Speech Encoding I.

Jan Černocký, Valentina Hubeika

{cernocky,ihubeika}@fit.vutbr.cz

DCGM FIT BUT Brno

Agenda

- Groups of encoders
- Quality measure
- Waveform coding
- Vocoders
- Vector quantization

Why ?

Most of the communication nowadays is *digital*.

- lowest possible bit-rate.
- highest possible quality.
- lowest possible delay.
- highest possible robustness to errors.
- lowest possible computation cost.

... some criteria are in contradiciton, thus need to tradeoff.

Encoding — simplest commercial subfield of ASP (automatic speech processing)

Standardization

- CCITT (Centre Consultatif International Téléphonique et Télégraphique), from which was found ITU-TSS (International Telecommunication Union — Telecom. Standardization Sector). Reccomendations Gxxx. `http://www.itu.int`
- US DoD (Department of Defense). Federal standards FSxxxx.
- ETSI (European Telecommunications Standards Institute), mobile telephony. `http://www.etsi.org`
- and other institutions, e.g. INMARSAT.

Groups of coders – Principal division

- **waveform coders** - sample by sample, high quality, not only for speech signals.
- **vocoders** facilitates findings on human speech production and perception (excitation and modification). Working in frames, where the signal is considered stationary and mostly making use of linear-predictive (LP) model. Good only for speech.
- **Hybrid** - alias for the CELP algorithms (GSM). Model based representation of modification, waveform encoding of excitation.
- **phonetic vocoders** – longer speech segments, contain units longer than frames (phonemes, automatically trained units). Based on recognition in the encoder and synthesis in the decoder. Laboratory prototypes. Speech signal only, language (speaker) dependent.

Division with Respect to Bit-Rate

Bit rate = bits per second (source coding, channel coding).

- fixed-rate (standard scheme)
- variable-rate (buffering, re-allocation of channel capacity among speech and data channels).

notation	bit rate
high rate	$>$ 16 kbit/s
medium rate	8 – 16 kbit/s
low rate	2.4 – 8 kbit/s
very low rate	$<$ 2.4 kbit/s

Division with Respect to Quality

- **broadcast** – better than analog telephony bit-rate $>$ 64 kbit/s.
- **network** or **toll** – quality of the analog telephone, band of 300–3400 Hz.
- **communications** – intelligible and preserving speaker characteristics.
- **synthetic** – unnatural, does not preserve speaker characteristics, lower intelligibility.

Quality Evaluations

- objective
- subjective

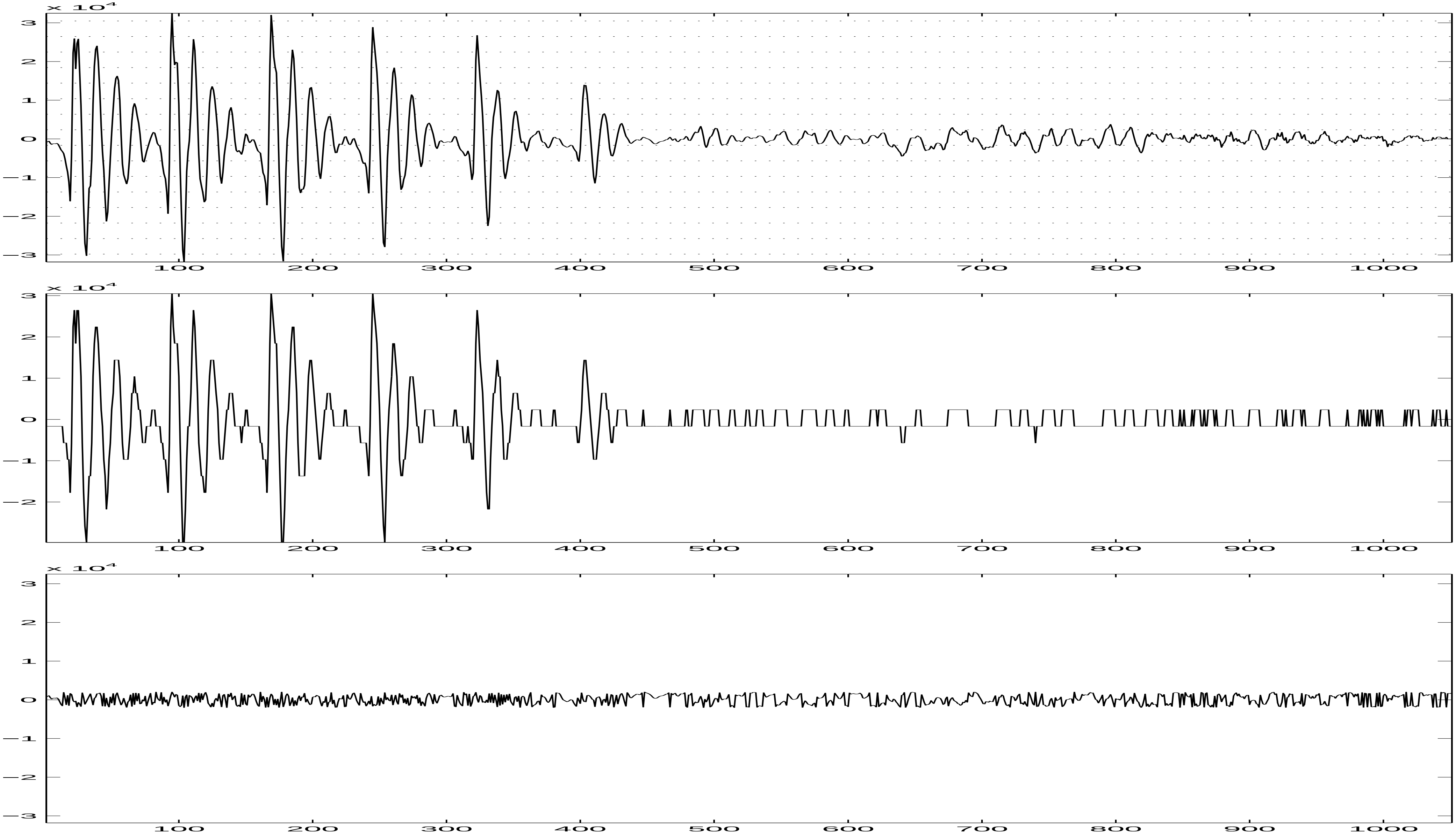
Objective Evaluations

signal-to-noise ratio SNR:

$$SNR = 10\log_{10}\left\{\frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1}[s(n)-\hat{s}(n)]^2}\right\}$$

Drawback: considers the global signal.

Example: encoding of the syllable "as" on 4 bits (16 quantization levels) SNR = 14.89 dB.


x 10^4
3
2
1
0
−1
−2
−3
100
200
300
400
500
600
700
800
900
1000
x 10^4
3
2
1
0
−1
−2
100
200
300
400
500
600
700
800
900
1000
x 10^4
3
2
1
0
−1
−2
−3
100
200
300
400
500
600
700
800
900
1000

Segmented Signal to Noise Ratio (SEGSNR)

$$SEGSNR = \frac{10}{N_{ram}} \sum_{i=0}^{N_{ram}-1} \log_{10} \left\{ \frac{\sum_{n=0}^{l_{ram}-1} s^2(il_{ram}+n)}{\sum_{n=0}^{l_{ram}-1} [s(il_{ram}+n) - \hat{s}(il_{ram}+n)]^2} \right\} \qquad (1)$$

Here, SNR is calculated separately for each frame (for the "classical" frame length of 160 samples):

20.04 19.63 14.35 0.21 4.26 -0.54(!)

and SEGSNR (average value) is **9.66 dB**, the obtained result is substantially worse but more informative than in case of SNR.

Logarithmic Spectral Distance

evaluates the distance between speech frames in *spectral domain*: (...can be simply calculated from LPC-cepstral coefficients – no integration, just a sum)

$$d_2 = \sqrt{\int_{-1/2}^{+1/2} \left|V(f)\right|^2 df}, \quad \text{where} \quad V(f) = 10\log G(f) - 10\log \hat{G}(f), \tag{2}$$

where $G(f)$ and $\log \hat{G}(f)$ are power spectral densities of the original and the coded frame.

Subjective Quality Evaluations

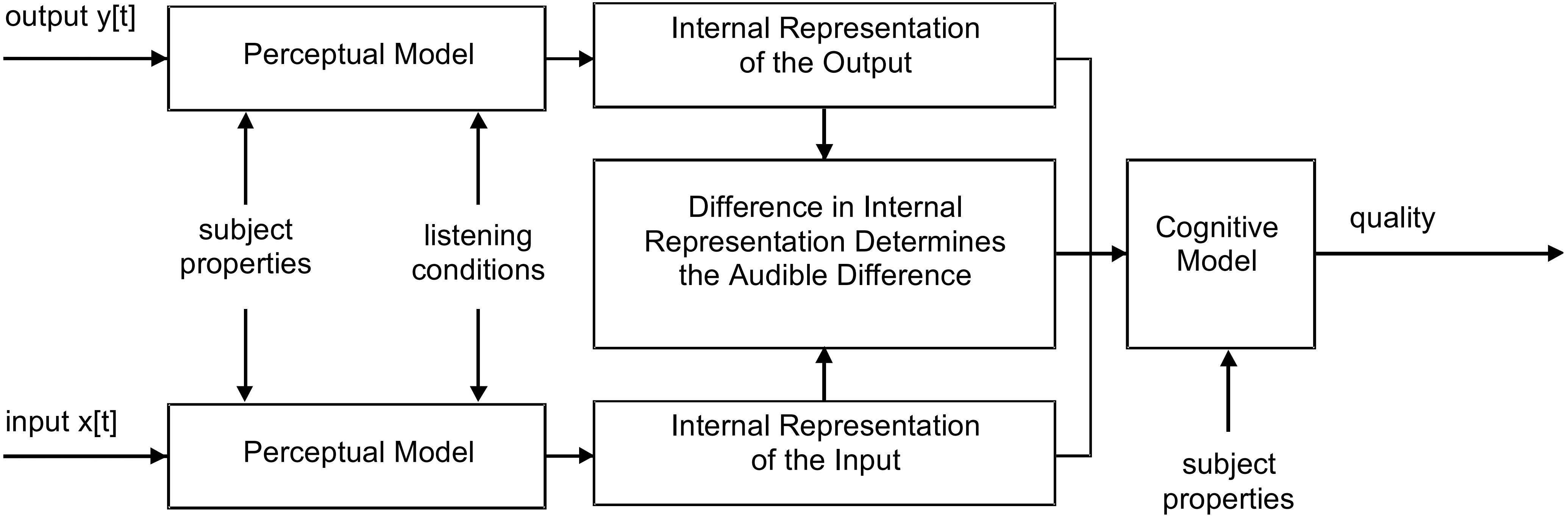
requires a group of trained listeners.

- DRT (Diagnostic Rhyme Test) – intelligibility measurement using pairs of similarly sounding words (e.g. meat×heat).
- DAM (Diagnostic Acceptability Measure) – a set of methods judging the overall quality of a communication system.
- MOS (Mean Opinion Score) – a group of 12-64 listeners evaluating the quality on a 5-point axis. The listeners are “calibrated” by signals of known MOS:

MOS	quality	remark
1	bad (unacceptable)	very annoying noise and artifacts in the signal
2	poor	...
3	fair	in between
4	good	...
5	excellent	not distinguishable from the original, w/o any audible nois

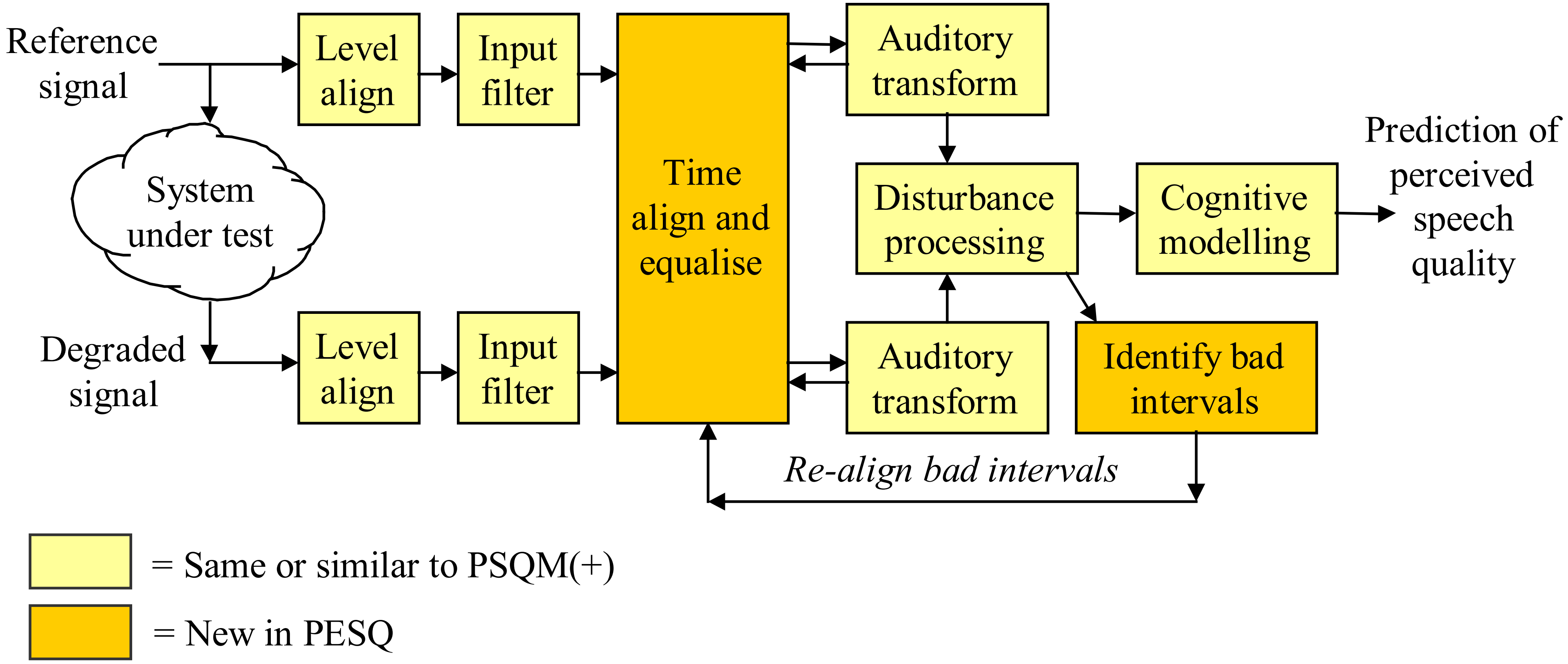
Quality Measure Inspired by Human Perception

Perceptual Speech Quality Measure – PSQM


output y[t]
Perceptual Model
Internal Representation of the Output
subject properties
listening conditions
Difference in Internal Representation Determines the Audible Difference
Cognitive Model
quality
input x[t]
Perceptual Model
Internal Representation of the Input
subject properties

source: OPTICOM Whitepaper on "State-of-the-Art Voice Quality Testing", 2000
ITU standard 1996, P.861.

Perceptual Evaluation of Speech Quality – PESQ



Reference signal

Degraded signal

System under test

Level align

Input filter

Level align

Input filter

Time align and equalise

Auditory transform

Auditory transform

Disturbance processing

Cognitive modelling

Identify bad intervals

Prediction of perceived speech quality

Re-align bad intervals

🟨 = Same or similar to PSQM(+)

🟧 = New in PESQ

source: OPTICOM Whitepaper on "State-of-the-Art Voice Quality Testing", 2000

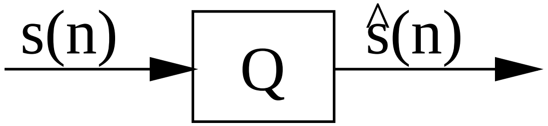
ITU standard P.862

A.W.Rix et al.: Perceptual evaluation of speech quality (PESQ) a new method for speech quality assessment of telephone networks and coders, Proc. ICASSP 2001.

WAVEFORM CODING

Pulse code modulation (PCM)

historical title, independent quantization of samples using fixed number of bits.



Linear quantization (constant quantization step):

$$SNR = 6B + K \tag{3}$$

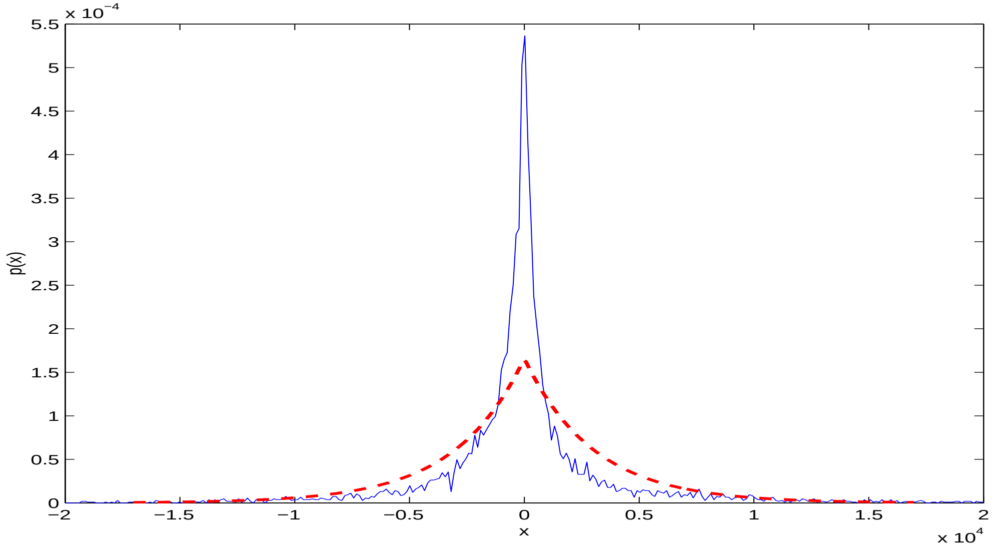
in [dB], where B is the number of bits and K is a constant depending on the character of the signal. For CD theoretically: 16×6=96 dB. Interpretation of the equation: adding 1 bit, SNR will improve by 6 dB.

For speech signals, linear quantization is not optimal:

1. speech contains a lot of "small samples", the probability density function (PDF) can be approximated by the Laplace distribution:

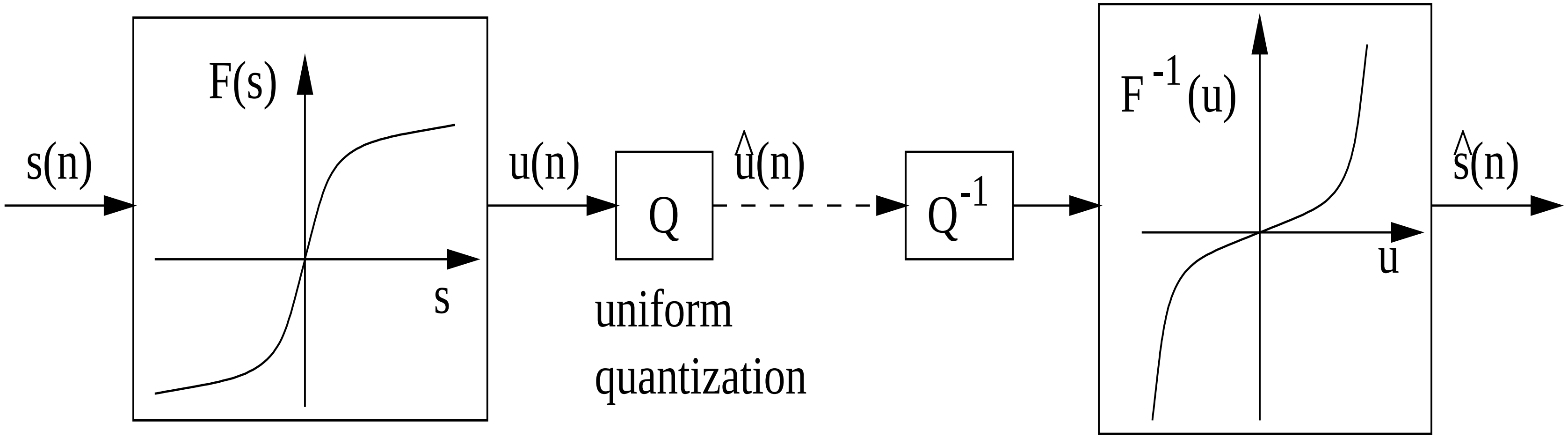
$$p(x) = \frac{1}{\sqrt{2}\sigma_x} e^{-\frac{\sqrt{2}|x|}{\sigma_x}}, \tag{4}$$

   where σ_x is standard deviation. Example: Czech sentence with discarded silence segments:



2. Perceptual studies show that human ear has a logarithmic sensitivity to the amplitude of acoustic pressure

logarithmic PCM, compression in the encoder and decompression in the decoder:



Nonlinearity cannot be defined as log: ($\log(0) = -\infty$), thus approximations:

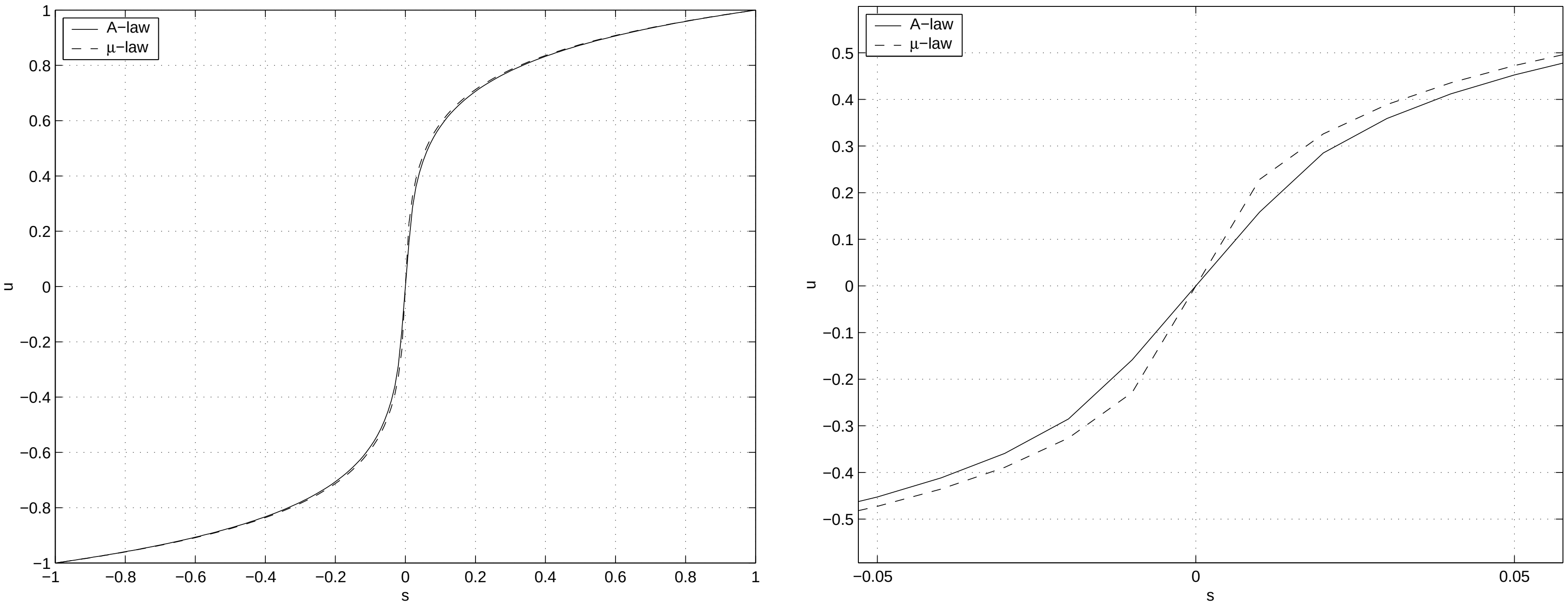
- Europe: **A-law**:

$$u(n) = S_{max} \frac{1 + \ln A \dfrac{|s(n)|}{S_{max}}}{1 + lnA} \text{ sign } [s(n)], \quad \text{where} \quad A = 87.56. \tag{5}$$

- USA: μ-law:

$$u(n) = S_{max} \frac{\ln\left(1 + \mu \dfrac{|s(n)|}{S_{max}}\right)}{\ln(1+\mu)} \text{ sign } [s(n)], \quad \text{where} \quad \mu = 255. \tag{6}$$

Comparison of A-law and μ-law:



$\Rightarrow$ both are practically identical and both improving SNR for small signals by 12 dB. For telephony applications, 8 bit log-PCM has similar quality as 13 bit linear. CCITT G.711

Adaptive Pulse-Code Modulation (APCM)

The quantization levels (uniform or non-uniform) are adapted for each block of signal samples. Information on quantization levels can be:

- sent to the decoder as additional information – *feed-forward*.
- computed from several past samples (as well available for the decoder) – *feed-back*.

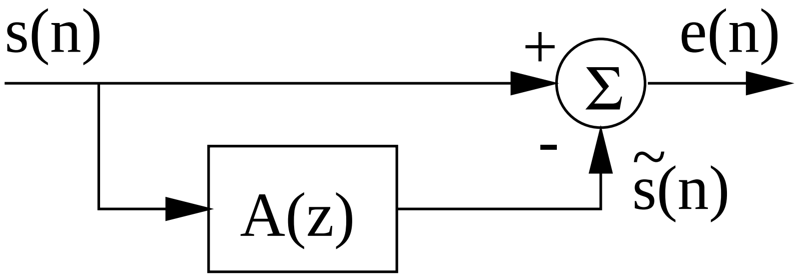
APCM is not used independently, but as a part of other coders (for example full-rate GSM: RPE-LTP).

Differential Pulse-Code Modulation (DPCM)

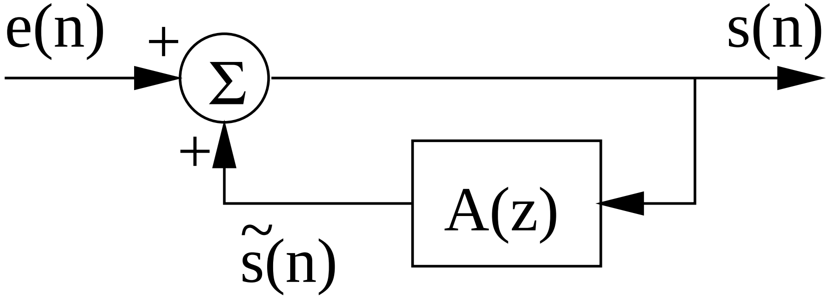
- Exploits statistical dependencies between samples (the only signal where samples are independent is white noise).
- The current sample is estimated from several previous samples. Encoding of the error signal.
- In case the estimate is good, the error signal has low energy and low amplitude, that means less bits are required for the encoding than for the original signal.

$$A(z) = \sum_{i=1}^{P} a_i z^{-i} \tag{7}$$

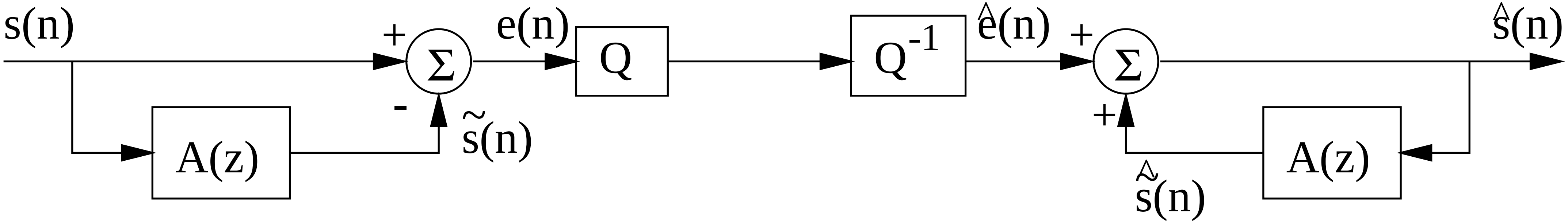
Error signal:



The decoder is simple: current sample is predicted from previous ones and the decoded error is added:

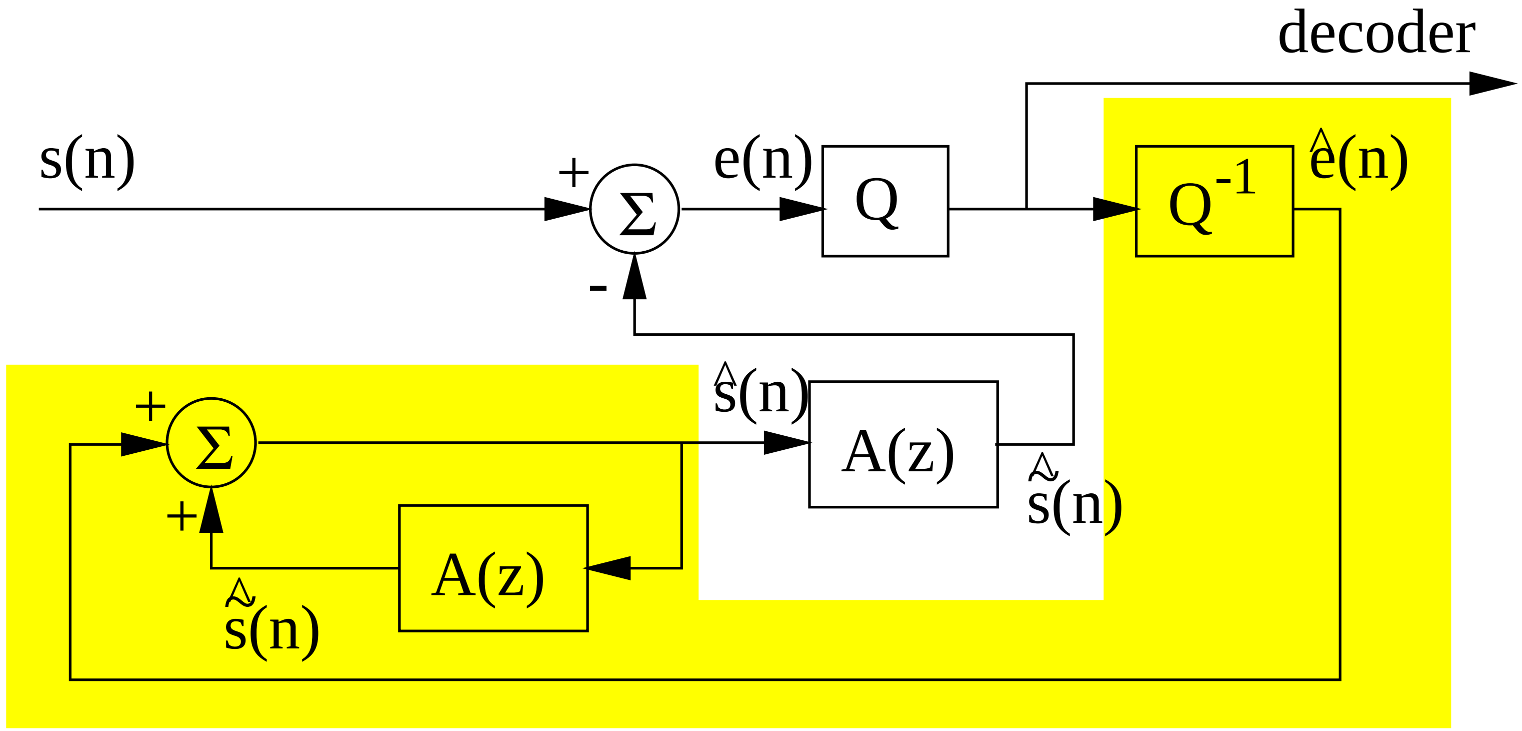


. . . the error signal is **quantized**



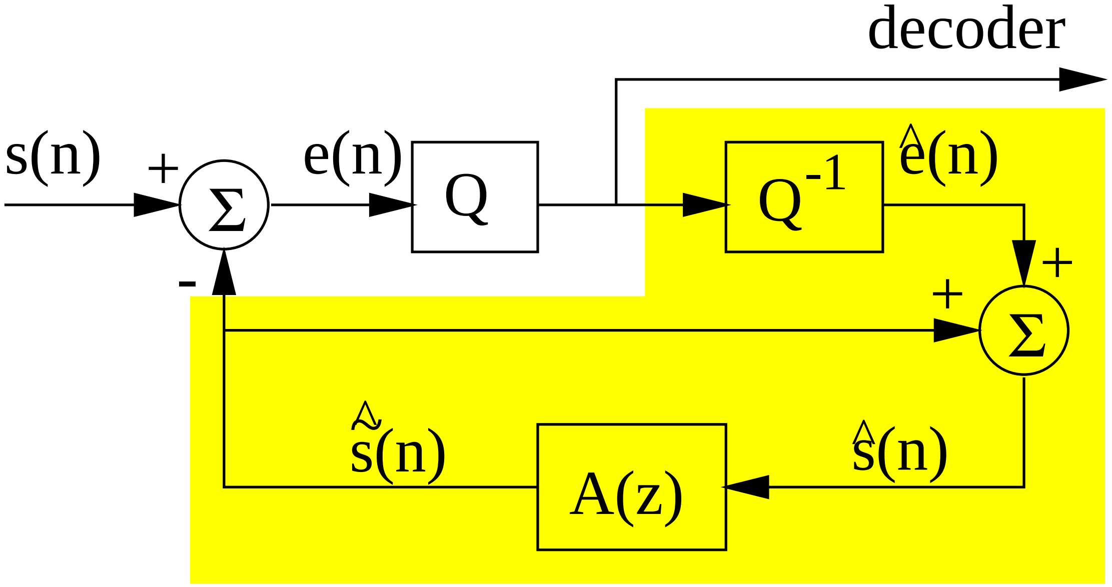
!!! In such case, the decoder would estimate the signal from **different** samples than the encoder!!!

- Encoder disposes of: $s(n)$.
- Decoder disposes of $\hat{s}(n)$ due to quantization of $e(n)$, not equal to $s(n)$!
- Decoder thus has to be **embedded** in the encoder. The output is used for the prediction.



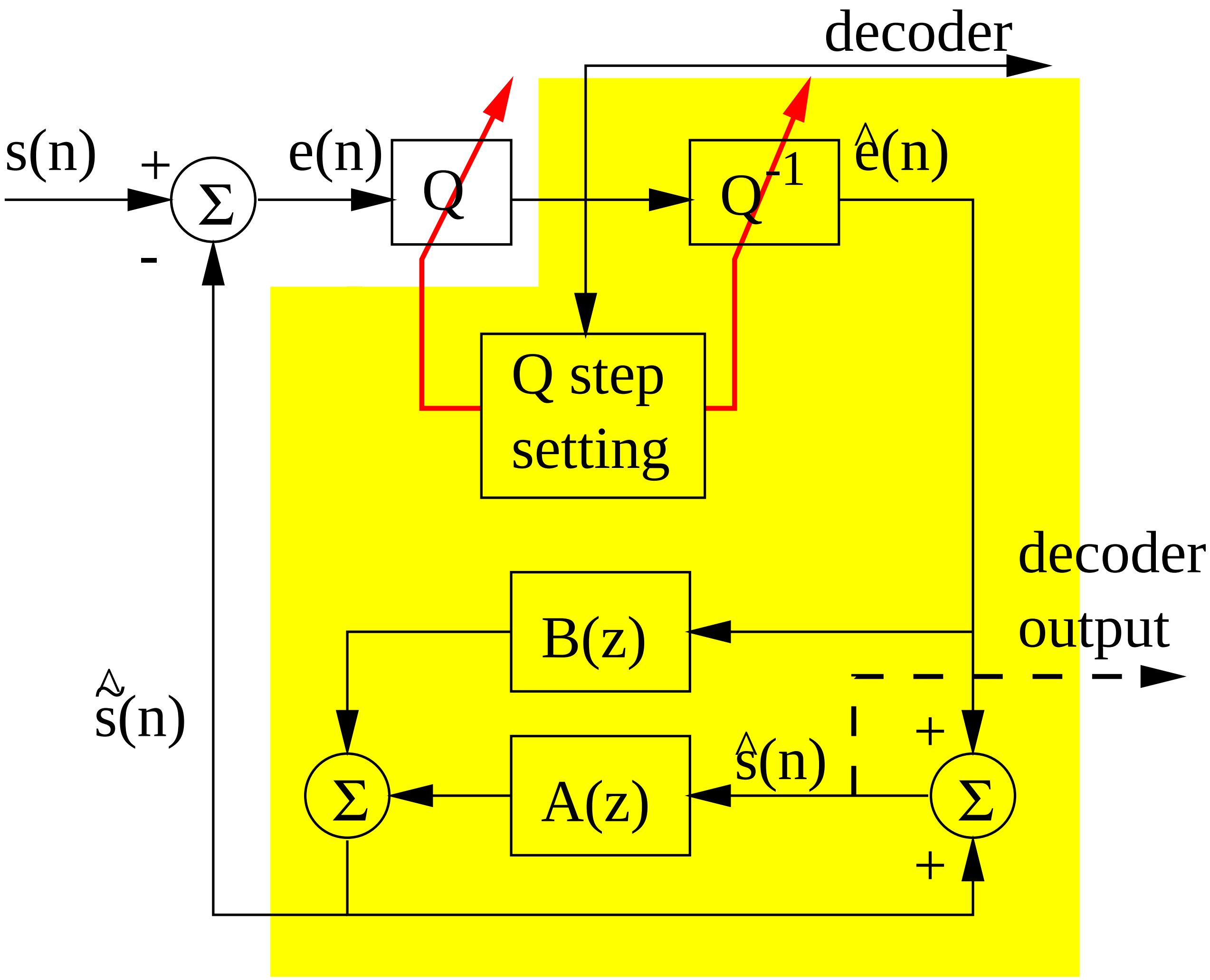
This scheme is not optimal as the filter $A(z)$ appears twice and it filters in both cases the same signal!

Optimization:



The coder again contains a complete decoder (yellow color) in itself.

Adaptive Differential Pulse-Code Modulation (ADPCM) G. 721



- difference to DPCM - adaptation of the quantization step.
- ADCPM is used in G.721: log-PCM (64 kbit/s) $\Rightarrow$ 32 kbit/s.
- G.721 is composed of 2 filters used for the error signal computation $e(n)$: $A(z)$ (as in the previous example), $B(z)$ estimates a sample of the error signal from several previous values of the error sample (IIR).
- Block "Q-step setting" regulates the quantization step.
- Decoder (yellow) is "hidden" in the encoder.
- Information of the Q step and the filter parameter identification is based on the encoder output (back-ward).

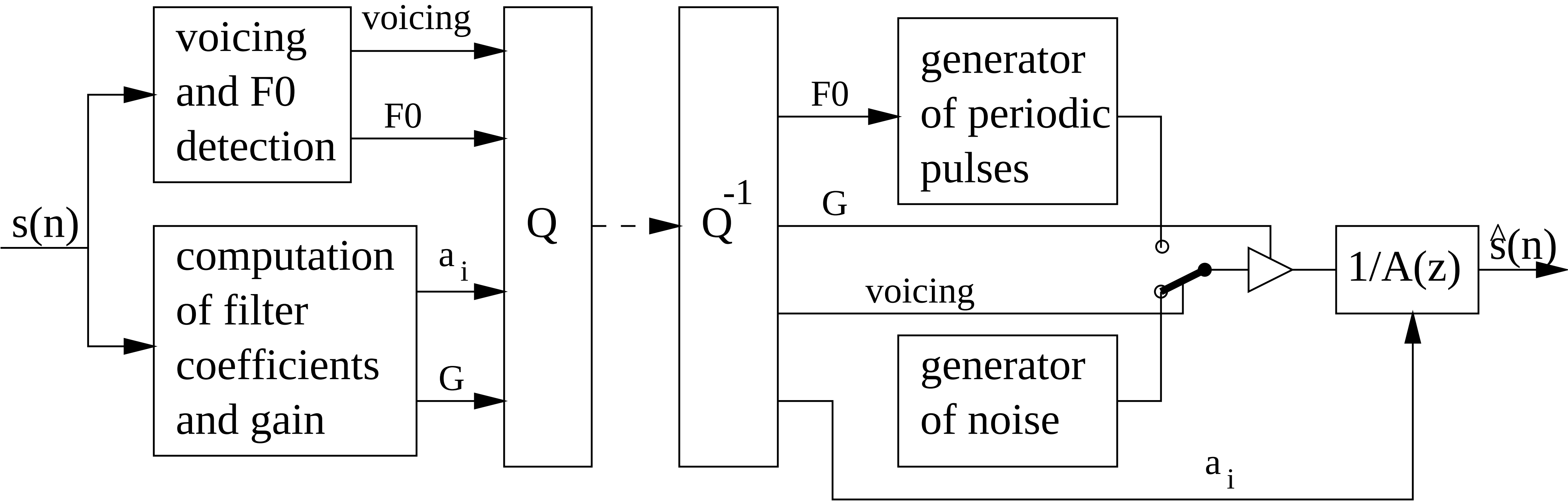
VOCODERS

Voice coders

- exploit findings of human speech production to reduce bit-rate.
- work well only for speech.
- make use of the excitation—filter model.

Encoder Based on the Linear Predictive Model - LPC

The input signal is segmented into frames, and for each frame, a set of predictor coefficients is estimated: $A(z) = 1 + \sum_{i=1}^{P} a_i z^{-i}$. Then, gain G of the filter and voiced/unvoiced flag is calculated. In case of a voiced frame, pitch period (lag) is estimated.



Excitation !!!

Example: US-DoD FS1015 standard: filter 1800 bps, excitation 600 bps, 2.4 kbps. The main drawback is poor modeling of excitation. Unnatural sound of speech. Big improvement in the CELP encoders.

Residual Excited Linear Prediction – RELP

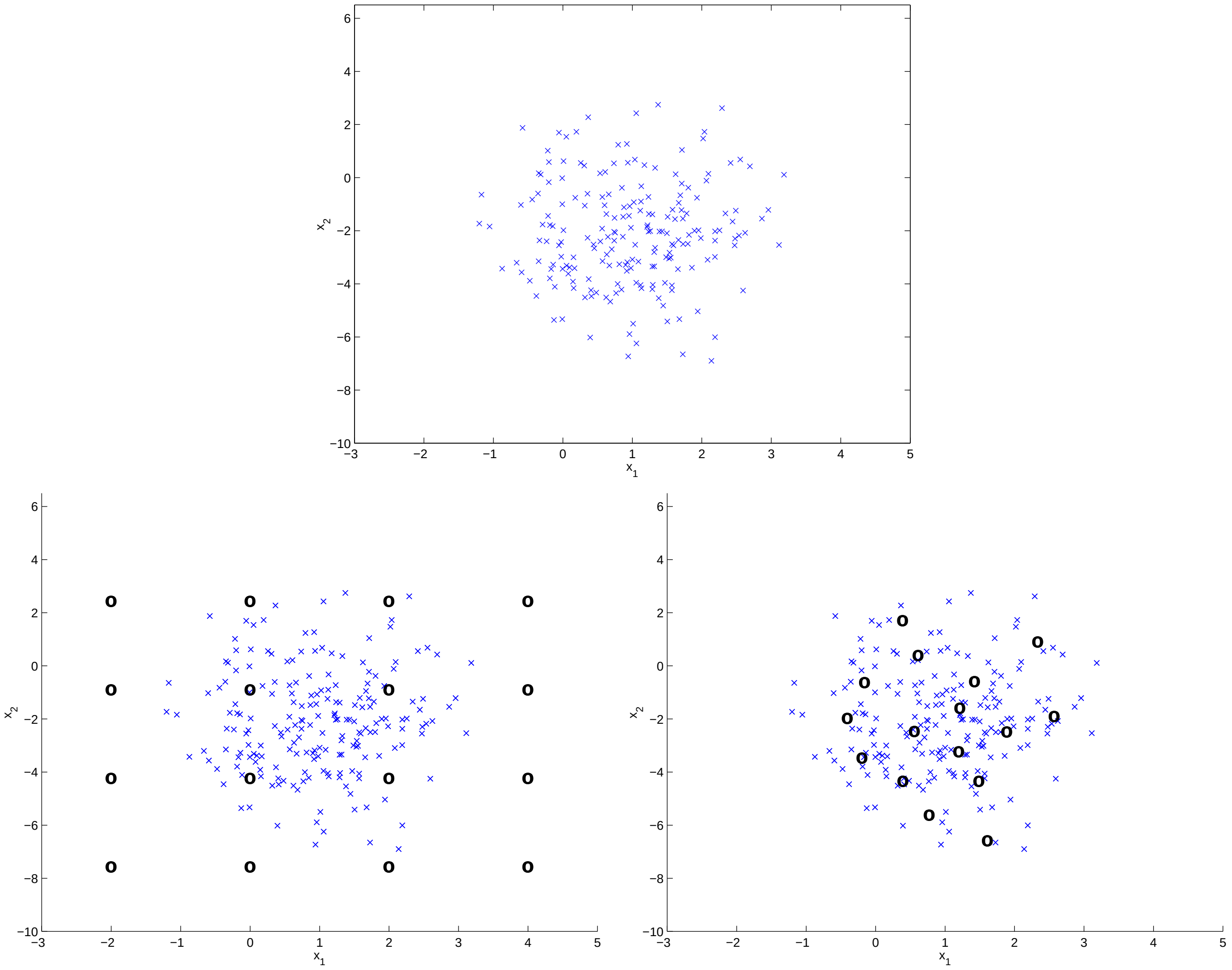
For each frame, parameters of the filter $A(z)$ are estimated and the error signal is calculated $e(n)$: $E(z) = A(z)S(z)$. This signal is transferred to the decoder, where then filtered by the filter $H(z) = \frac{1}{A(z)}$. If the filter coefficients a_i nor the error signal $e(n)$ are not quantized, the resulting signal is equivalent to the input signal $s(n)$.

Increasing of the bit-rate — Bad. . .

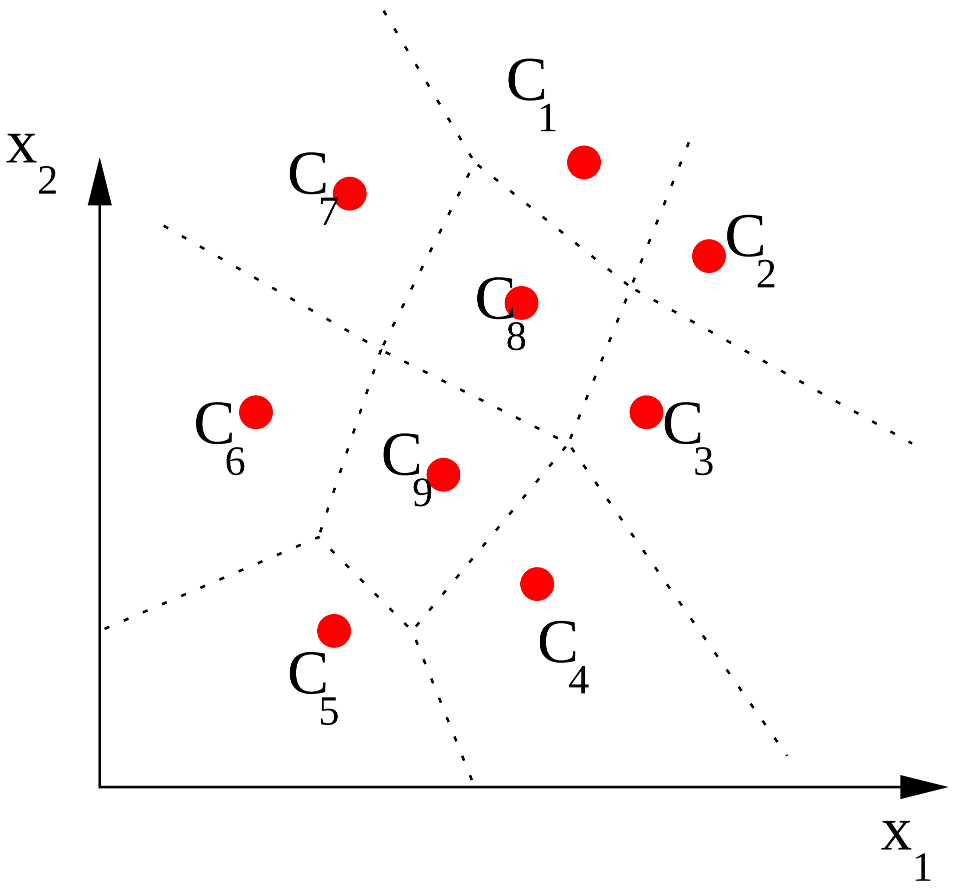
VECTOR QUANTIZATION

Why ?

- Encoding of P-dimensional vectors is very costly (P float numbers results in $P \times 4$ byte. . .)
- Scalar quantization of the single components results in wasting bits.
- In the P-dimensional space, it is more efficient to use code-vectors ("typical") and map the input vectors to these code-vectors.


6
4
2
0
−2
−4
−6
−8
−10
−3
−2
−1
0
1
2
3
4
5
x_1
x_2
o

Code vectors, centroids, Voronoi cells...



Code-Vectors Training — K-means

The code-vectors are trained on the training data.

- Given N training vectors,
- train a codebook $\mathbf{Y}$ of the size K.

- **Initialization:** $k = 0$, define $\mathbf{Y}(0)$.
- **Step 1:** Alignment of the vectors to the cells — "encoding":

$$Q[\mathbf{x}] = \mathbf{y}_i(k) \;\text{ if }\; d(\mathbf{x}, \mathbf{y}_i(k)) \leq d(\mathbf{x}, \mathbf{y}_j(k)) \;\text{ pro }\; j \neq i,\; j \in 1 \ldots K \tag{8}$$

To measure the distance $d(\mathbf{x}, \mathbf{y}_j)$, Euclidean distance can be used:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})} = \sqrt{\sum_{k=1}^{P} |x_k - y_k|^2}. \tag{9}$$

Note, that code-vectors and cells carry the index of the current generation of the codebook.

- **Step 2:** Calculate relative improvement of distortion – total distance:

$$D_{VQ} = \frac{1}{N} \sum_{n=1}^{N} d\left(\mathbf{x}(n), Q[\mathbf{x}(n)]\right). \tag{10}$$

- **Step 3:** If the calculated relative improvement is smaller than the threshold calculated as:

$$\frac{D_{VQ}(k-1) - D_{VQ}(k)}{D_{VQ}(k)} \leq \varepsilon, \tag{11}$$

  STOP and denote the k-th generation to be the resulting codebook. $\mathbf{Y} = \mathbf{Y}(k)$. If the improvement is still significant, continue:

- **Step 4:** Calculate new centroids of the cells and use them as new code-vectors:

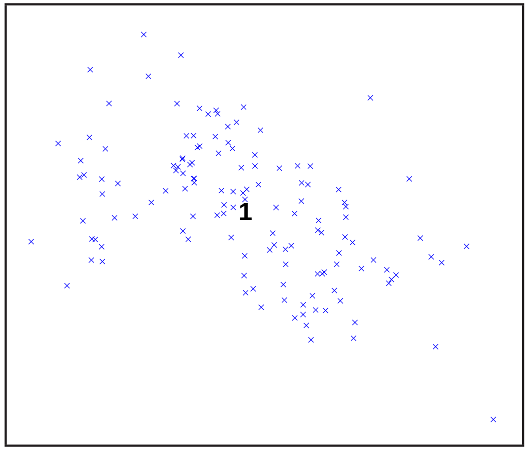
$$\mathbf{y}_i(k+1) = Cent(C_i(k)) = \frac{1}{M_i(k)} \sum_{\mathbf{x} \in C_i(k)} \mathbf{x}, \tag{12}$$

  where $M_i(k)$ is the number of the training vectors, that for the k-th generation of the codebook are aligned to the cell i. Simple arithmetic average. Increment: $k = k + 1$. GOTO Step 1.

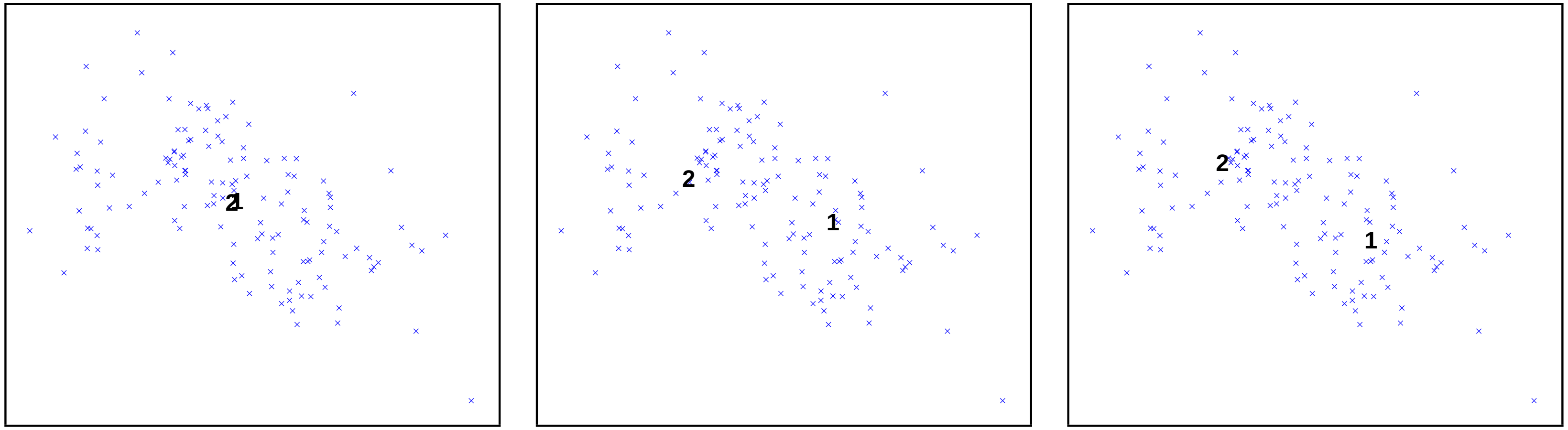
Linde-Buzo-Gray — LBG

- The K-means algorithms can be suboptimal due to incorrect initialization (how to initialize $\mathbf{Y}(0)$ – random initialization ? random selection from the input vectors?)
- Can happen, that during training a code-vector will end up in having no training vectors $\Rightarrow$ division by zero $\Rightarrow$ crash :-(
- LBG solves the problem by gradual splitting of the vectors in the codebook:

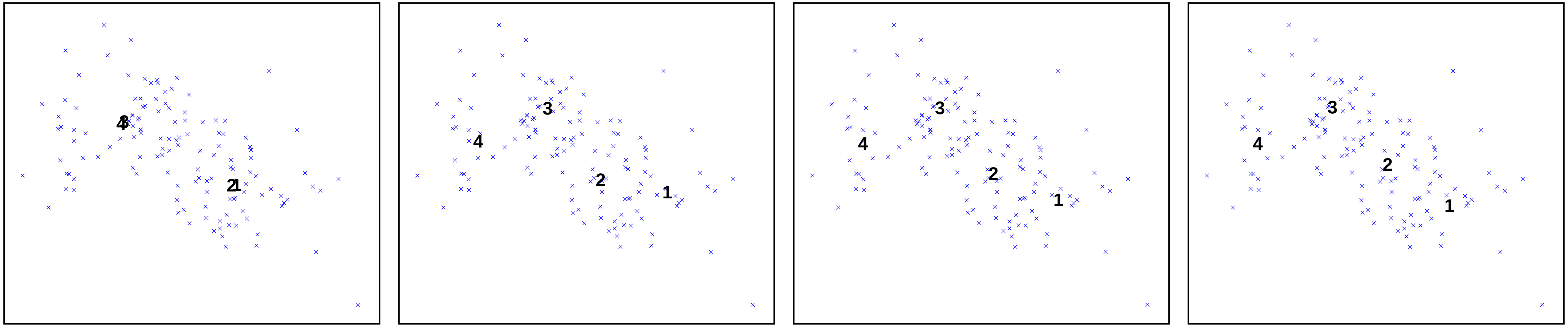
$r = 0, L = 1$



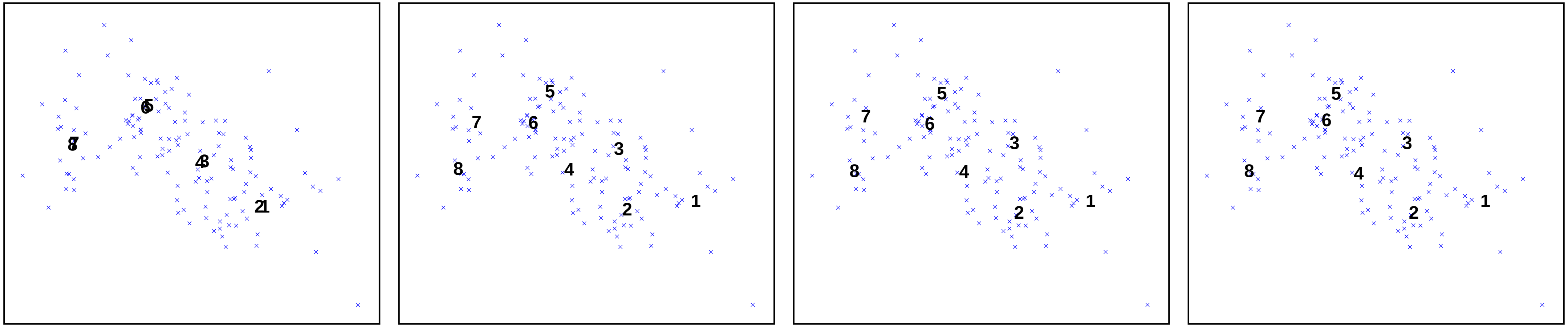
$r = 1, L = 2$



$r = 2, L = 4$



$r = 3, L = 8$



Utilization of Vector Quantization

1. quantization of the filter coefficients a_i (most often PARCOR, LAR or LSF are derived).
2. excitation encoding in CELP-like algorithms, blocks of samples passed through a short-term $A(z)$ and a long-term $B(z)$.

Flavors of Vector Quantization

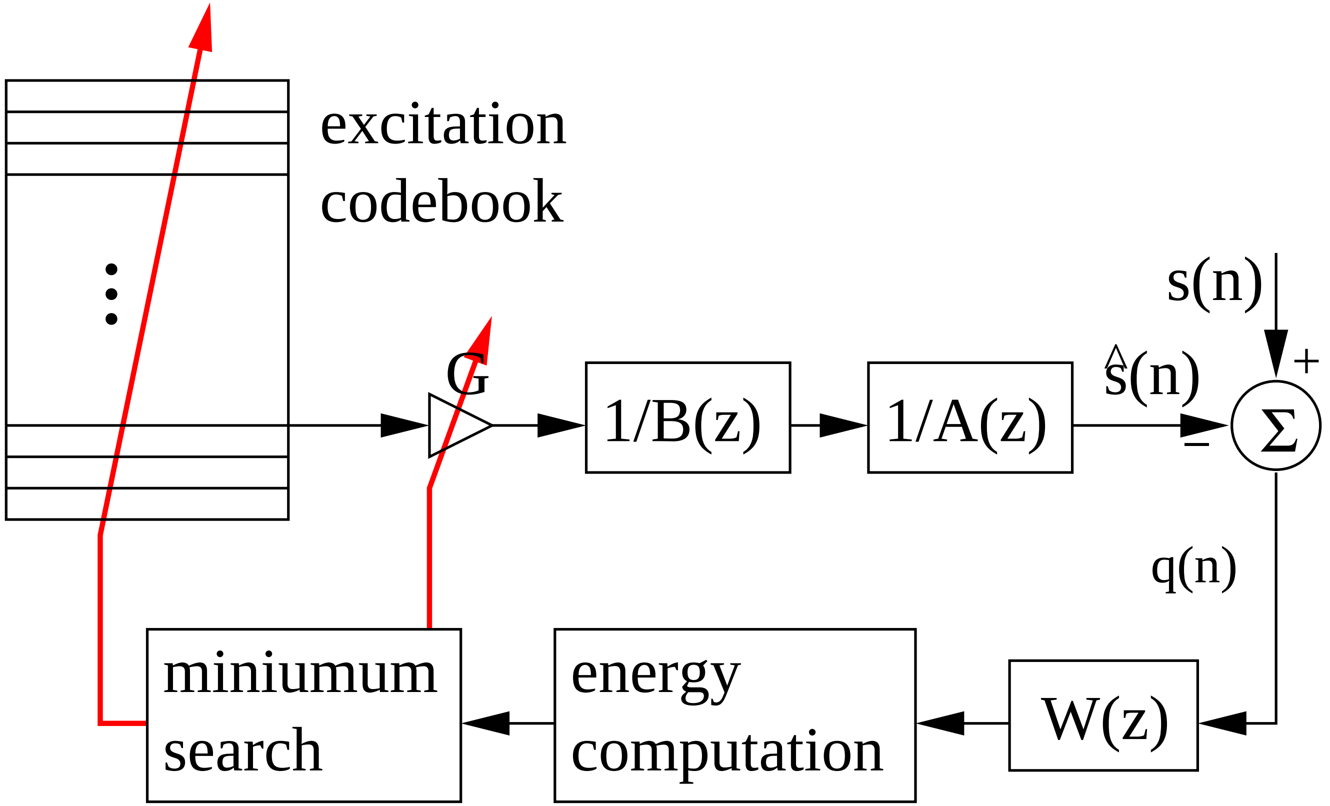
The algorithms of both VQ training and coding are computationally quite expensive, the following tricks make VQ more efficient:

- split-VQ: split of a vector into several sub-vectors with less coefficients. Thus it results in having several smaller codebooks. (typically 3-3-4 for P=10).
- algebraic VQ: The code vectors are not distributed arbitrarily in the space, but their positions are deterministic (e.g. uniform on a hyper-plane). Thus the input vector does not have to be compared to all the code vectors.
- random codebook (for training): for large K quality of the trained codebook approaches a randomly defined codebook $\Rightarrow$ no need in training!
- tree-structured VQ: remember all generations of LBG and assume, that if the vector $\mathbf{y}^k$ was in the generation k of the codebook aligned to some code vector, then in the generation $k+1$ the vector $\mathbf{y}^k$ can be assigned only to the child code vectors $\Rightarrow$ suboptimal, but requires only $2\log_2 K$ comparisons instead of K.
- multi-stage VQ: 2 codebooks are used, the second codebook is used for quantization of the first codebook error. While decoding, vectors from both codebooks are summed.

CELP – Codebook-Excited Linear Prediction

Excitation is coded using **code-books**

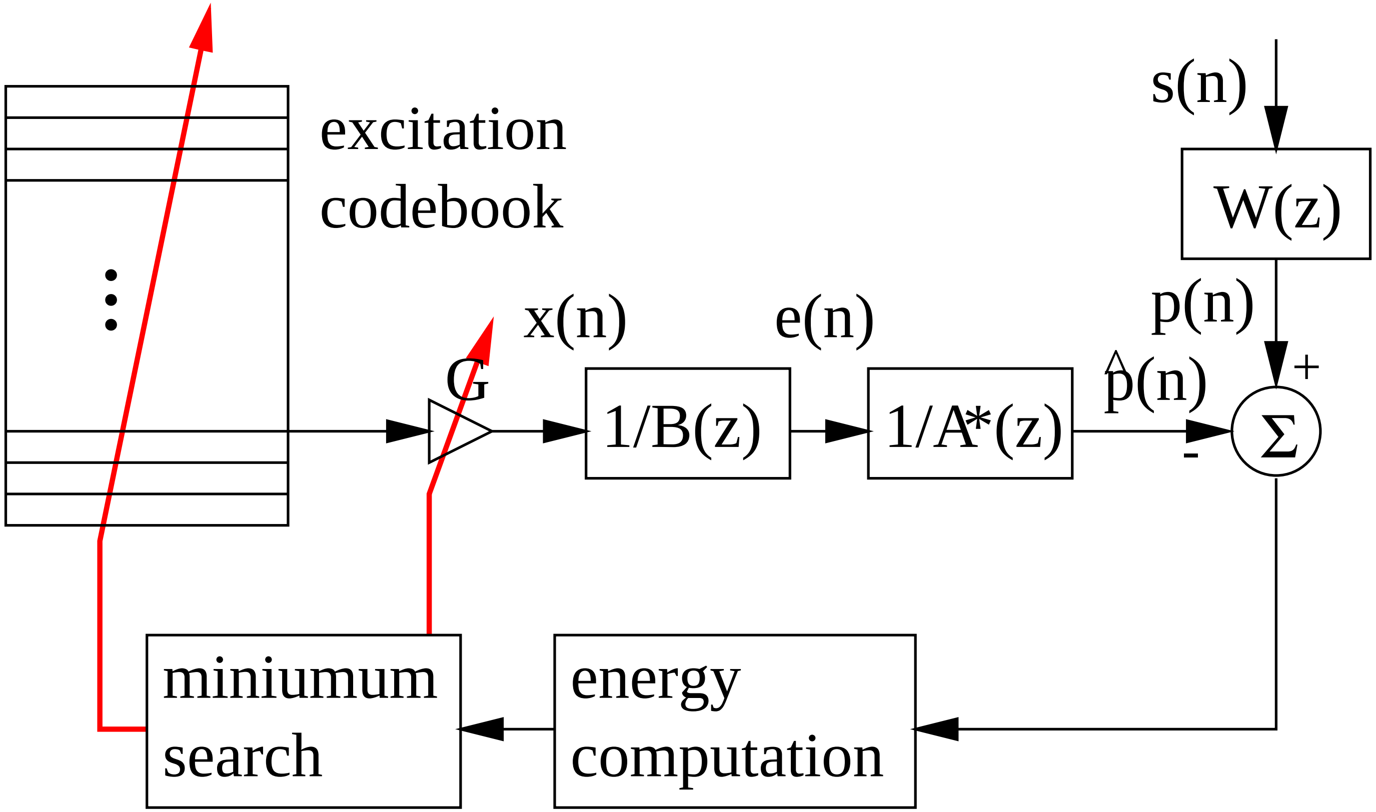
Basic structure with a perceptual filter is:



. . . Each tested signal has to be filtered by $W(z) = \frac{A(z)}{A^\star(z)}$ — too much work!

Perceptual filter on the Input

Filtering is a linear operation, thus $W(z)$ can be moved to both sides: to the input and to the signal after the sequence $\frac{1}{B(z)} - \frac{1}{A(z)}$. Can be simplified to: $\frac{1}{A(z)}\frac{A(z)}{A^\star(z)} = \frac{1}{A^\star(z)}$ — this is a new filter that has to be used after $\frac{1}{B(z)}$.


excitation
codebook
s(n)
W(z)
x(n)
e(n)
p(n)
G
1/B(z)
1/A*(z)
$\hat{p}(n)$
+
-
Σ
miniumum
search
energy
computation