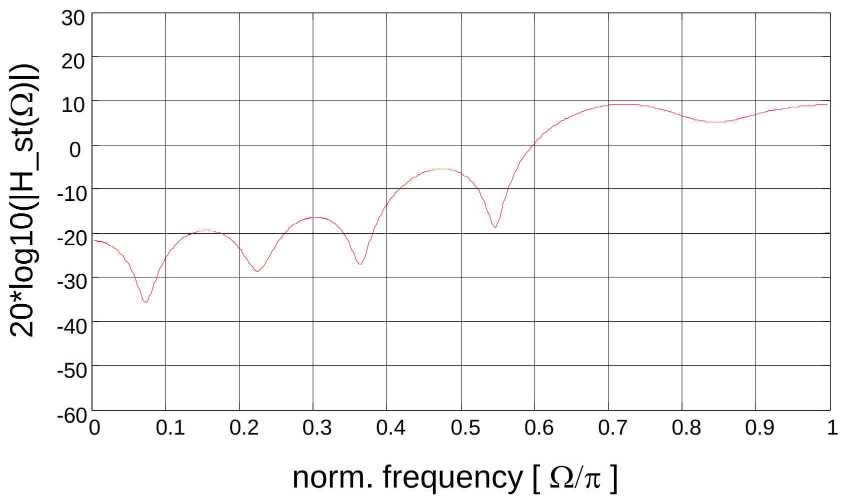
- **Common properties** of Hybrid coders:
  - Prediction coefficients are transmitted as side information.
  - Residual signal is quantized rather coarsely or even replaced by codebook vectors.
  - Typically, short and long-term prediction is used.

Signal Form Cdoder: Only error
Vocoder: only coefficients
Hybrid: error + coeff

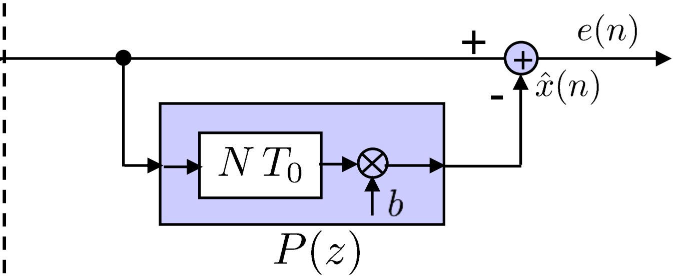
Hybrid Coder: Short- and Long-Term prediction

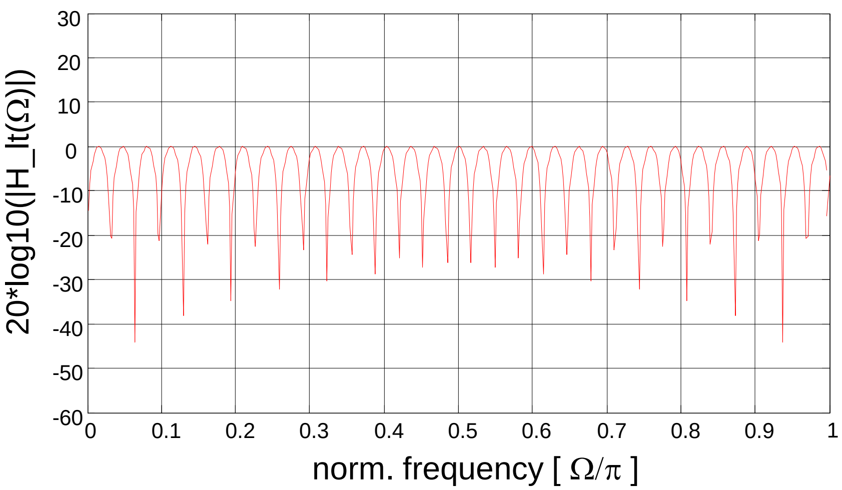
Short-Term prediction:





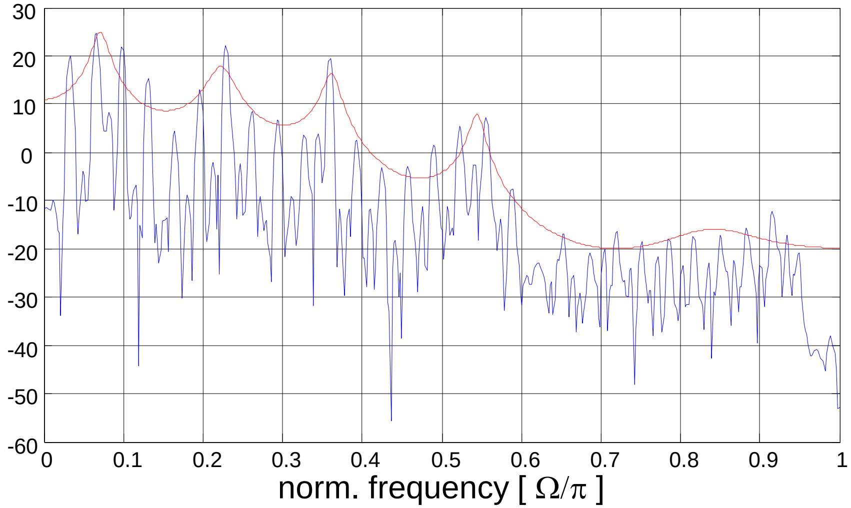
Long-Term prediction:



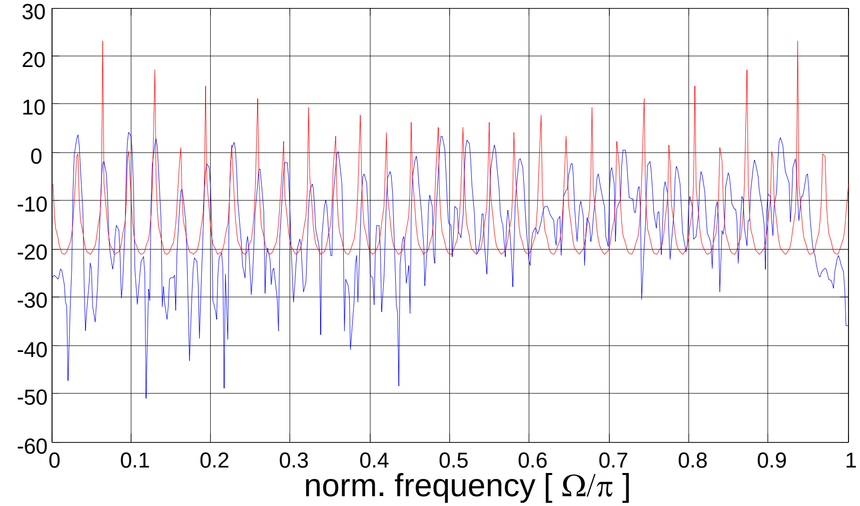


Hybrid Coder: Short (ST)- and Long-Term (LT) prediction

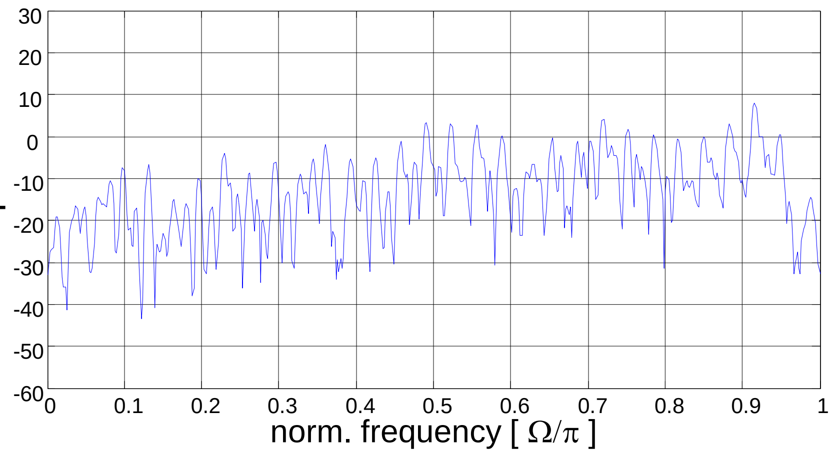
- Voiced speech spectrum + ST envelope



- Short-term residual spectrum + LT envelope



- Spectrum of residual signal after ST and LT prediction



Pitch not exactly periodic
=> worse approximation
for higher frequencies

Hybrid Coder

- **Quantization** of the residual signal:

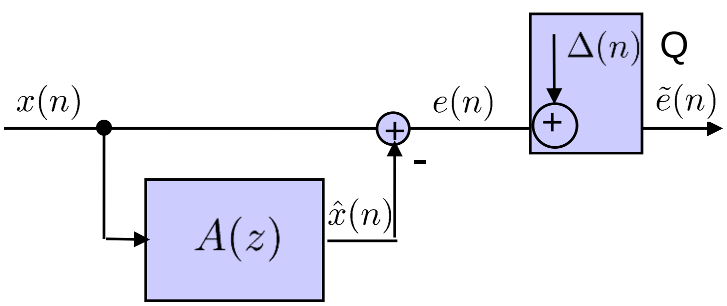
  - Scalar quantization (comparable to structures known from DPCM)

  - Vector quantization based on appropriate codebooks
    (for residual signal vectors).

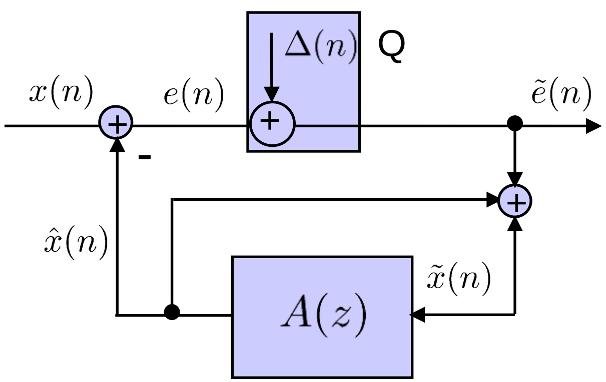
- Scalar quantization depends on the place of the quantizer:

  Closed or open loop structures:

General open-loop structure:



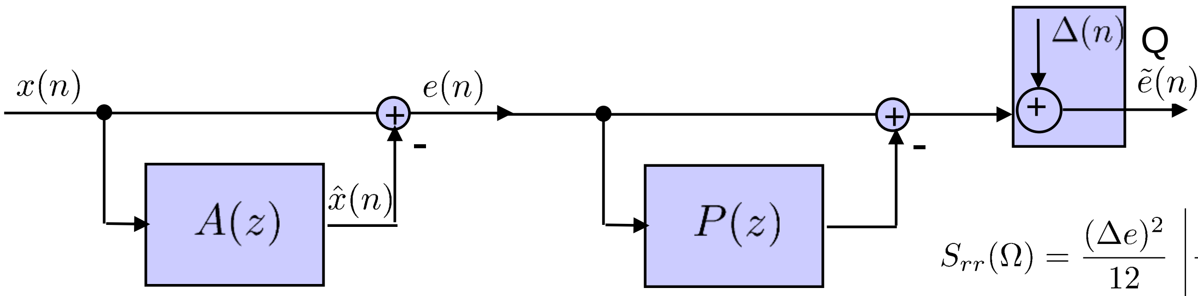
General closed-loop structure:



Combinations of open and closed loop structures in short- and long term prediction

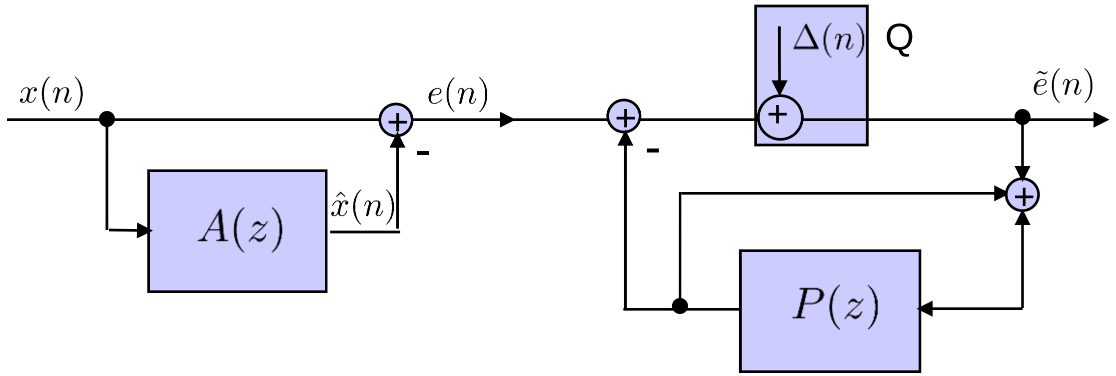
- Open loop structure for short- and long-term prediction:

$S_{rr}(\Omega)$: PSD of quant. noise after the decoder



$$S_{rr}(\Omega) = \frac{(\Delta e)^2}{12} \left| \frac{1}{\left(1 - A(e^{j\Omega})\right)\left(1 - P(e^{j\Omega})\right)} \right|^2$$

- Open loop structure for short term and closed loop structure for long-term prediction:

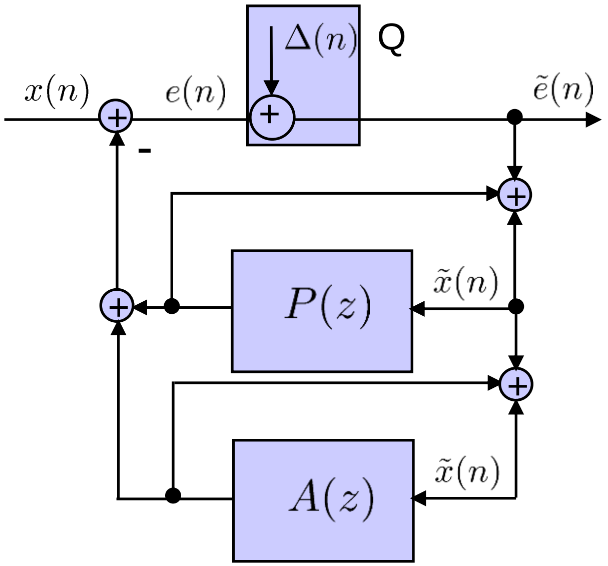


$$S_{rr}(\Omega) = \frac{(\Delta e)^2}{12} \left| \frac{1}{1 - A(e^{j\Omega})} \right|^2$$

Combinations of open and closed loop structures in short- and long term prediction

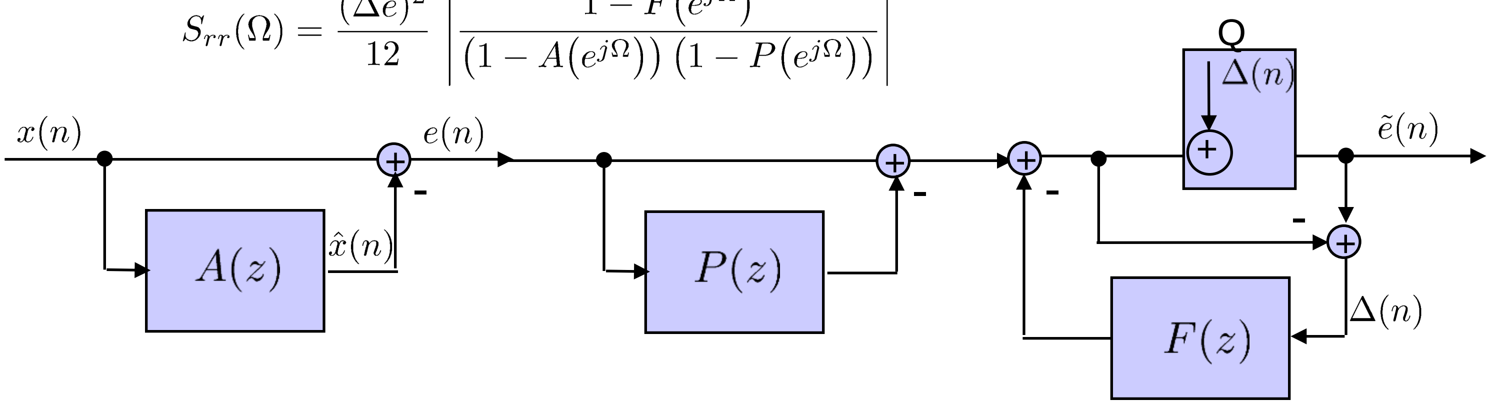
- Closed loop structure for short- and long-term prediction:

$$S_{rr}(\Omega) = \frac{(\Delta e)^2}{12}$$



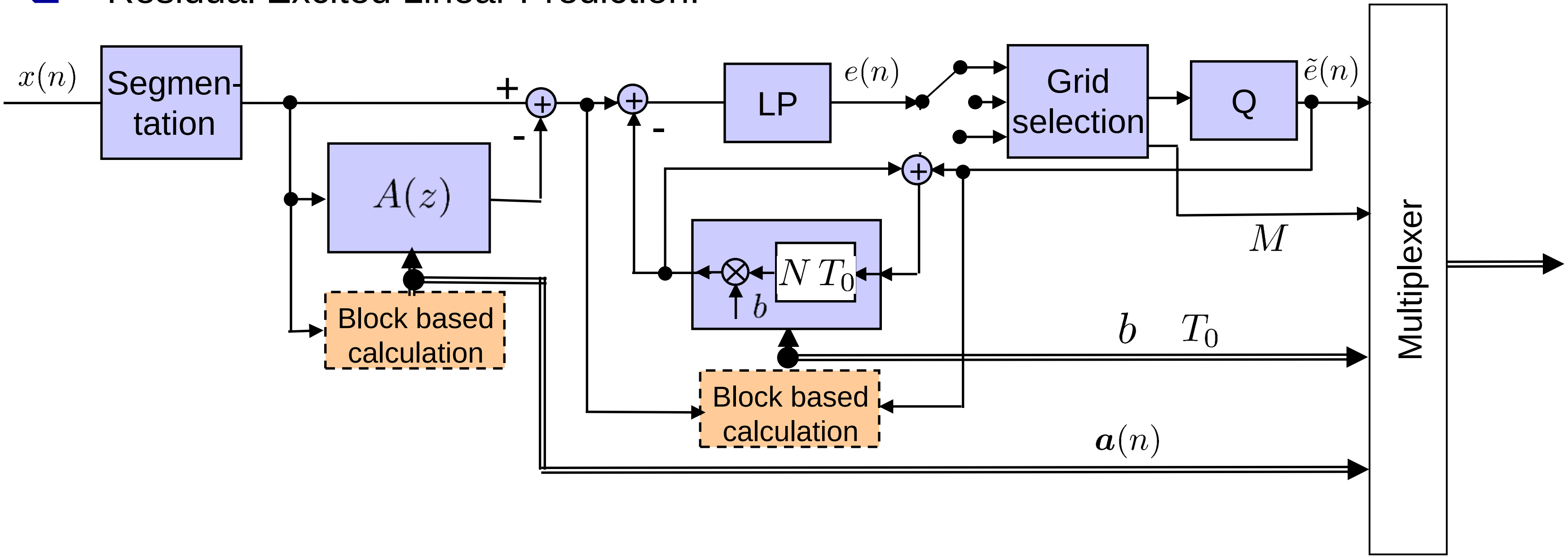
- Flexible noise shaping for open- and closed loop structure:

$$S_{rr}(\Omega) = \frac{(\Delta e)^2}{12} \left| \frac{1 - F(e^{j\Omega})}{\left(1 - A(e^{j\Omega})\right)\left(1 - P(e^{j\Omega})\right)} \right|^2$$



RELP coding: The GSM full-rate codec (ETSI GSM 06.10)

- Residual Excited Linear Prediction:

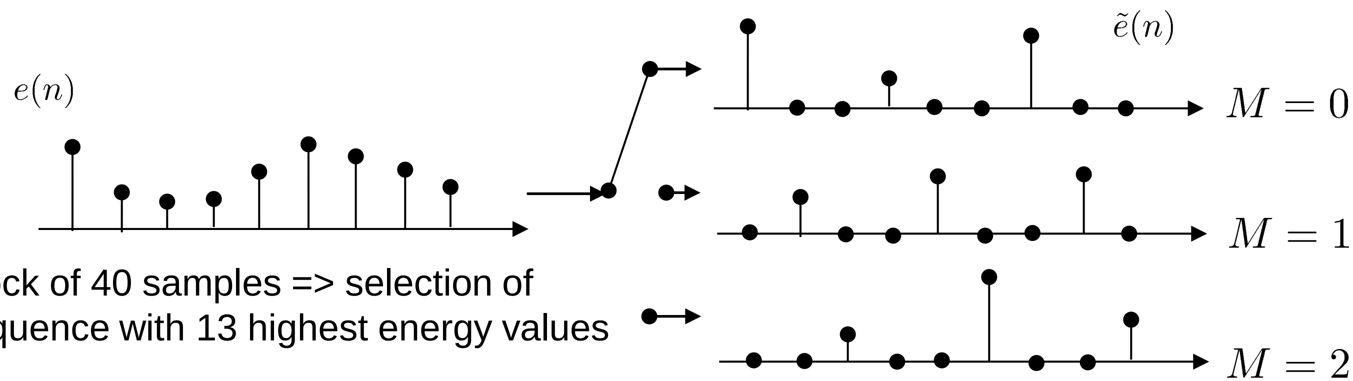


1. Segmentation in blocks of 20 msec (160 samples, fs = 8 kHz)
   Block based calculation of short term prediction A(z) with 8 coefficients

2. Segmentation: 4x40 samples (5 msec) for long term prediction and grid selection

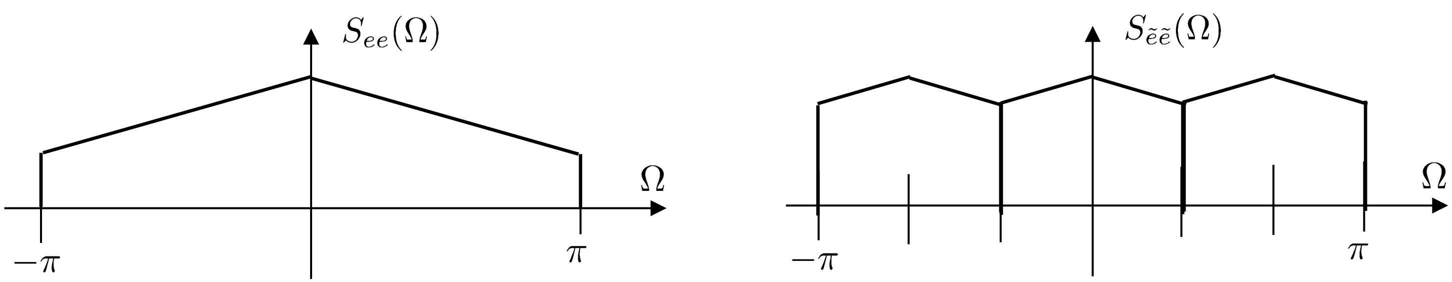
RELP coding: The GSM full-rate codec (ETSI GSM 06.10)

- **Grid selection:**
  Keep every third value, set other two to zero. Choice of highest energy sequence.
  This is equivalent to down- and upsampling without anti-imaging filtering:



Block of 40 samples => selection of sequence with 13 highest energy values

- Low frequency components are preserved, higher components are extrapolated.
  Assumptions: Nearly white PSD after predictor and lower sensitivity to higher components of the human ear.



RELP coding: The GSM full-rate codec (ETSI GSM 06.10)

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Bit coding scheme:**

  1) 8 prediction coefficients for **short-term prediction** every 20 msec:
  coded as (2x6 Bit, 2x5 bit, 2x4 Bit, 2x3 Bit) 36 Bit / 20 msec = **1.8 kBit / sec**.

  2) b and T_0 (2 Bit + 7 Bit) 9 Bit / 5 msec = **1,8 kBit / sec** for **long-term prediction**.

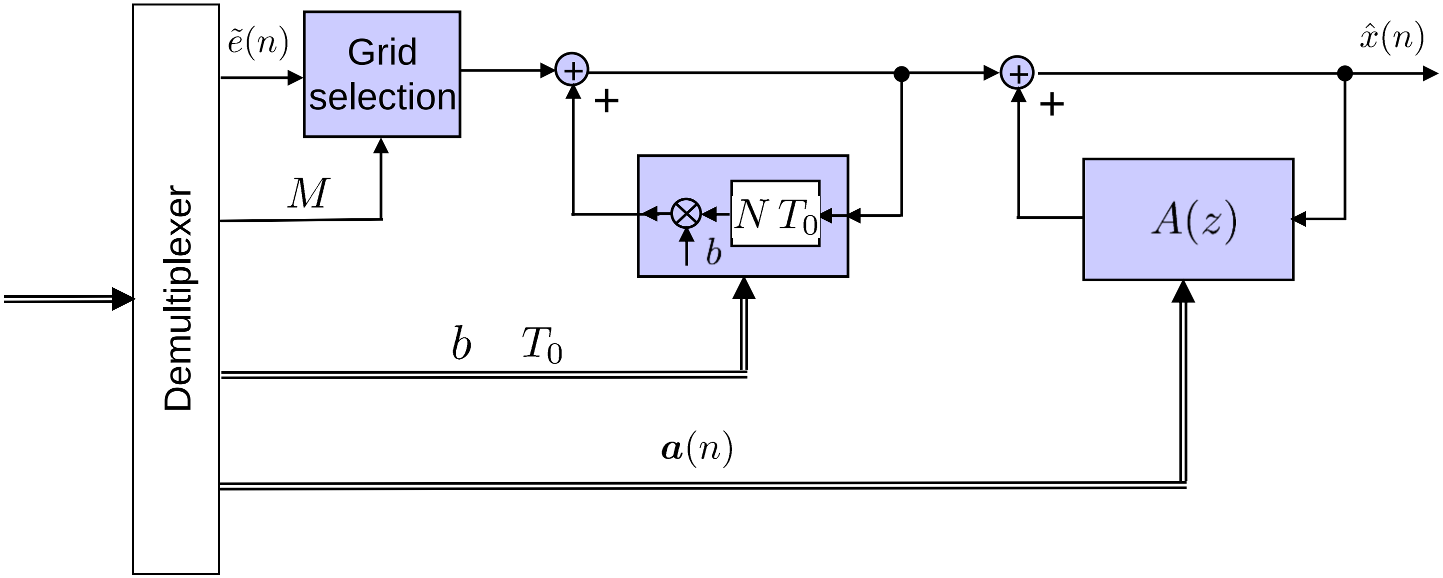
  3) **Residual signal:** 13 values (every third of 40 values). Coding with 3 Bit each after normalization. Normalization value (6 Bit) and sequence selection values M (2 Bit)
  => (13x3 + 6 + 2) Bit / 5 msec = **9,4 kBit / sec**

  => **total: 13 kBit / sec** => 13/8 Bit / sample = **1,65 Bit / sample**

RELP coding: The GSM full-rate codec (ETSI GSM 06.10)

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Decoder:**



Vector quantization

- Known optimization criterion for Vector Quantization:

$$\frac{1}{N}\sum_{n=0}^{N-1} \min_{i=0\ldots K-1} ||\boldsymbol{x}(n) - \boldsymbol{c}_i||^2 \rightarrow \min$$

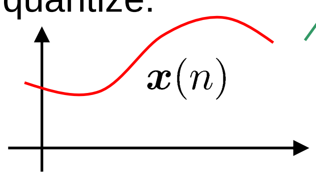
- Gain Shape Vector Quantization:

$$\frac{1}{N}\sum_{n=0}^{N-1} \min_{i=0\ldots K-1} ||\boldsymbol{x}(n) - g_i\,\boldsymbol{c}_i||^2 \rightarrow \min$$

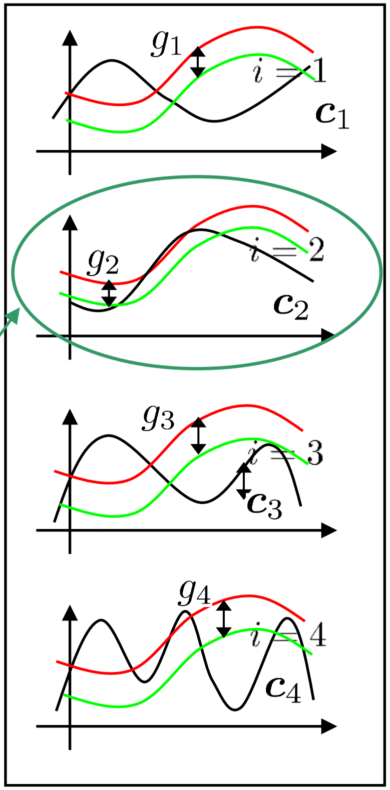
with: $g_i = \dfrac{\boldsymbol{x}^{\mathrm{T}}(n)\boldsymbol{c}_i}{\boldsymbol{c_i}^{\mathrm{T}}\boldsymbol{c_i}}$

- Concept: Code the form of the vectors, independent of the power.

Vector of residual signal to quantize:

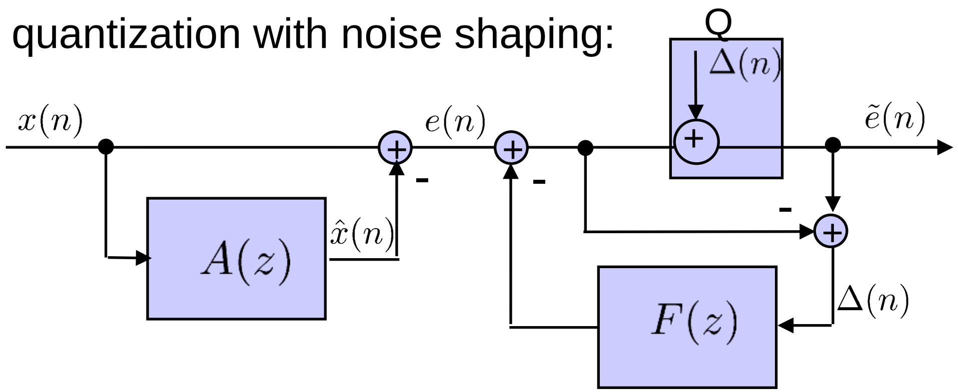


Codebook for residual signal:

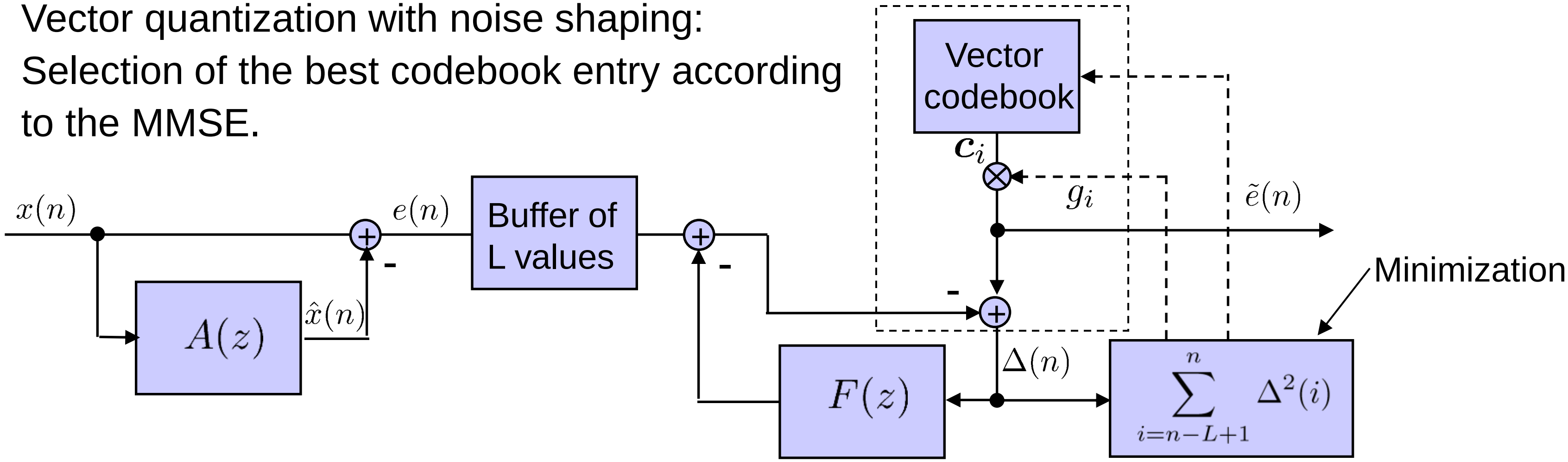


Vector quantization: CELP (code excited linear prediction)

- Scalar quantization with noise shaping:

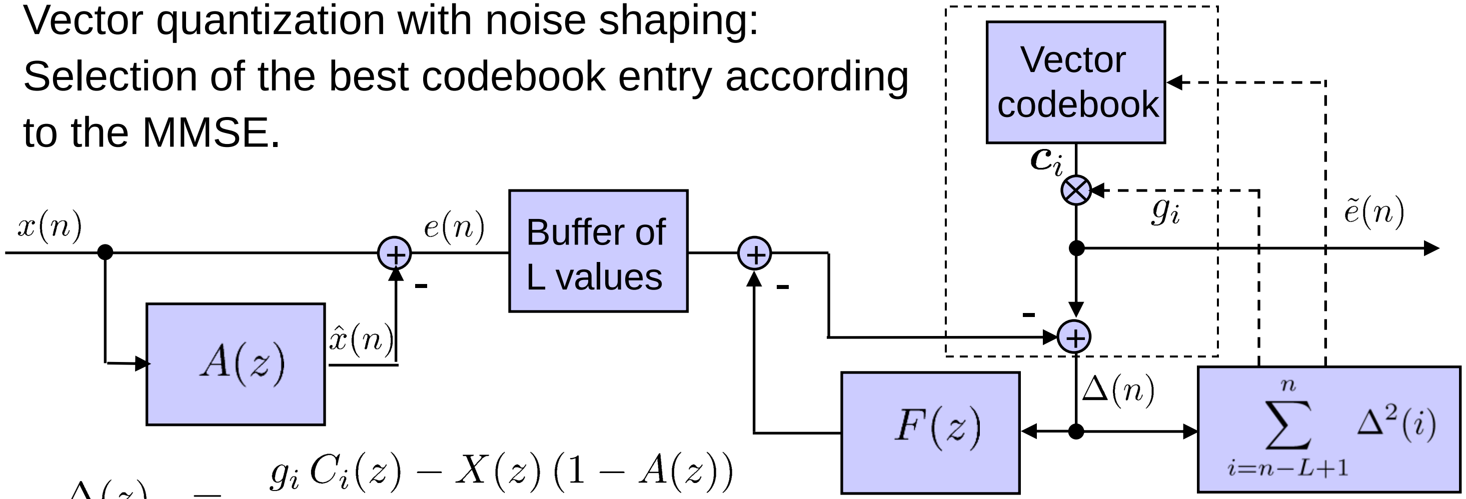


- Vector quantization with noise shaping:
  Selection of the best codebook entry according to the MMSE.



Vector quantization: CELP (code excited linear prediction)

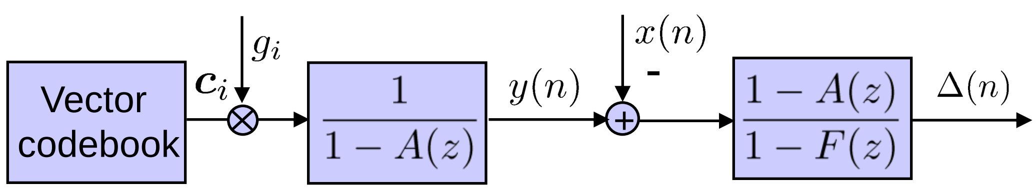
- Vector quantization with noise shaping:
  Selection of the best codebook entry according to the MMSE.



$$\Delta(z) = \frac{g_i\, C_i(z) - X(z)\,(1 - A(z))}{1 - F(z)}$$

$$= (Y(z) - X(z))\,\frac{1 - A(z)}{1 - F(z)} \qquad \text{with: } Y(z) = \frac{g_i\, C_i(z)}{1 - A(z)}$$

- Equivalent Description:



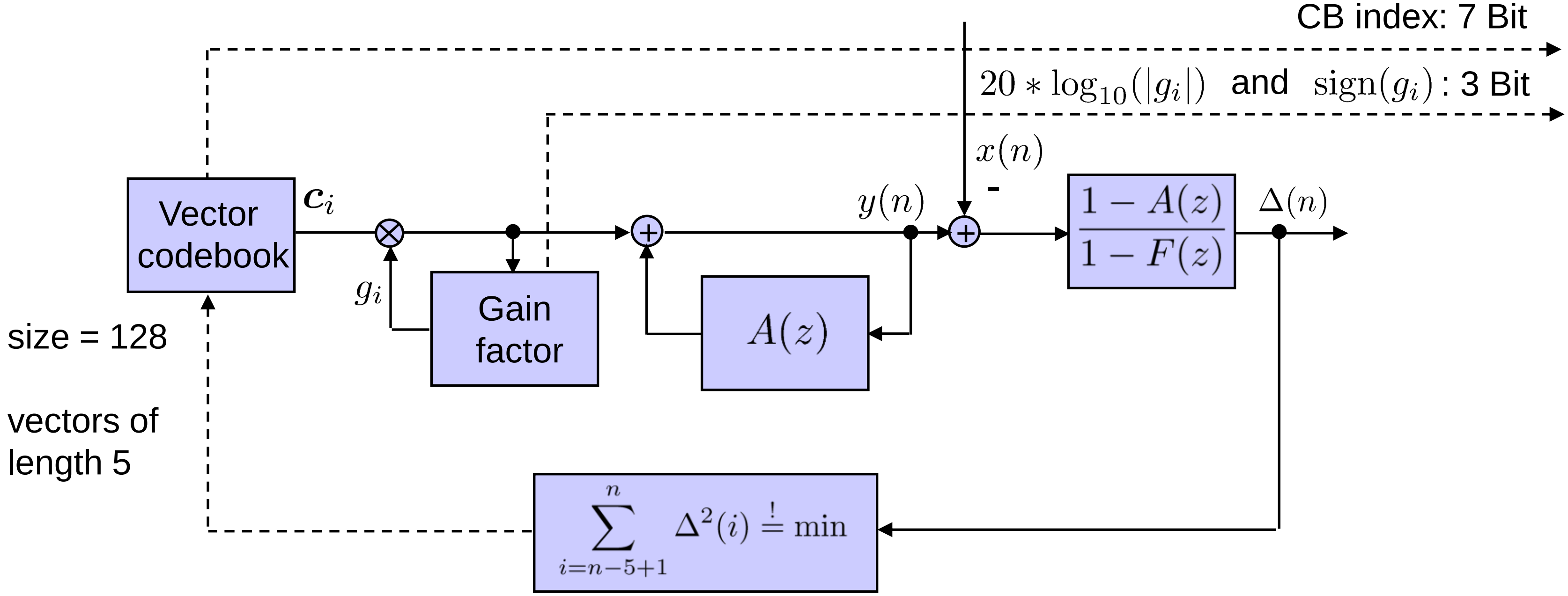
Synthesis Filter

Principle: Analysis-by-synthesis

=> Synthesize codebook output signal y(n).

=> Spectral weighting according to the inverse of the noise shaping filter.

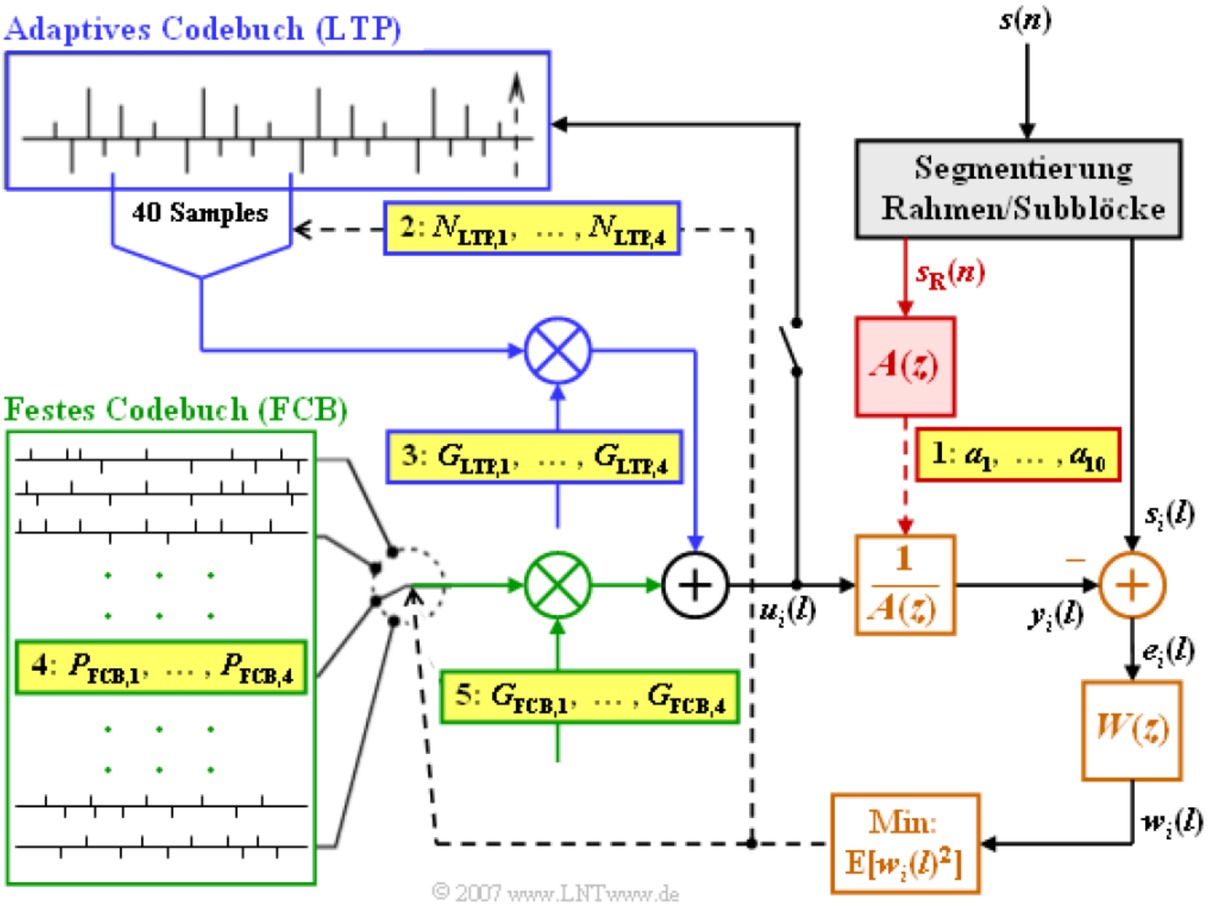
Low delay CELP coder: ITU-T G.728



- Predictor of order 50
- 10 Bit / 5 samples => 16 kBit / sec for 8 kHz sampled data
- No transmission of predictor coefficients

AMR (Adaptive Multi-Rate) Codec

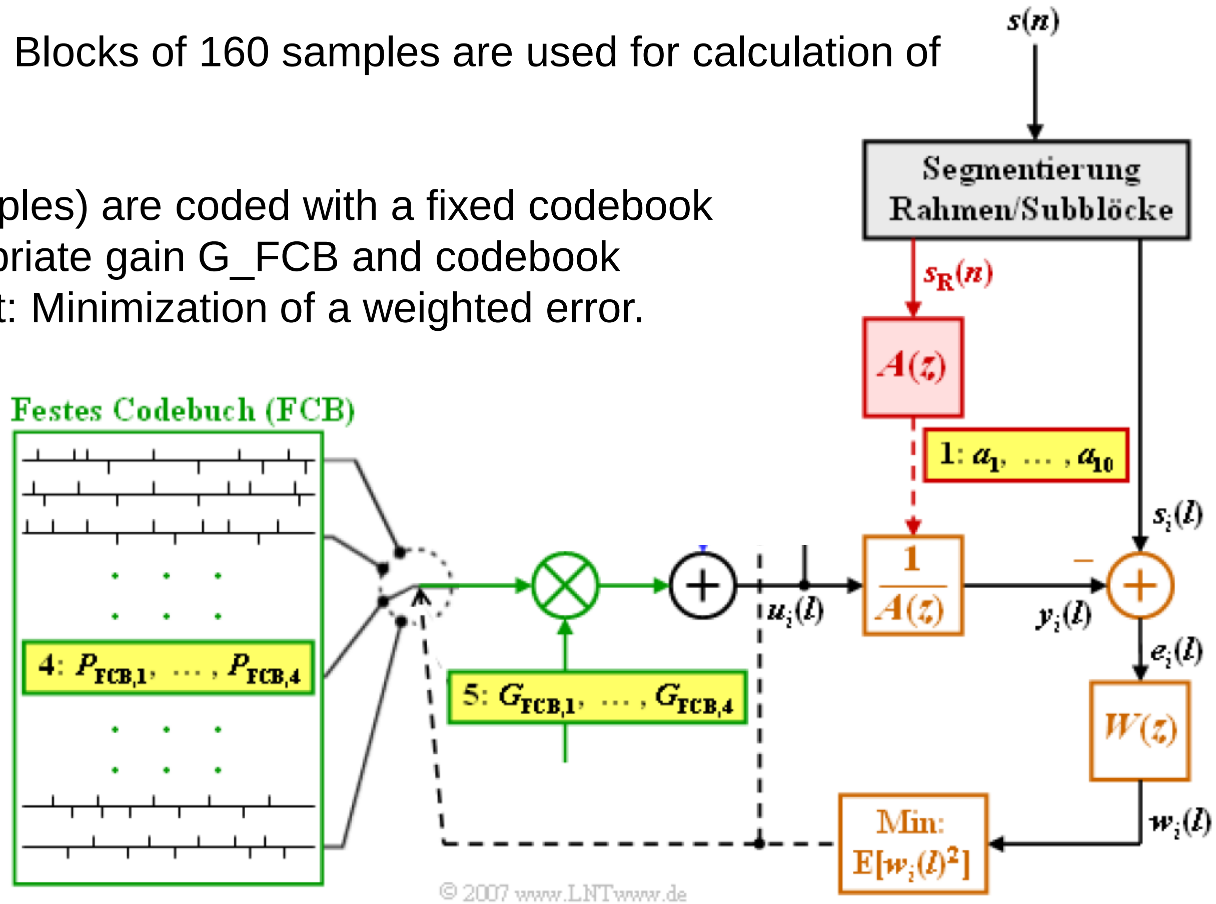
- **AMR block diagram (overview):**



AMR Codec

- **AMR block diagram:**
  - 1) Analog to GSM coder: Blocks of 160 samples are used for calculation of a prediction filter A(z).
  - 2) 4 Sub-blocks (40 samples) are coded with a fixed codebook by selecting an appropriate gain G_FCB and codebook indexes P_FCB Target: Minimization of a weighted error.

.



AMR Codec

TECHNISCHE UNIVERSITÄT DARMSTADT

❑ AMR block diagram:

❑ 3) **Fixed codebook:**

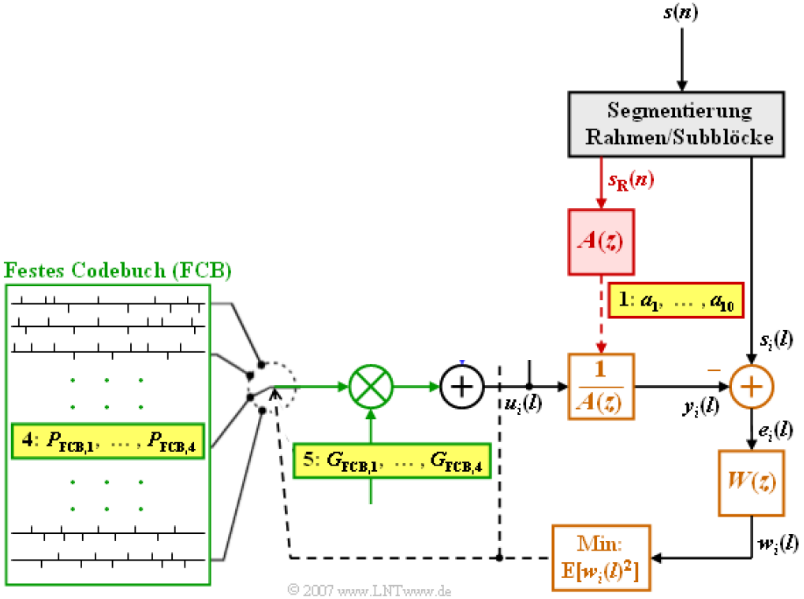
10 of 40 non-zero samples (+1/-1).

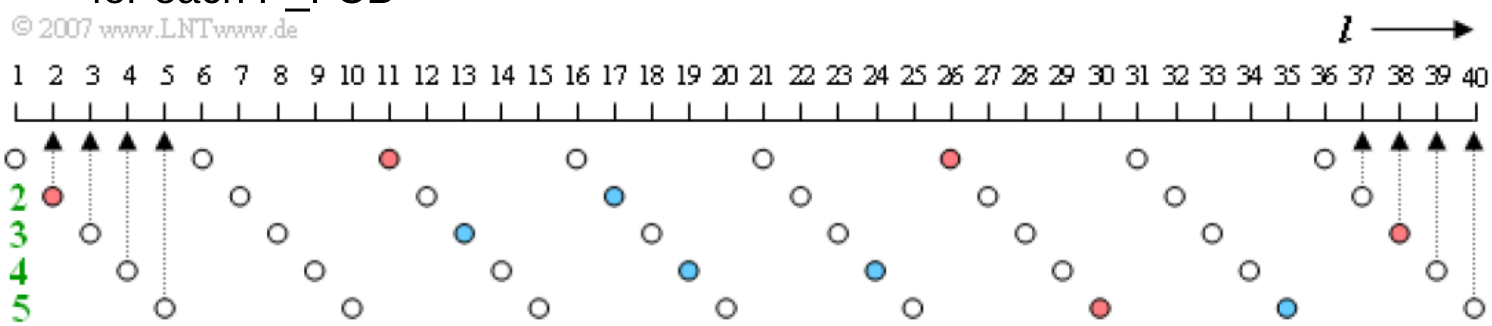
8 possible values in 5 lines (=> 3 bit for the position)

2 selected in 5 lines = 10x3 bit = 30 bit

Sign: 1 additional bit per line indicating the sign of the first value; rest coded by increasing or decreasing sample index

=> 30 + 5 = **35 bit** (for 40 sample blocks) for each P_FCB

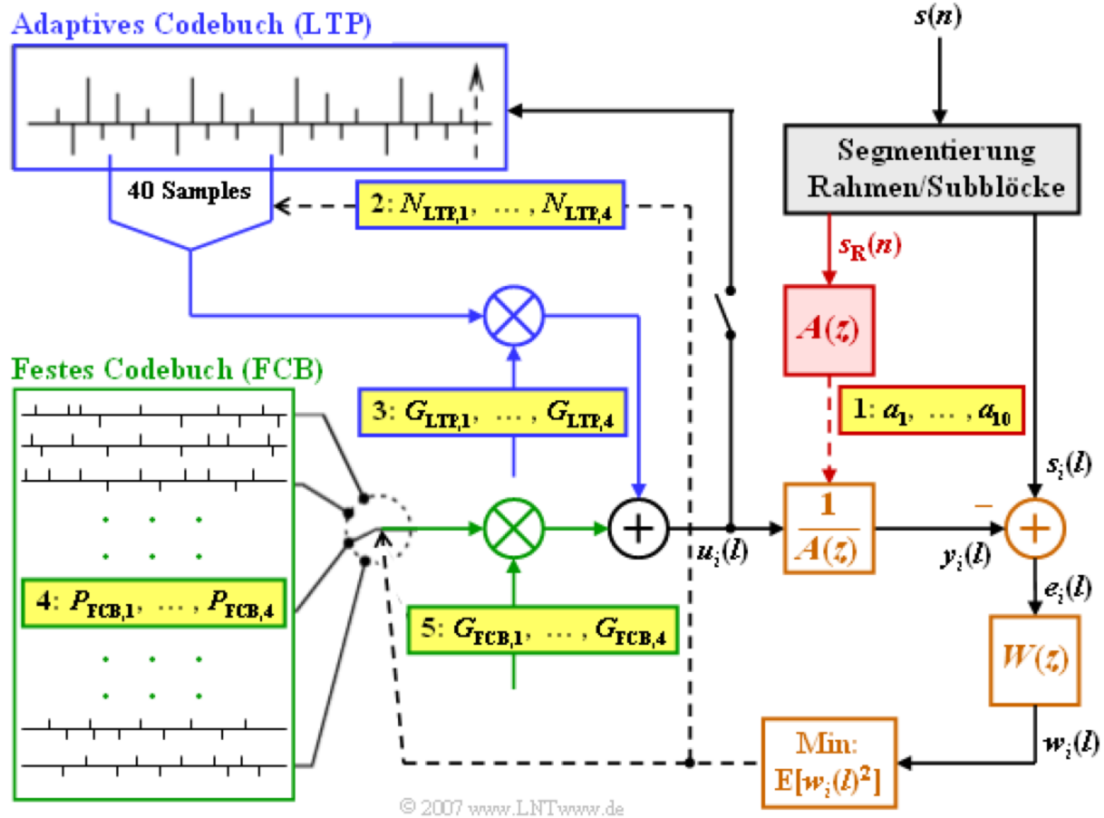




Example: indexes: 2, 5,   0, 3,   2, 7,   3, 4,   5, 6 (10 values, each 3 Bit)

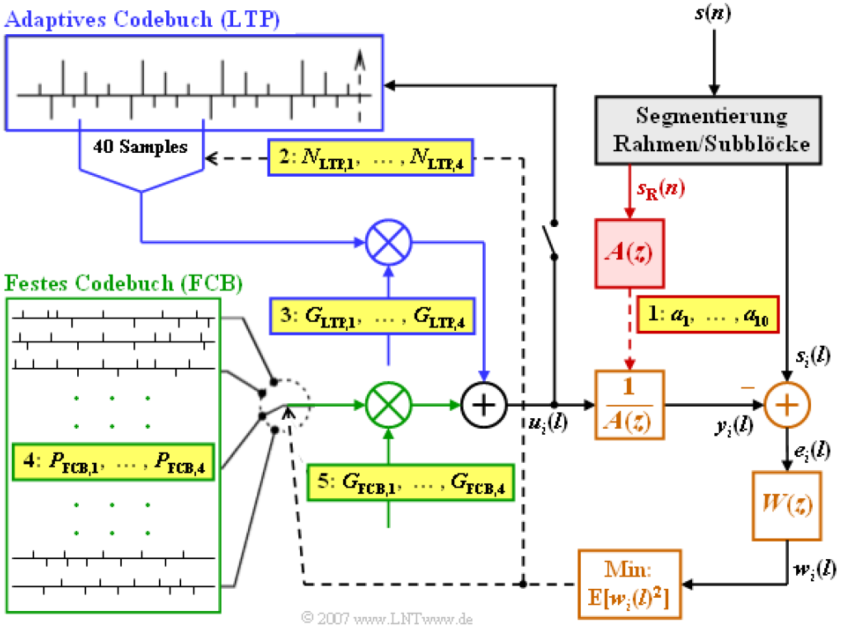
AMR Codec

- 4) **Adaptive codebook:**
  Long-term prediction.
  Codebook contains previous signal values.
    N_LTP: latency
    G_LTP: Gain.



AMR block diagram:

5) **Overall split of bits for the 12.2 kbit/s mode (244 Bit / 160 samples at 8 kHz):**



Coding scheme for 4x40 = 160 samples

AMR-Parameter	Bezeichnung	Modus 12.2 kbit/s
LPC-Filterkoeffizienten	$a_1, \ldots, a_{10}$	38
LTP-Verzögerung	$N_{LTP,1}, \ldots, N_{LTP,4}$	$9+6+9+6=30$
LTP-Verstärkung	$G_{LTP,1}, \ldots, G_{LTP,4}$	$4 \cdot 4 = 16$
FCB-Pulskennzeichnung	$P_{FCB,1}, \ldots, P_{FCB,4}$	$4 \cdot 5 \cdot 7 = 140$
FCB-Verstärkung	$G_{FCB,1}, \ldots, G_{FCB,4}$	$4 \cdot 5 = 20$
Gesamt		244 Bit

Adaptive codebook

Fixed codebook

Source:

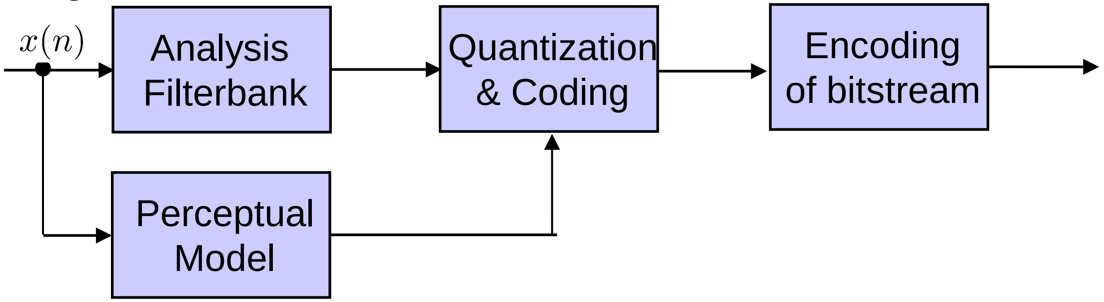
https://www.lntwww.de/Beispiele_von_Nachrichtensystemen/Sprachcodierung

Frequency domain or sub-band coders:

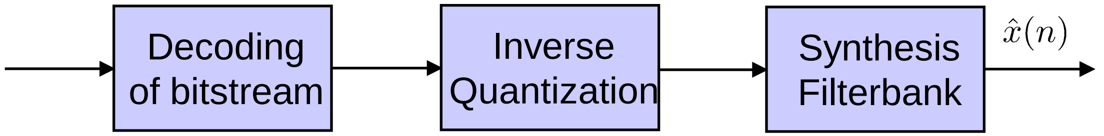
Processing in the frequency domain / frequency sub-bands
=> Explores psycho-acoustic masking
=> Introduces a higher processing latency

MP3 Coding

- MP3 Coding stands for MEPG1/2-Layer3:
- Principle: "Perceptual Audio Coding"
  - Uses psychoacoustics, i.e., masking to reduce the bit rate.
  - Quantization of the specific frequency bands according to masking thresholds.
- Encoding:

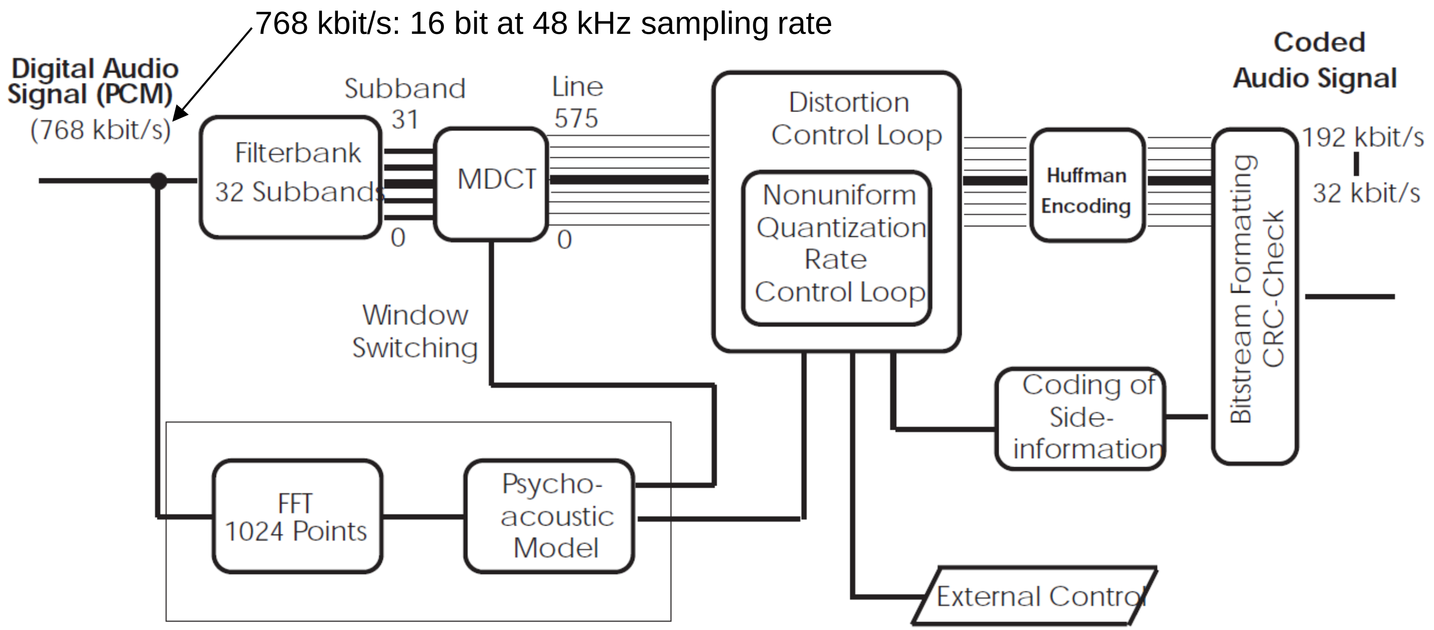


- Decoding:



MP3 Coding

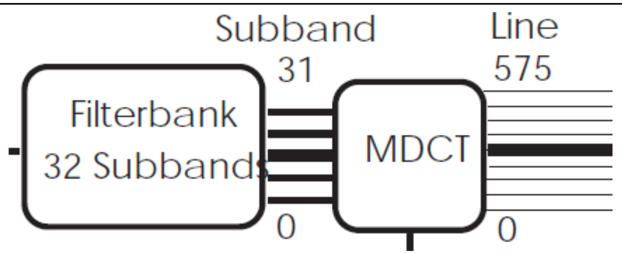
- MP3 Coding Block diagram [2]:



- Flexible sampling frequencies:
  32 / 44.1 / 48 kHz sampling rates and also half rates
  in MPEG 2, i.e., 16 / 22.05 / 24 kHz

Frequency decomposition

- Two-stage Filterbank:



- First stage: Decomposition into 32 frequency bands by a polyphase filter-bank. Filter length: 512 taps. => latency **511 samples** (at 48 kHz). Followed by a subsampling of 32.

- Second stage: - MDCT (modified discrete cosine transform). Further decomposition into **18 bands** of each of the 32 polyphase bands (Filter length: 36 taps).
  => 32*18 = 576 freq. bands in total (Followed by a subsampling of 18).
  - For **transient signals only 6 bands** => 12 samples look-ahead for the selection of the appropriate resolution.
  - Latency: 32*(35+12) = **1504 samples** (at 48 kHz).

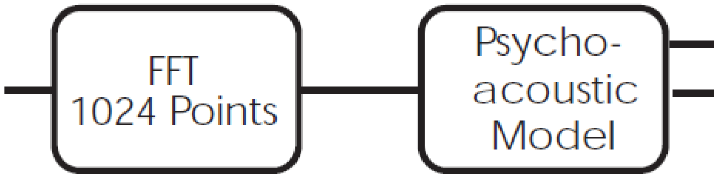
Subsampling of 1. stage
Look ahead
Filter delay (36 samples filter)

- Overall latency: 511 + 1504 + 18*32 = 2591 samples (= **54 ms**)

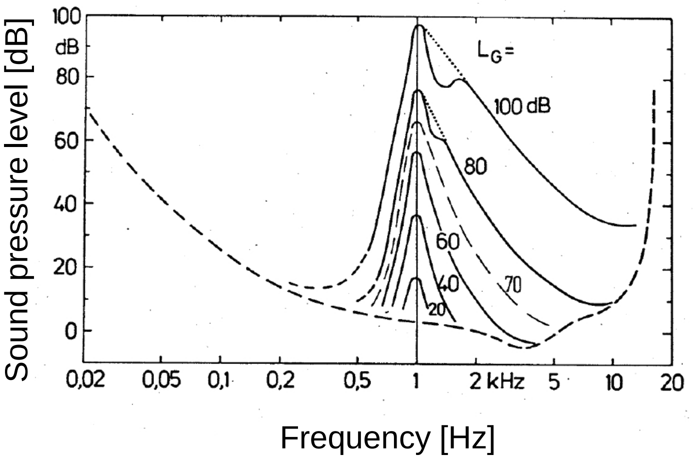
Coding of two blocks together (block size: 18*32 = 576)

Psychoacoustic model

- Calculation of masking thresholds:



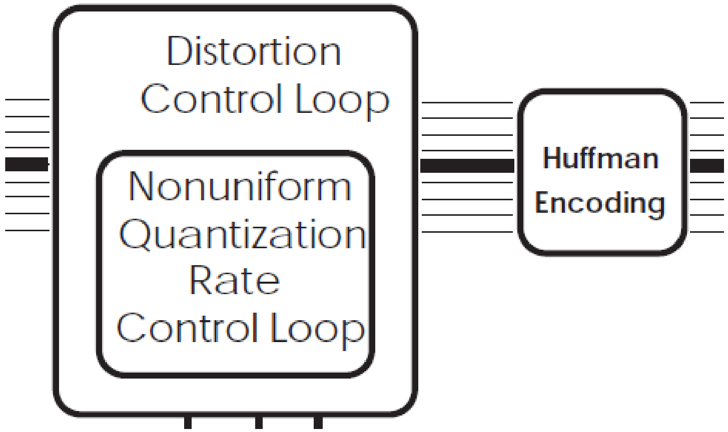
- Frequency masking:



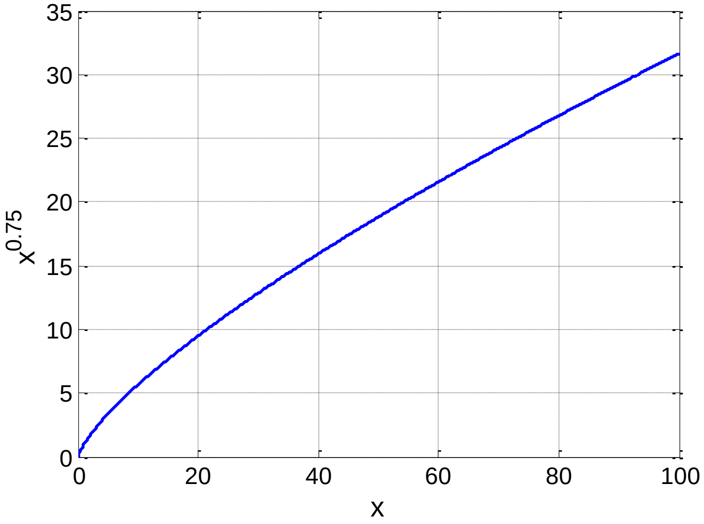
- Result: Determines the allowed quantization noise in each frequency band which is masked by signal excitation.

Quantization & Coding

- Quantization in two steps:



- 1) Power-law quantizer:



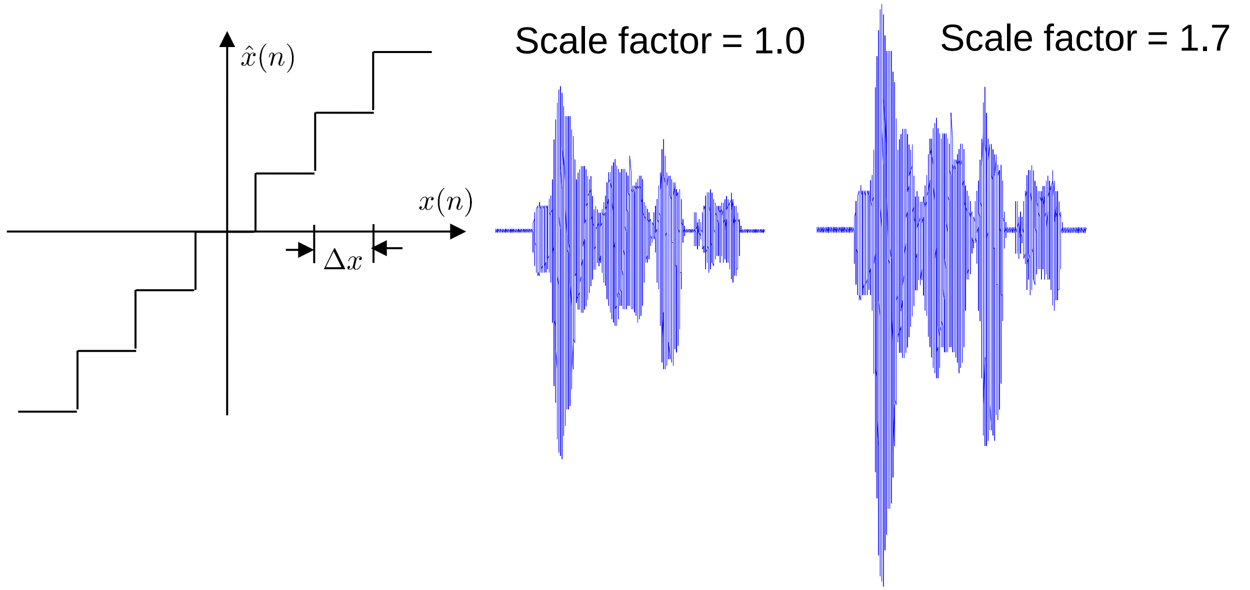
- 2) Definition of an **overall gain factor** for all frequency bands
  => adjusts the number of coded bits.

- 3) Definition of **scale factors** for each band to not exceed the quantization
  noise in the respective band.

- => There is an iteration performed between those definitions to not exceed the
  overall number of coded bits.

Gain / scale factors

- The higher the overall gain and the band selective scale factors:
  => the higher is the number of bits and the less the quantization noise (relative to the signal power):



Huffman Coding

- **Principle:**
  - Huffman coding is one example of Entropy Coding.
  - The higher the probability of a symbol or a value to code, the lower should be the number of bits to code the symbol or value.
  - Target: minimization of the mean codeword length.
  - Example:

Selection of the best of several possible tables (side information necessary)

codeword Length	codeword	X	P(X)
3	110	1	0.15
3	111	2	0.15
2	00	3	0.2
2	01	4	0.25
2	10	5	0.25

X: Symbol to code

P(X): Symbol probability

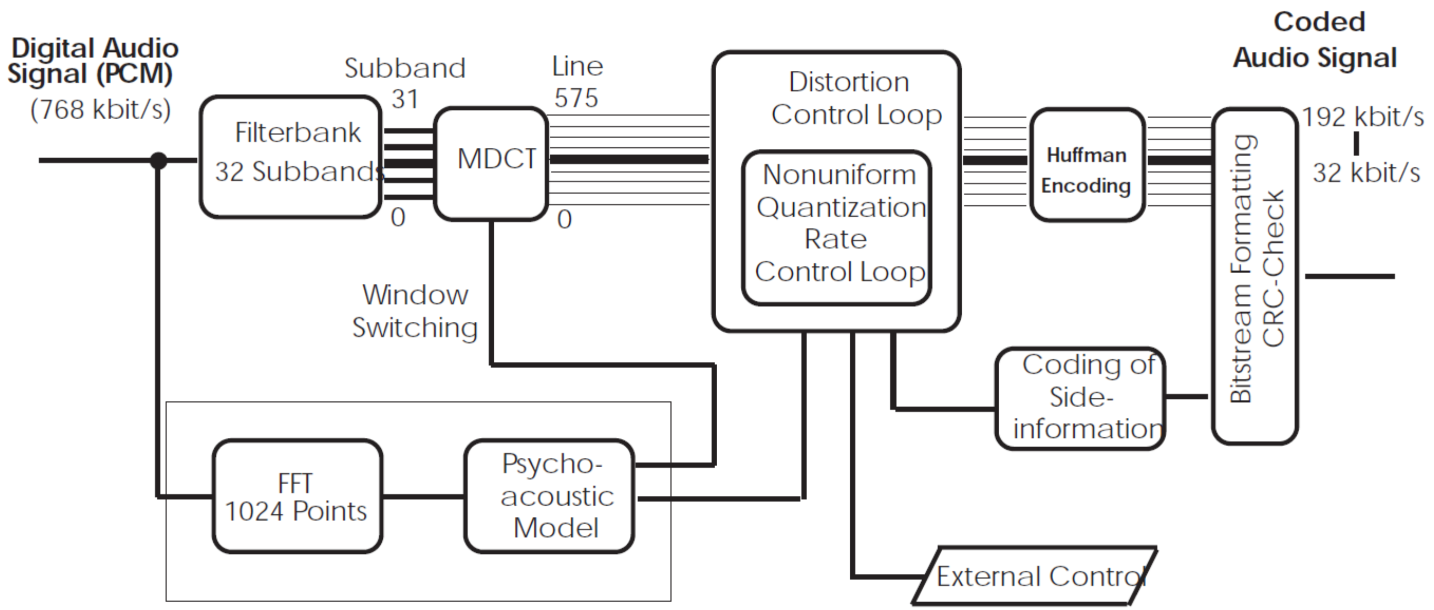
w_i Codeword length

Mean codeword length:

$$L = \sum_{i=1}^{N} P(X_i) * w_i \quad = 2.3 \text{ Bit}$$

MP3 Coding

- MP3 Coding Block diagram:



- Side information:
  e.g., overall gain, scale factors, Huffman table index.

MPEG2 – AAC (Advanced audio coding)

- Modifications compared to mp3:

  - 1) **Higher frequency resolution:**
    - One stage MDCT (mod. discrete cosine transform) with 1024 frequency bands and a 2048 filter length (compared to 18*32 = 576 in MP3)
    - Switching to 128 MDCT frequency bands and a 256-filter length for transient signals

    => Switching look ahead necessary (to decide about the mode) equal to 576 samples.

    => **Latency:** 2047 + 576 samples = 2623 samples => **54.6 ms** (fs = 48 kHz)

  - 2) **Frequency domain prediction:**
    Prediction over time, independent in each frequency band.

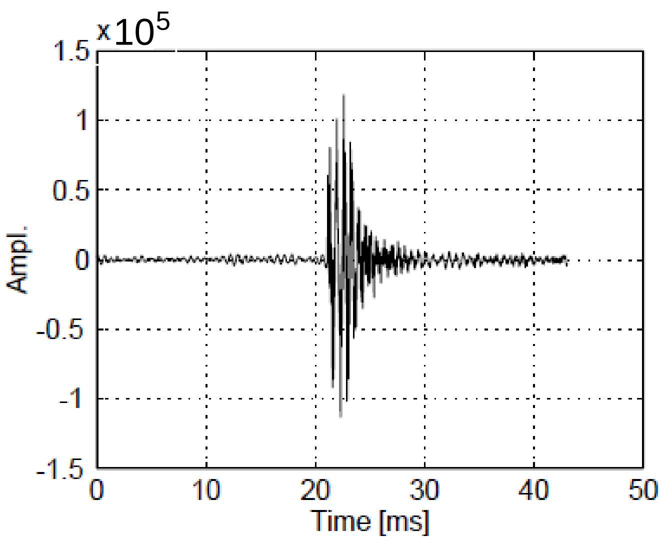
  - 3) **Joint stereo coding:**
    Mid-side coding:
    Coding of sum and difference of the left and the right channel
    Advantage: Quantization noise is correlated in the left and right decodes signals and as such perceived from the frontal direction where is the origin of the main intense signals.

MPEG2 - AAC

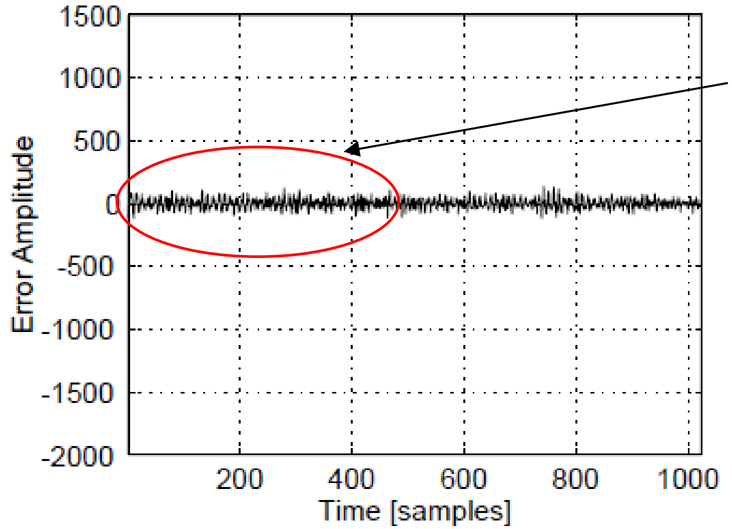
- **AAC: Advanced audio coding.**

- Modifications compared to mp3:
  - 4) **TNS: Temporal noise shaping:**
    Typical problem of frequency domain prediction / quantization:
    Pre-echoes, i.e., Spread of quantization noise over a complete signal block

  - Transient signal to code:



Spread of quantization noise over time:



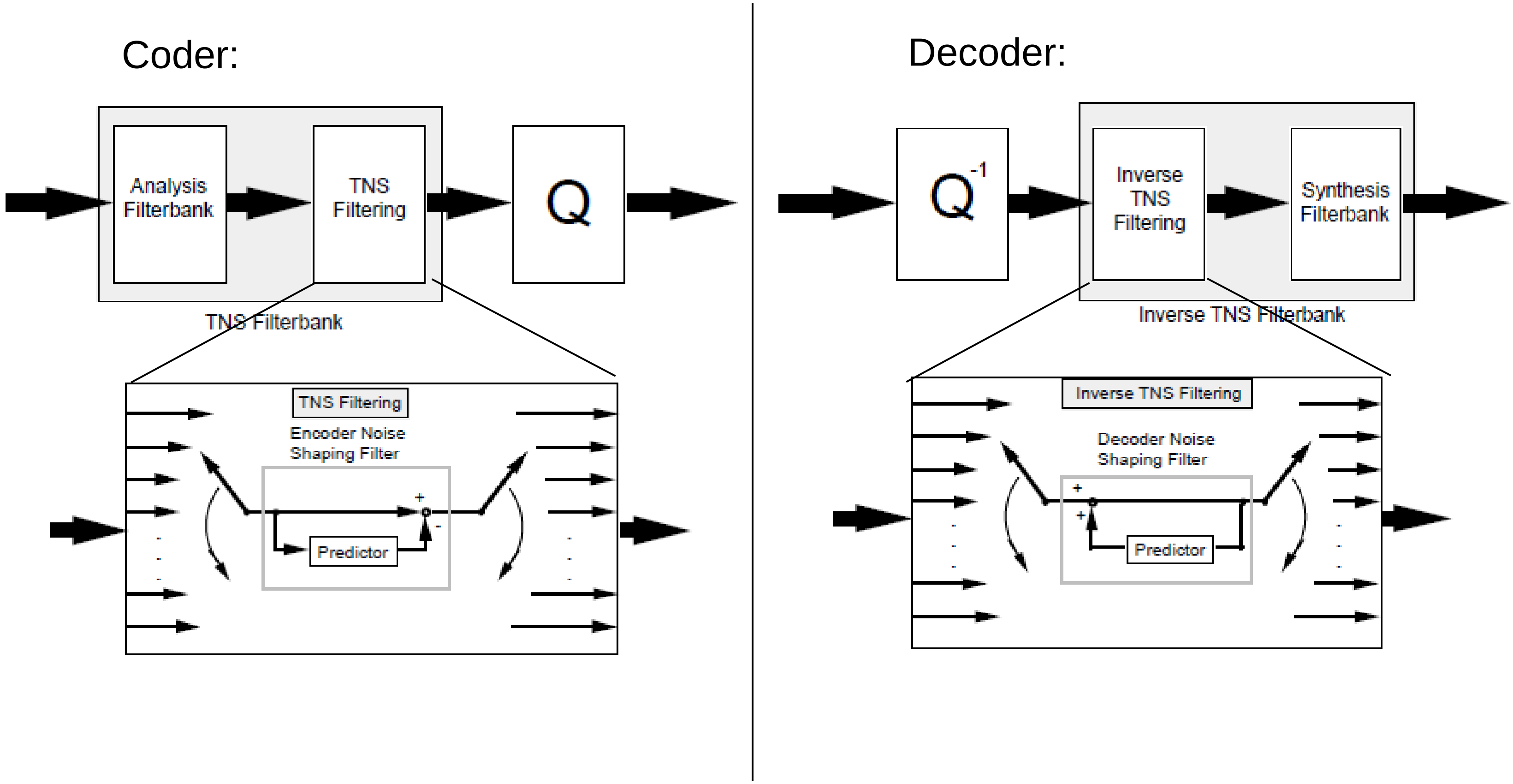
High quantization noise before the transient signal => pre-echoes

TNS: Temporal noise shaping

- **Concept of TNS:**
  - Shape the time domain coding noise according to the time shape of the signal to quantize.
  - Consider the time / frequency duality:
    - Time domain prediction (open loop structure)
      => Spectral domain: Noise shaped according to the signal spectral shape.
  - => Concept: Apply a prediction over the frequency values
    => Time domain quantization noise shape is according to the time domain shape of the quantized signal.

TNS: Temporal noise shaping

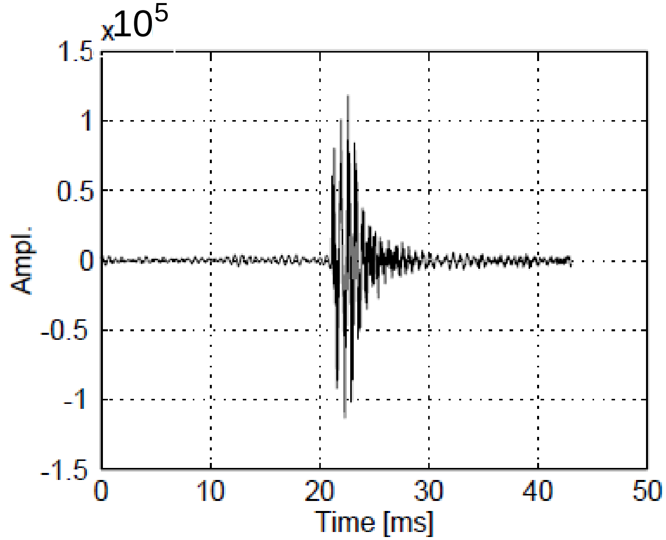
Concept of TNS:



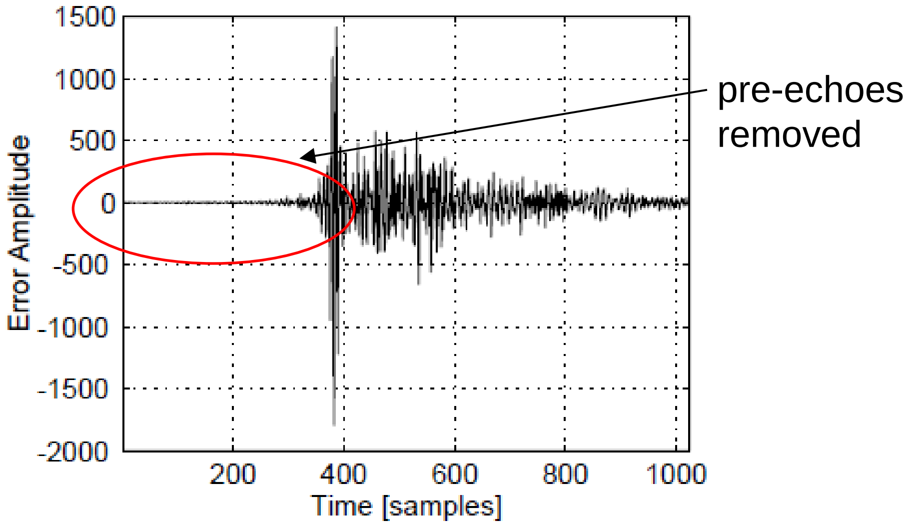
TNS: Temporal noise shaping

- **Results:**

Transient signal to code:

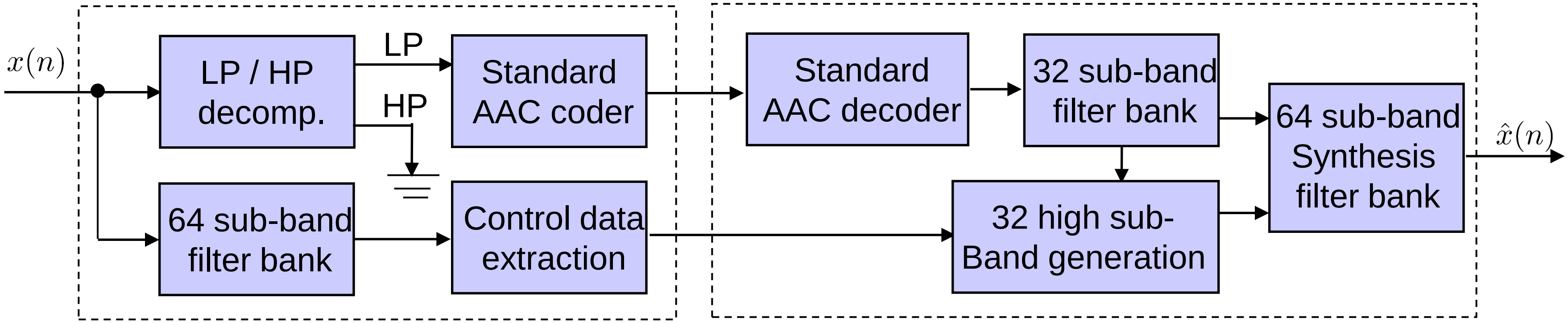


Quantization noise spread according to the signal to code => masking!



MPEG4 – High Efficiency AAC (HE AAC)

- **Principle**:
  - Core AAC running in the lower half of the frequency range (after LP filtering and sub-sampling by a factor of 2).
  - Missing high frequency components are recovered by a "spectral band replication" (SBR).
  - SBR comparable to bandwidth extension. SBR uses the spectral content of the low frequency components and some "control data", i.e., information of the spectral high frequency content, which is extracted before sub-sampling.
  - Computational complexity reduction due to AAC running at ½ rate.



  - Latency: 2623 + 288 + 192 = 3102 samples (at 24 kHz) => **129 ms**.

6*32 samples look ahead for HF signal recovering

32 sub-band analysis in the decoder

MPEG4 – AAC Low Delay (AAC LD)

- **Principle**:
  - Several modifications to reduce standard AAC latency to 20 ms:
  - 1) One stage MDCT with
    - either 480 frequency bands and a 960 filter length
    - or 512 frequency bands and a 1024 filter length
  - 2) No switching to a lower frequency resolution for transient signals
    => no look ahead necessary.
  - => Overall latency is 959 (= **20 ms**) or 1023 (= **21.3 ms**)

References

[1] K. Brandenburg, O. Kunz, A. Sugiyama: „MPEG-4 natural audio coding“, Signal Processing: Image Communication 15 (4-5) (2000) 423-444.

[2] K. Brandenburg : “MP3 and AAC Explained”, AES 17th Int. Conf. on High Quality Audio Coding, 2012

[3] J. Herre: “Temporal Noise Shaping, Quantization, and Coding Methods in Perceptual Audio Coding: A Tutorial Introduction”, AES 17th Int. Conf. on High Quality Audio Coding, 2012

[4] M. Lutzky et. al.: “A guideline to audio codec delay”, Audio Engineering Society 116th Convention, Berlin, Germany, 2004

Summary & Outlook

TECHNISCHE UNIVERSITÄT DARMSTADT

- Extensive view on audio coding schemes.
- Target of all audio coding schemes:
  - Remove redundancy of the signal which is coded.
  - Transmit only the relevant information.
- Coding schemes are based on prediction error filtering:
  - Signal form coder
  - Vocoder
  - Hybrid coder
- Sub-band coding schemes:
  - MP3 and AAC coding methods
- **Next week:** Noise reduction and dereverberation.