Speech Coding II.

Jan Černocký, Valentina Hubeika

{cernocky|ihubeika}@fit.vutbr.cz

DCGM FIT BUT Brno

Agenda

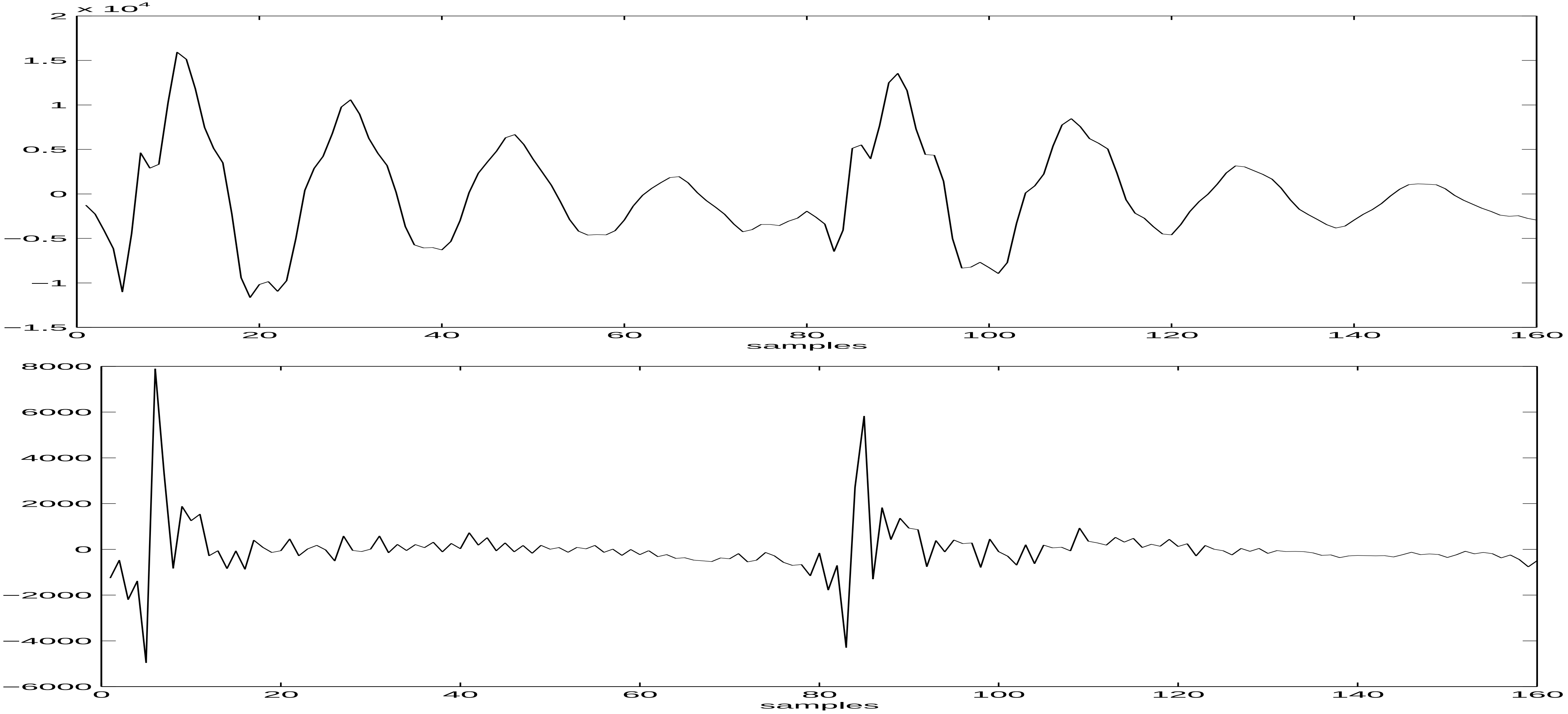
- Tricks used in coding – long-term predictor, analysis by synthesis, perceptual filter.
- GSM full rate – RPE-LTP
- CELP
- GSM enhanced full rate – ACELP

EXCITATION CODING

- LPC reaches a low bit-rate but the quality is very poor ("robot" sound).
- caused by primitive coding of excitation (only voiced/unvoiced), whilst people have both components (excitation and modification) varying over time.
- in modern (hybrid) coders a lot of attention is paid to coding of excitation.
- following slides: tips in excitation coding.

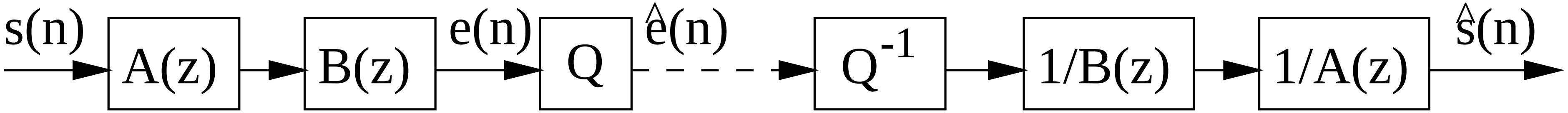
Long-Term Predictor (LTP)

The LPC error signal attributes behaviour of noise, however only within short time intervals. For voiced sounds in longer time span, the signal is correlated in periods of the fundamental frequency:



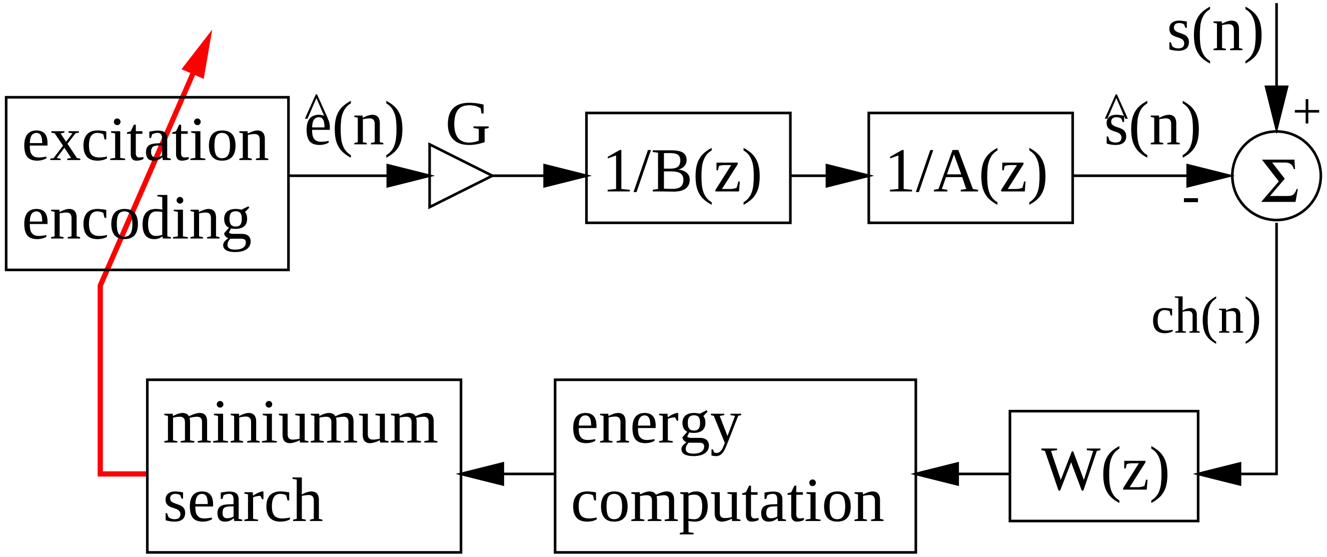
$\Rightarrow$ long-term predictor (LTP) with the transfer function $-bz^{-L}$. Predicts the sample $s(n)$ from the preceding sample $s(n-L)$ (L is the pitch period in samples – lag).

Error signal: $e(n) = s(n) - \hat{s}(n) = s(n) - [-bs(n-L)] = s(n) + bs(n-L)$, thus $B(z) = 1 + bz^{-L}$.



Analysis by Synthesis

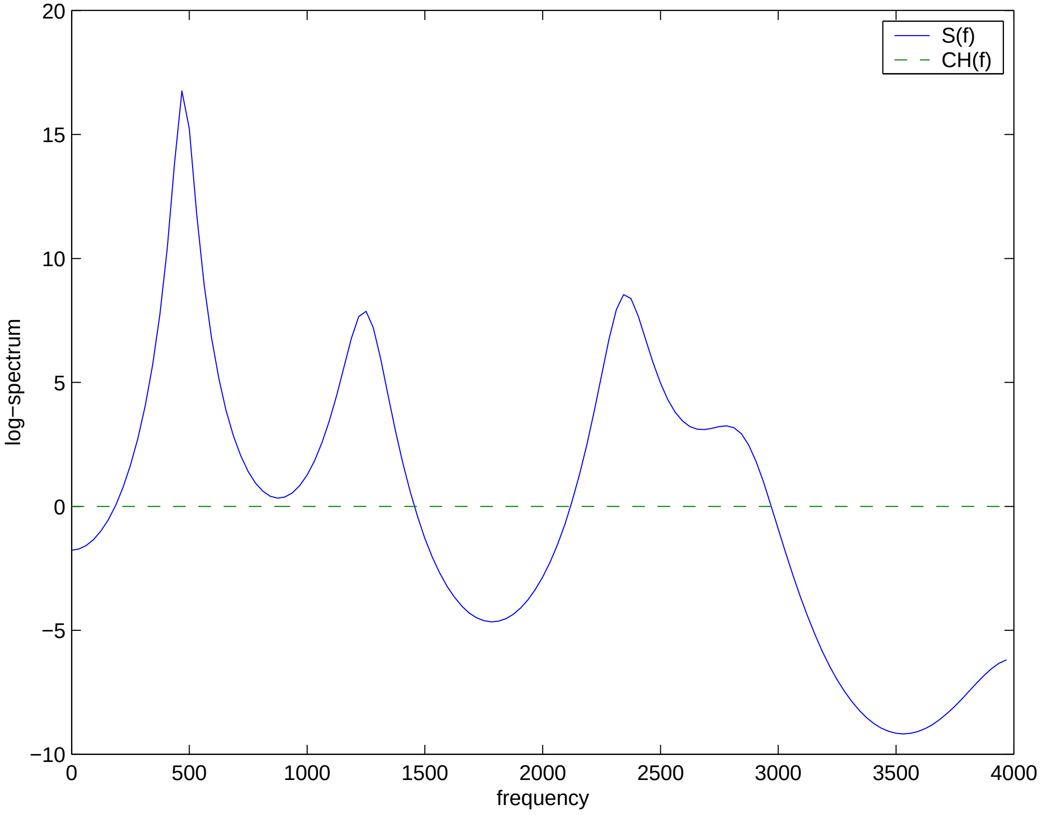
In most coders, the optimal excitation (after removing short-term and long-term correlations) cannot be found analytically. Thus all possible excitations have to be generated, the output (decoded) speech signal produced and compared to the original input signal. The excitation producing minimal error wins. The method is called closed-loop. To distinguish between the excitation $e(n)$ and the error signal between the generated output and the input signals, the later (error signal) is designated as $ch(n)$.

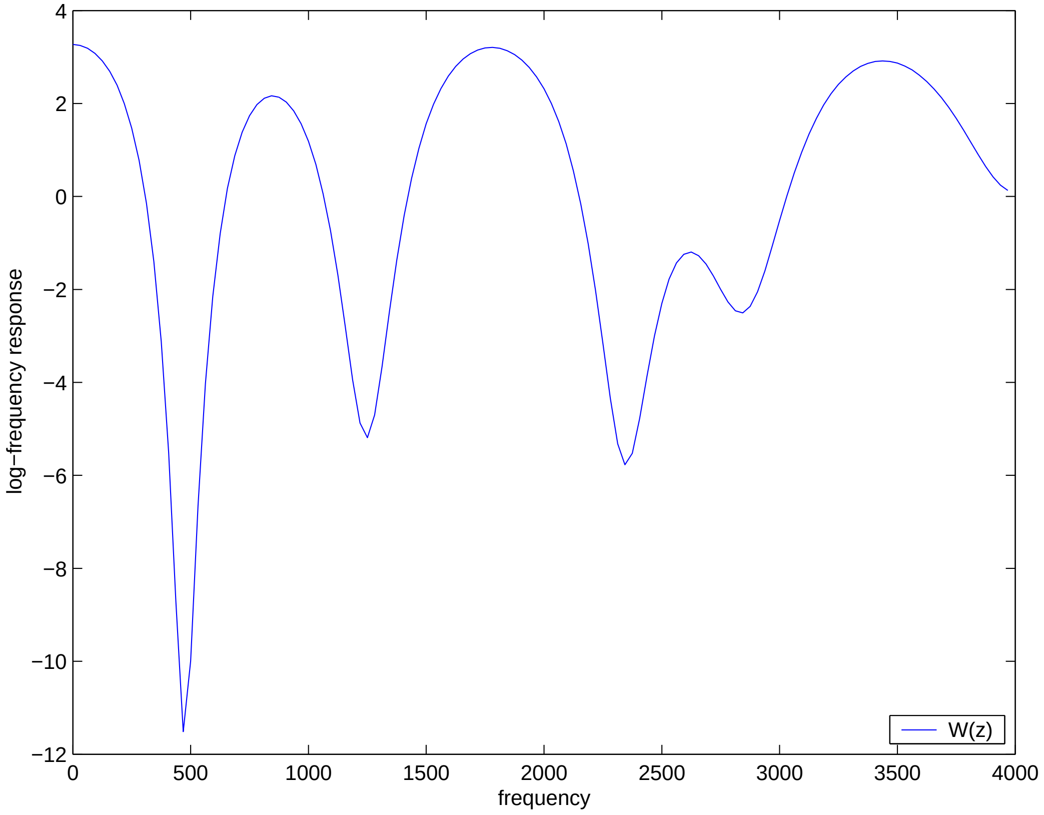


Perceptual filter $W(z)$

In the closed-loop method, the synthesized signals is compared to the input signal. PF – modifies the error signal according to human hearing.

Comparison of the error signal spectrum to the smoothed speech spectrum:





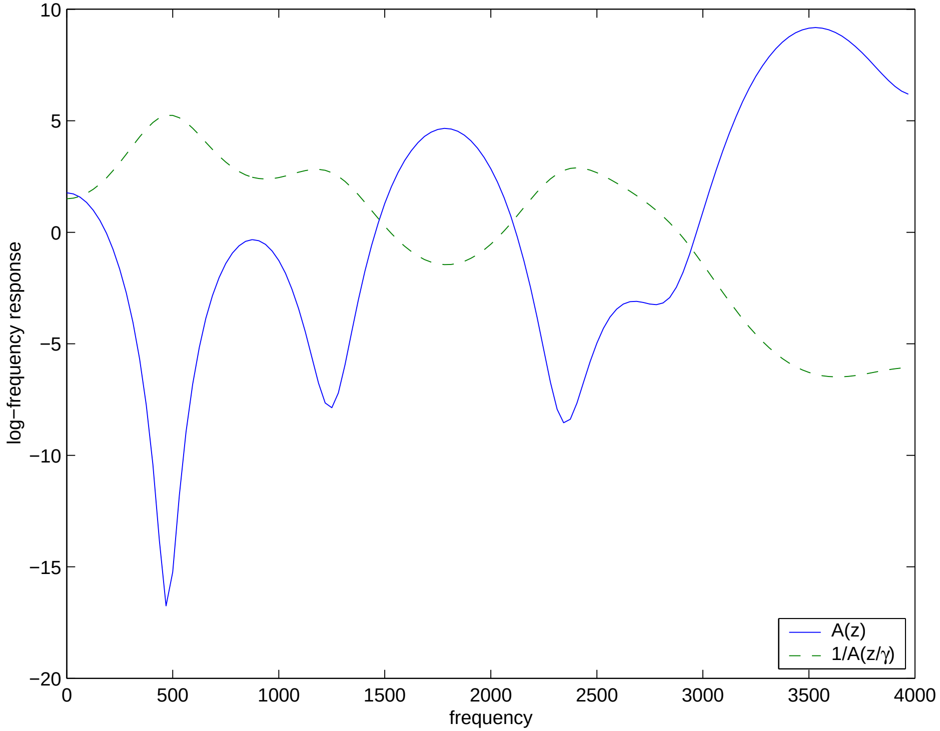
in formant regions, speech is much "stronger" than the error ⇒ the error is *masked*. In "valleys" between formants, there is an opposite situation: for example in the 1500-2000 H band, the error is *stronger* than speech - leads to an audible artifacts.

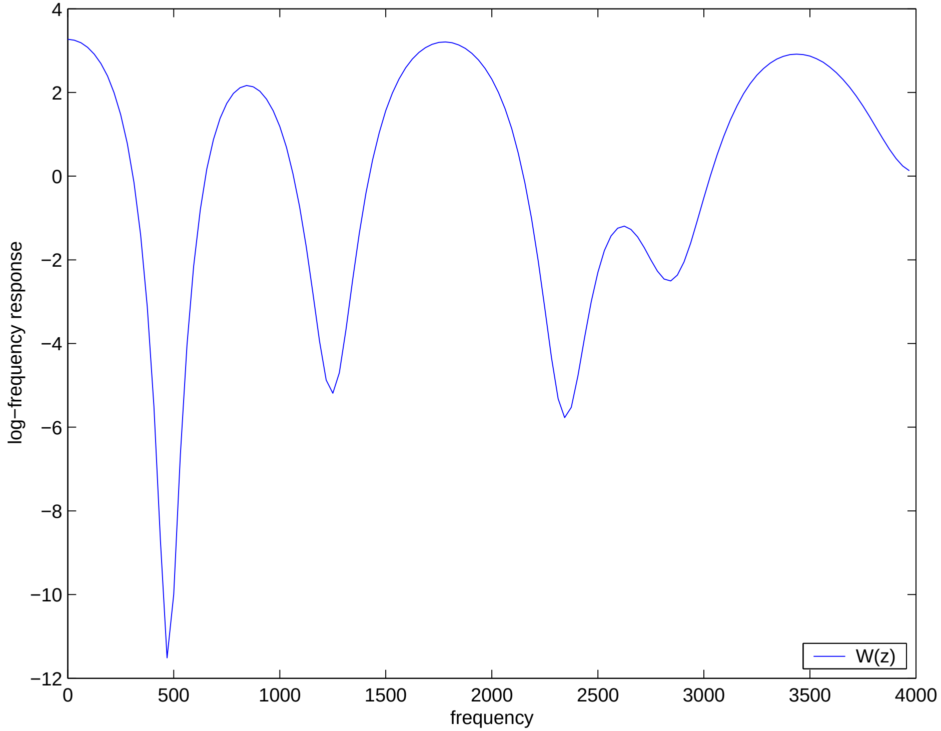
⇒ a filter that attenuates error signal in formant (not important) regions and amplify it in "sensitive" regions.

$$W(z) = \frac{A(z)}{A(z/\gamma)}, \quad \text{where} \quad \gamma \in [0.8, 0.9] \tag{1}$$

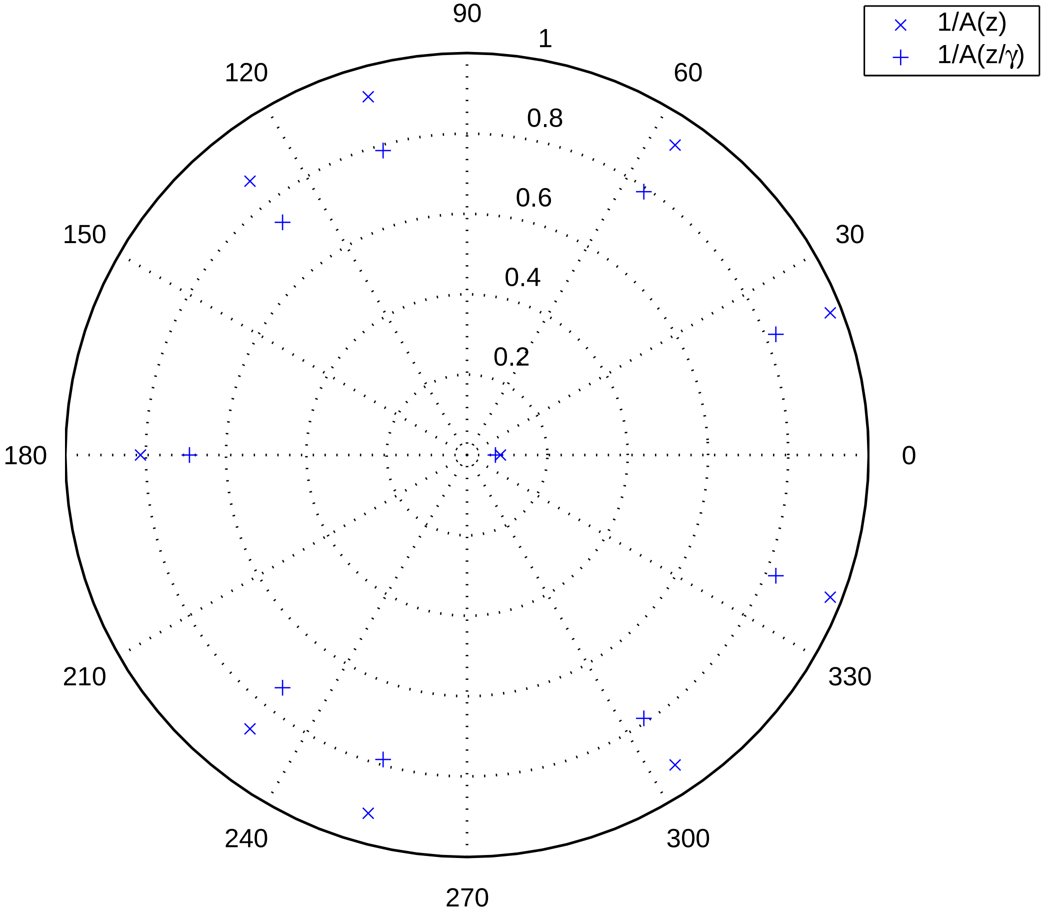
How does it work ?

- the nominator $A(z)$ (which is actually an "inverse LPC filter") has the characteristic inverse to the smoothed speech spectrum.
- filter $\frac{1}{A(z/\gamma)}$ has a similar characteristic as $\frac{1}{A(z)}$ (smoothed speech spectrum), but as the poles are placed close to the origin of the unit circle, the picks of the filter are not sharp.
- multiplication of these components results in TF $W(f)$ with the required characteristic.





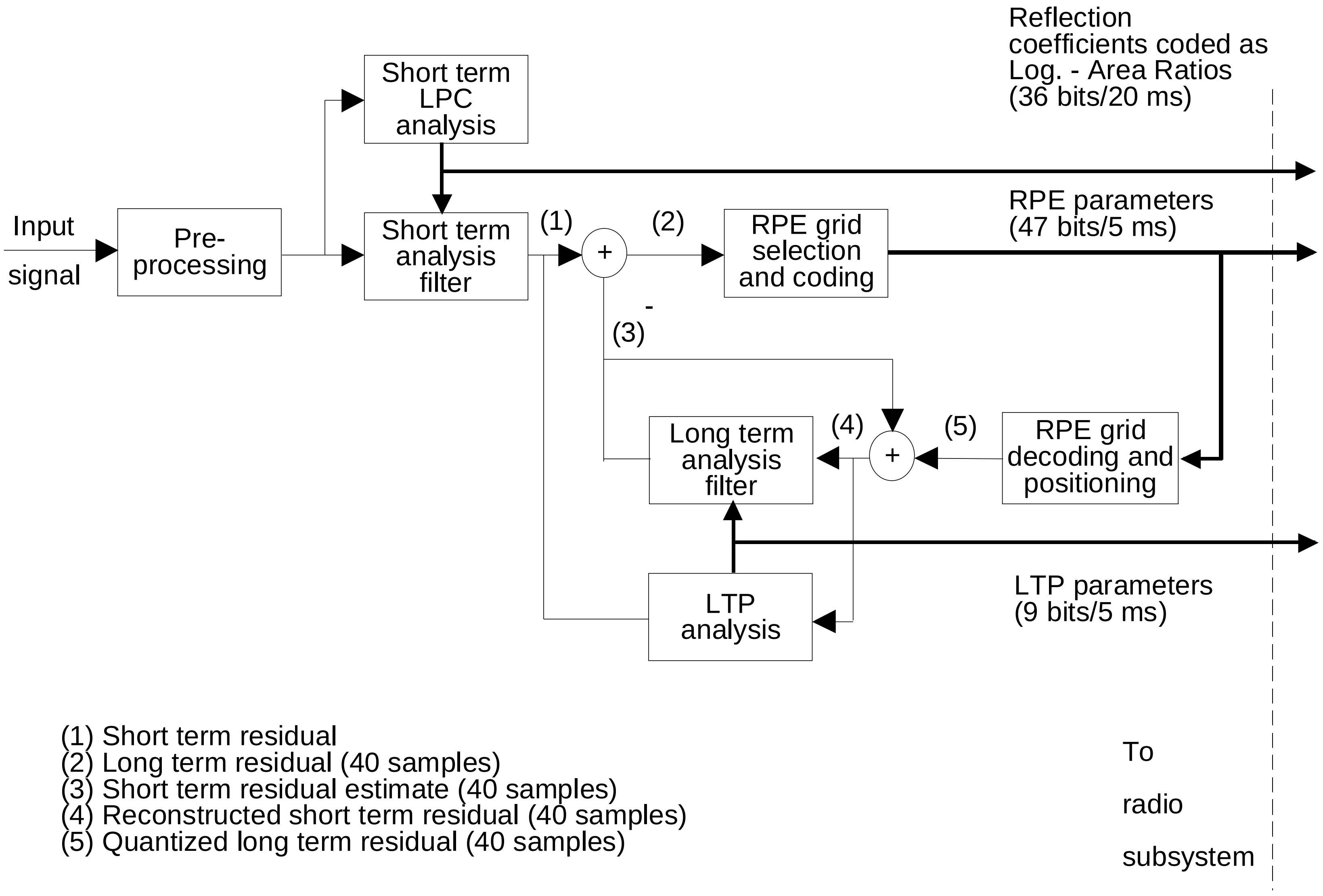
poles of the transfer function: $\frac{1}{A(z)}$ and $\frac{1}{A(z/\gamma)}$:



Excitation Coding in Shorter Frames

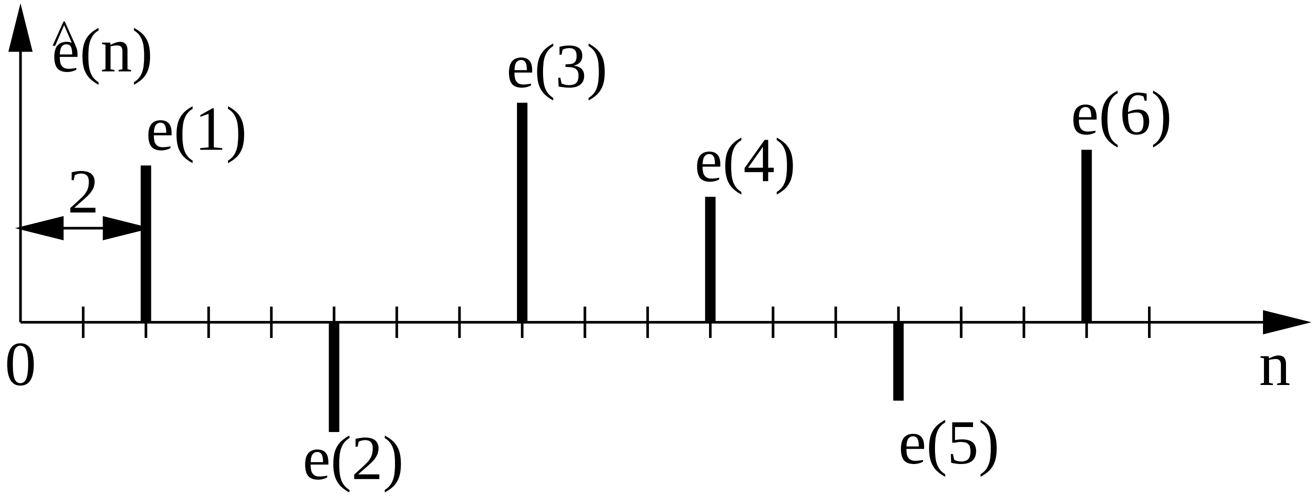
While LP analysis is done over a frame with a "standard" length (usually 20 ms – 160 samples for $F_s = 8\ kHz$), excitation is often coded in shorter frames – typically 40 samples.

ENCODER I: RPE-LTP



(1) Short term residual
(2) Long term residual (40 samples)
(3) Short term residual estimate (40 samples)
(4) Reconstructed short term residual (40 samples)
(5) Quantized long term residual (40 samples)

- Regular-Pulse Excitation, Long Term Prediction, GSM full-rate ETSI 06.10
- Short-term analysis (20 ms frames with 160 samples), the coefficients of a LPC filter are converted to 8 LAR.
- Long-time prediction (LTP - in frames of 5 ms 40 samples) - lag a gain.
- Excitation is coded in frames of 40 samples, where the error signal is down-sampled with the factor 3 (14,13,13), and only the position of the first impulse is quantized (0,1,2,3(!))
- The sizes of the single impulses are quantized using APCM.



- The result is a frame on 260 bits × 50 = 13 kbit/s.
- For more information see the norm 06.10, available for downloading at `http://pda.etsi.org`

Decoder RPE-LTP

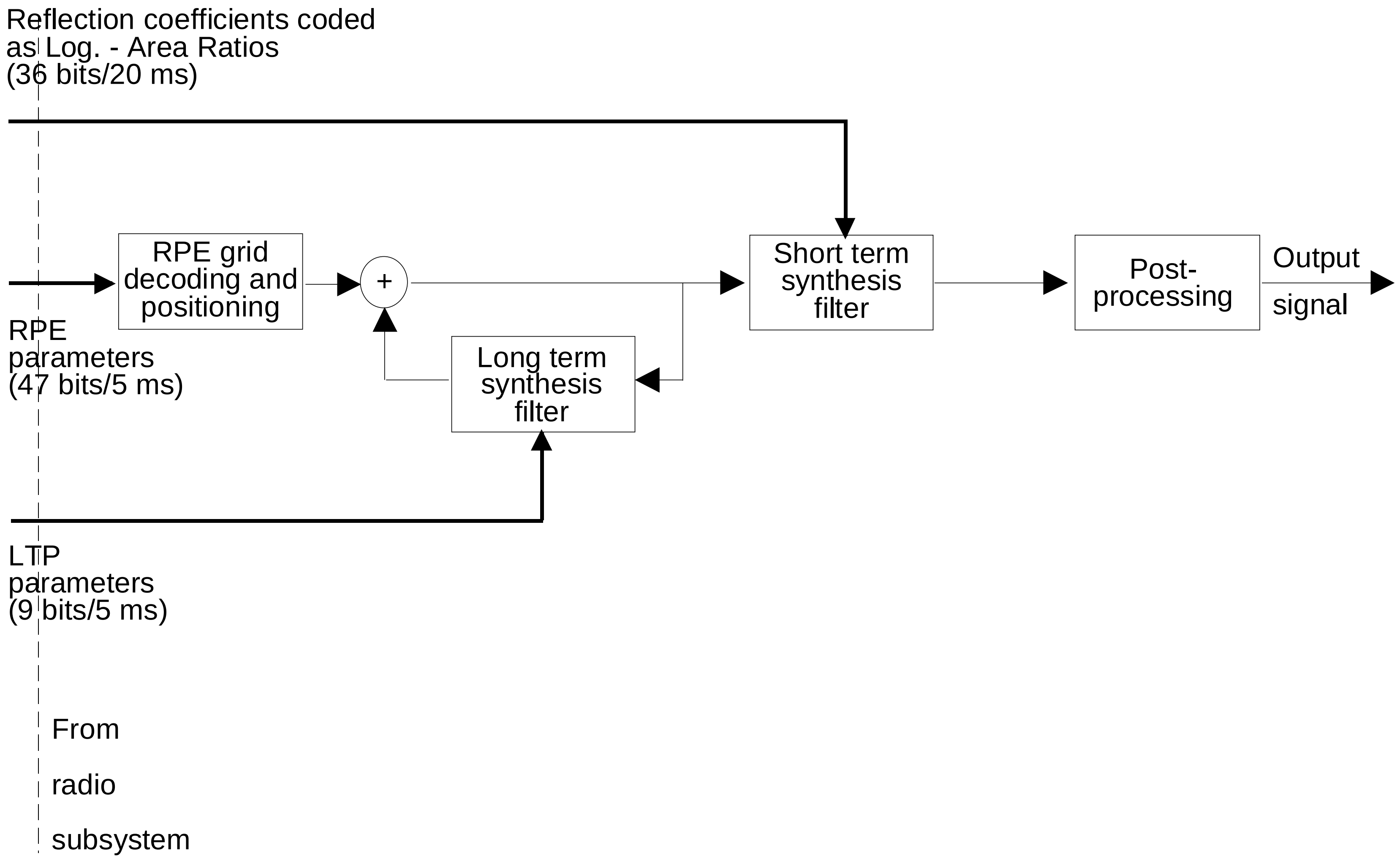
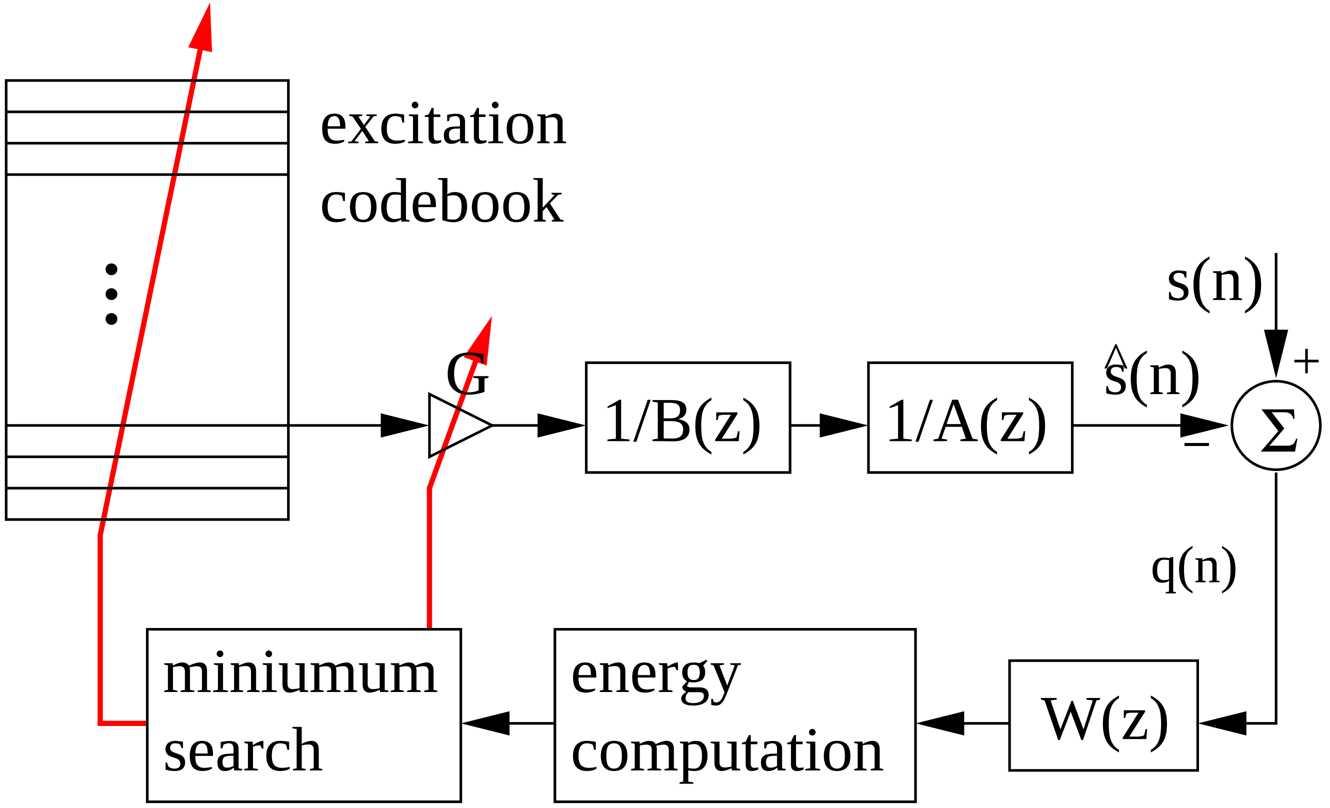


Figure 1.2: Simplified block diagram of the RPE - LTP decoder

CELP – Codebook-Excited Linear Prediction

Excitation is coded using **code-books**

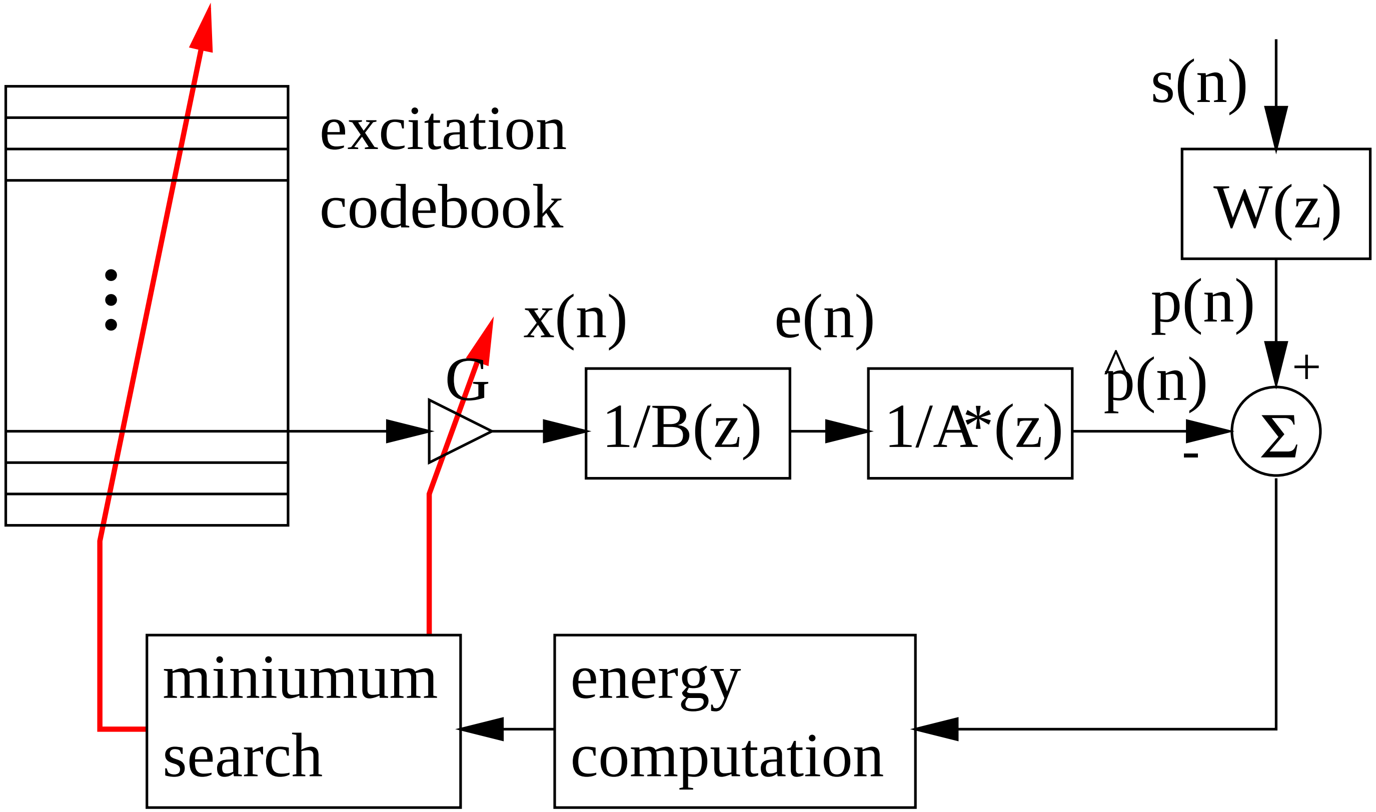
Basic structure with a perceptual filter is:



. . . 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 $\frac{1}{B(z)}$.



Null Input Response Reduction

Filter $\frac{1}{A^{\star}(z)}$ responds not only to excitation but also has memory (delayed samples) Denote the impulse response as $h(i)$:

$$\hat{p}(n) = \underbrace{\sum_{i=0}^{n-1} h(i)e(n-i)}_{\text{tento rámec}} + \underbrace{\sum_{i=n}^{\infty} h(i)e(n-i)}_{\text{minulé rámce } \hat{p}_0(n)} \tag{2}$$

Response from the previous samples does not depend on the current input – the response can be calculated only once and then subtracted from the examined signal (input filtered by $W(z)$), which results in :

$$\hat{p}(n) - \hat{p}_0(n) = \sum_{i=0}^{n-1} h(i)e(n-i) \tag{3}$$

Further approach: consider long-term predictor:

$$\frac{1}{B(z)} = \frac{1}{1 - bz^{-M}}, \tag{4}$$

where M is the optimal lag. $e(n)$ can be defined as

$$e(n) = x(n) + be(n-M), \tag{5}$$

in the filter equation thus instead of $e(n-i)$ we get:

$$\hat{p}(n) - \hat{p}_0(n) = \sum_{i=0}^{n-1} h(i)[x(n-i) + be(n-M-i)]. \tag{6}$$

then expand:

$$\hat{p}(n) - \hat{p}_0(n) = \sum_{i=0}^{n-1} h(i)x(n-i) + b\sum_{i=0}^{n-1} h(i)e(n-M-i), \tag{7}$$

because filtering is a linear operation. The second expression represents **past excitation** filtered by $\frac{1}{A^\star(z)}$ and multiplied by LTP gain b. The LTP in the CELP coder can be interpreted as another code-book containing past excitation. Here, it is considered as an actual code-book with rows $e(n-M)$ (or $\mathbf{e}^{(M)}$ in vector notation). In real applications, it is just a delayed exciting signal.

CELP Equations

$$\hat{p}(n)-\hat{p}_0(n)=\sum_{i=0}^{n-1}h(i)e(n-i)=\sum_{i=0}^{n-1}h(i)\left[g^{(j)}u^{(j)}(n-i)+be(n-i-M)\right]=\hat{p}_2(n)+\hat{p}_1(n), \tag{8}$$

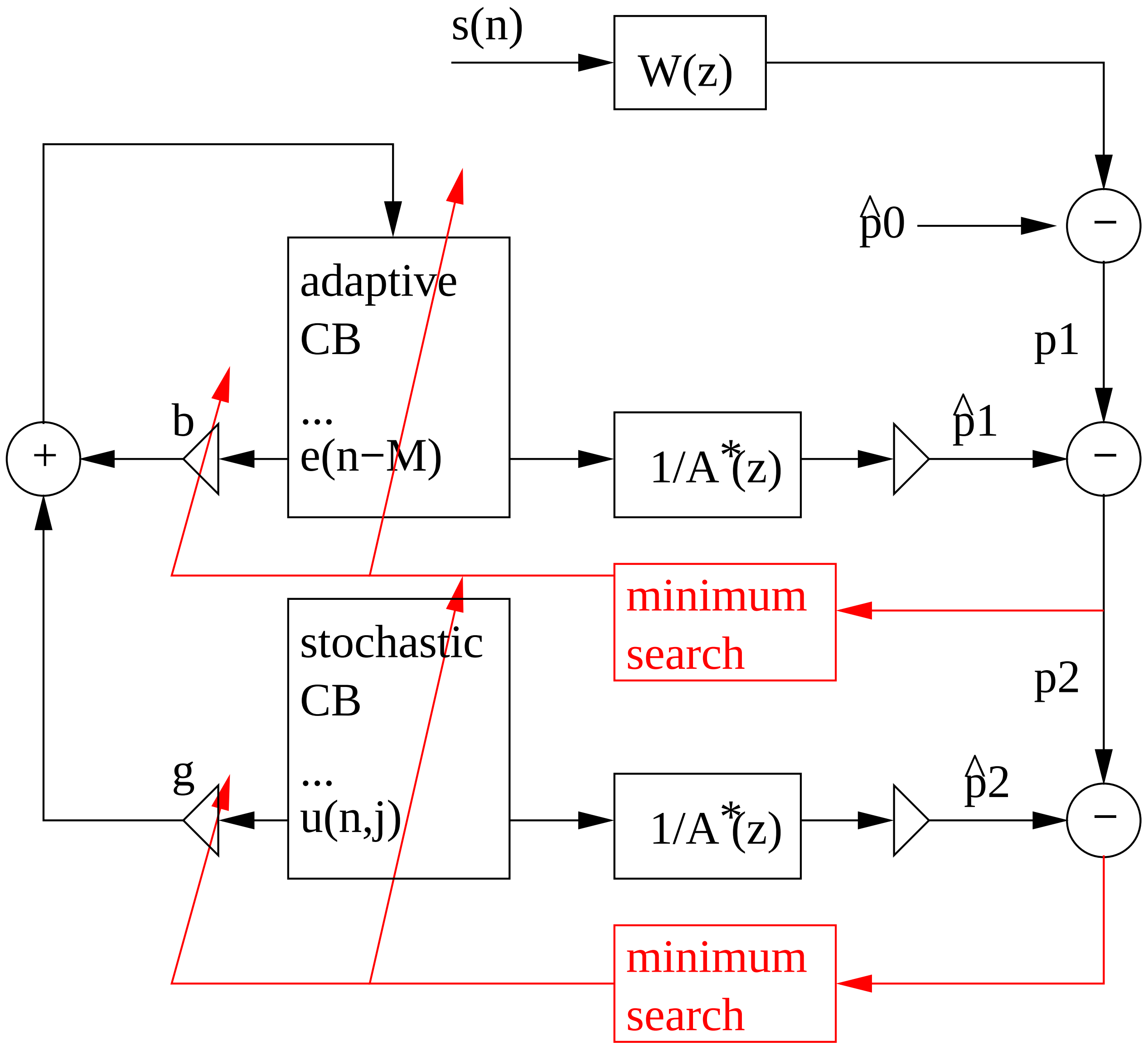
Task is to find:

- gain for the stochastic code-book g
- the best vector from the stochastic code-book $\mathbf{u}^{(j)}$
- gain of the adaptive code-book b
- the best vector from the best adaptive code-book $\mathbf{e}^{(M)}$.

Theoretically, all the combinations have to be tested ⇒ no ! Suboptimal procedure:

- first find, $\hat{p}_1(n)$ by searching in the adaptive code-book. The result is lag M and gain b.
- Subtract $\hat{p}_1(n)$ from $p_1(n)$ and get "an error signal of the second generation", which should be provided by the stochastic code-book.
- Find $\hat{p}_2(n)$ in the stochastic code-book. The result is the index j and gain g.
- As the last step usually, gains are optimized (contribution from both code-books).

Final Structure of CELP:



CELP equation – optional. . .

Searching the adaptive codebook

We want to minimize the error E_1 that is (for one frame):

$$E_1 = \sum_{N=0}^{N} (p_1(n) - \hat{p}_1(n))^2. \tag{9}$$

We can rewrite it in more convenient vector notation and develop $\hat{p}_1$ into the gain g and excitation signal $q(n-M)$, where $q(n)$ is the filtered version of $e(n-M)$:

$$q(n) = \sum_{i=0}^{n-1} h(i)e(n-i-M) \tag{10}$$

When we multiply this signal by the gain b, we get the signal

$$\hat{p}_1(n) = bq(n). \tag{11}$$

The goal is to find the minimum energy:

$$E_1 = ||\mathbf{p}_1 - \hat{\mathbf{p}}_1||^2 = ||\mathbf{p}_1 - b\mathbf{q}^{(M)}||^2. \tag{12}$$

which we can rewrite using inner products of involved vectors:

$$E_1 =< \mathbf{p}_1 - b\mathbf{q}^{(M)}, \mathbf{p}_1 - b\mathbf{q}^{(M)} >=< \mathbf{p}_1, \mathbf{p}_1 > -2b < \mathbf{p}_1, \mathbf{q}^{(M)} > +b^2 < \mathbf{q}^{(M)}, \mathbf{q}^{(M)} > . \tag{13}$$

We have to begin by finding an optimal gain for each M by a derivation with respect to b (the derivation must be zero for minimum energy):

$$\frac{\delta}{\delta b}\left[< \mathbf{p}_1, \mathbf{p}_1 > -2b < \mathbf{p}_1, \mathbf{q}^{(M)} > +b^2 < \mathbf{q}^{(M)}, \mathbf{q}^{(M)} >\right] = 0 \tag{14}$$

from here, we can directly write the result for lag M:

$$b^{(M)} = \frac{< \mathbf{p}_1, \mathbf{q}^{(M)} >}{< \mathbf{q}^{(M)}, \mathbf{q}^{(M)} >} \tag{15}$$

We need now to determine the optimal lag. We will substitute found $b^{(M)}$ in Eq. 13 and we obtain: the minimization of

$$\min\left[<\mathbf{p}_1,\mathbf{p}_1> - \frac{2<\mathbf{p}_1,\mathbf{q}^{(M)}><\mathbf{p}_1,\mathbf{q}^{(M)}>}{<\mathbf{q}^{(M)},\mathbf{q}^{(M)}>} + \frac{<\mathbf{p}_1,\mathbf{q}^{(M)}>^2<\mathbf{q}^{(M)},\mathbf{q}^{(M)}>}{<\mathbf{q}^{(M)},\mathbf{q}^{(M)}>^2}\right]. \tag{16}$$

we can drop the first term, as it does not influence the minimization and convert minimization into the maximization of:

$$M = \arg\max \frac{<\mathbf{p}_1,\mathbf{q}^{(M)}>^2}{<\mathbf{q}^{(M)},\mathbf{q}^{(M)}>} \tag{17}$$

Searching the stochastic codebook

First, all codebook vectors $\mathbf{u}^{(j)}(n)$ must be filtered by the filter $\frac{1}{A^{\star}(z)} \longrightarrow \mathbf{q}^{(j)}$. Then, we have to search $\hat{\mathbf{p}}_2(n)$ minimizing the energy of error:

$$E_2 = ||\mathbf{p}_2 - \hat{\mathbf{p}}_2||^2 = ||\mathbf{p}_2 - g\mathbf{q}^{(j)}||^2$$

Solution:

- $< \mathbf{p}_2 - g\mathbf{q}^{(j)}, \mathbf{q}^{(j)} >= 0$

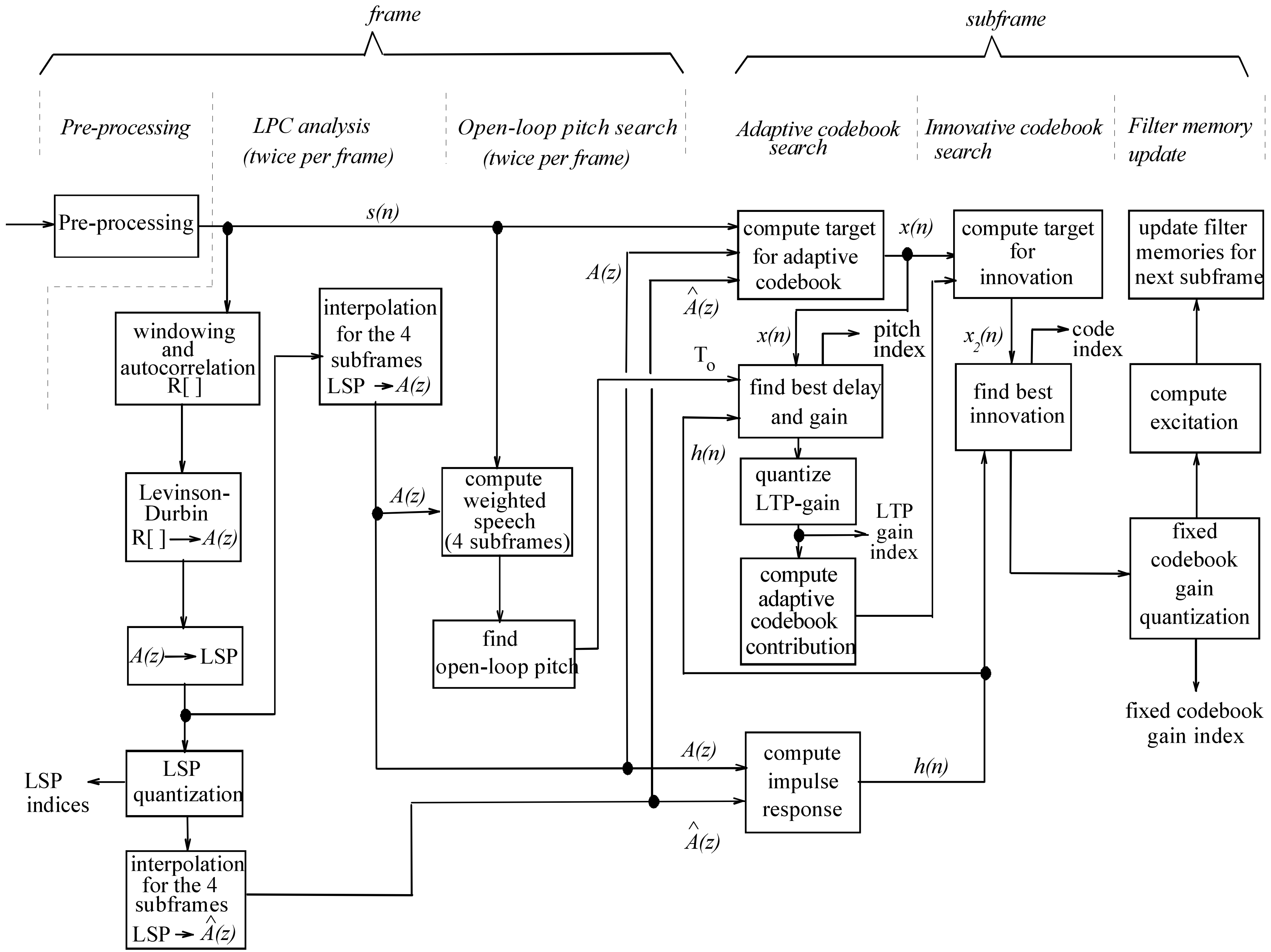
$$g = \frac{< \mathbf{p}_2, \mathbf{q}^{(j)} >}{< \mathbf{q}^{(j)}, \mathbf{q}^{(j)} >}.$$

- $\min E_2 = \min < \mathbf{p}_2 - g\mathbf{q}^{(j)}, \mathbf{p}_2 >= ||\mathbf{p}_2||^2 - \frac{<\mathbf{p}_2, \mathbf{q}^{(j)}>^2}{<\mathbf{q}^{(j)}, \mathbf{q}^{(j)}>}.$

$$j = \arg\max \frac{< \mathbf{p}_2, \mathbf{q}^{(j)} >^2}{< \mathbf{q}^{(j)}, \mathbf{q}^{(j)} >}.$$

Example of a CELP encoder ACELP – GSM EFR

- classical CELP with an “intelligent code-book”.
- Algebraic Codebook Excited Linear Prediction GSM Enhanced Full-rate ETSI 06.60



frame

subframe

Pre-processing

LPC analysis (twice per frame)

Open-loop pitch search (twice per frame)

Adaptive codebook search

Innovative codebook search

Filter memory update

Pre-processing

$s(n)$

windowing and autocorrelation R[ ]

Levinson-Durbin R[ ] → $A(z)$

$A(z)$ → LSP

LSP quantization

LSP indices

interpolation for the 4 subframes LSP → $\hat{A}(z)$

interpolation for the 4 subframes LSP → $A(z)$

$A(z)$

compute weighted speech (4 subframes)

find open-loop pitch

T_o

compute target for adaptive codebook

$A(z)$

$\hat{A}(z)$

$x(n)$

find best delay and gain

pitch index

$h(n)$

quantize LTP-gain

LTP gain index

compute adaptive codebook contribution

compute impulse response

$A(z)$

$\hat{A}(z)$

$h(n)$

compute target for innovation

$x_2(n)$

find best innovation

code index

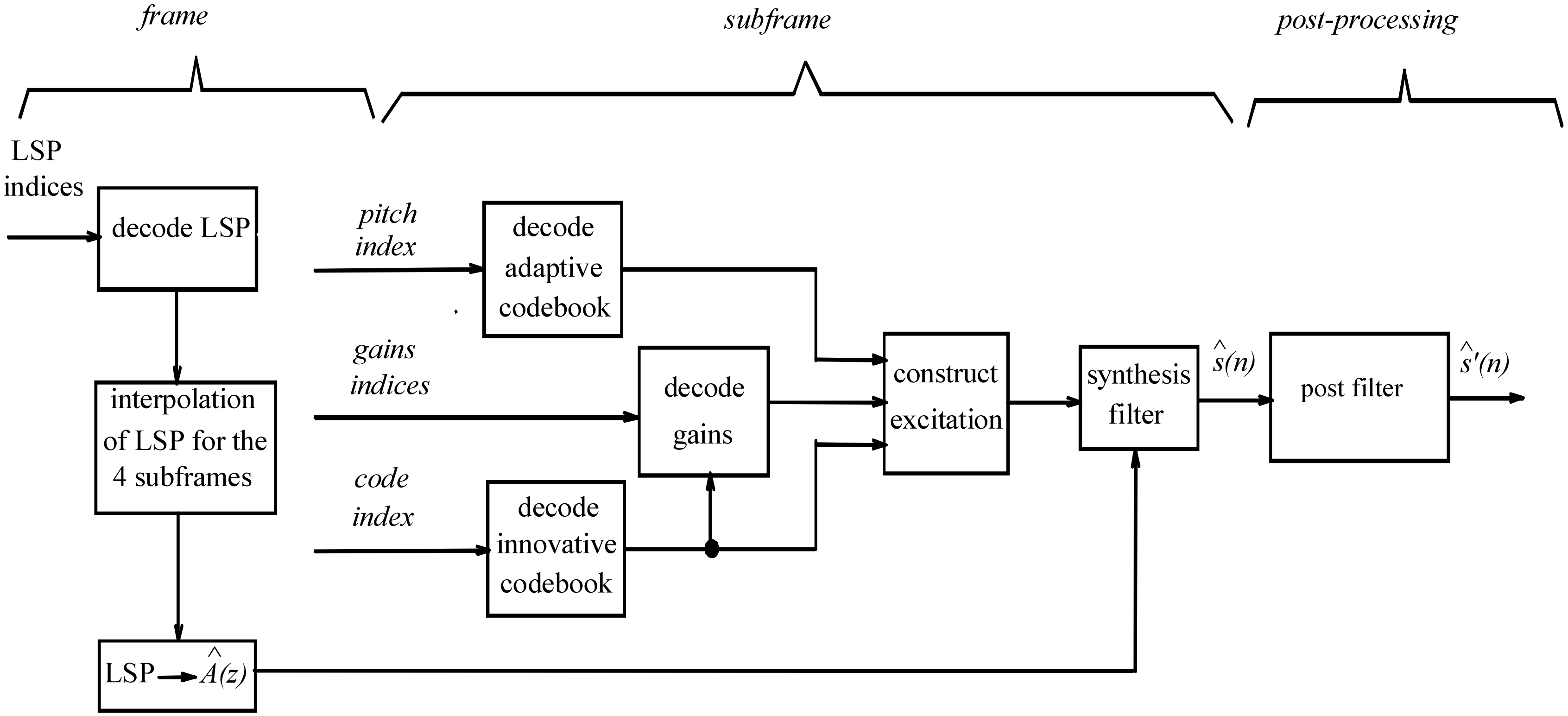
update filter memories for next subframe

compute excitation

fixed codebook gain quantization

fixed codebook gain index

- Again frames of 20 ms (160 samples)
- Short-term predictor – 10 coefficients a_i in two sub-frames, converted to line-spectral pairs, jointly quantized using split-matrix quantization (SMQ).
- Excitation: 4 sub-frames of 40 samples (5 ms).
- Lag estimation first as open-loop, then as closed-loop from the rough estimation, fractional pitch with the resolution of 1/6 of a sample.
- stochastic codebook: algebraic code-book - can contain only 10 nonzero impulses, that can be either +1 or -1 ⇒ fast search (fast correlation - only summation, no multiplication), etc.
- 244 bits per frame $\times$ 50 $=$ 12.2 kbit/s.
- for more information see norm 06.60, available fro download at `http://pda.etsi.org`
- decoder...


frame
subframe
post-processing
LSP indices
decode LSP
interpolation of LSP for the 4 subframes
LSP → $\hat{A}(z)$
pitch index
decode adaptive codebook
gains indices
decode gains
code index
decode innovative codebook
construct excitation
synthesis filter
$\hat{s}(n)$
post filter
$\hat{s}'(n)$

Additional Reading

- Andreas Spanias (Arizona University):
  `http://www.eas.asu.edu/~spanias`
  section Publications/Tutorial Papers provides a good outline: “Speech Coding: A Tutorial Review”, a part of it is printed in Proceedings of the IEEE, Oct. 1994.
- at the same pade in section Software/Tools/Demo - Matlab Speech Coding Simulations, software for FS1015, FS1016, RPE-LTP and more. Nice playing.
- Norm ETSI for cell phones are available free for download from:
  `http://pda.etsi.org/pda/queryform.asp`
  As the keywords enter fro instance “gsm half rate speech”. Many norms are provided with the source code of the coders in C.