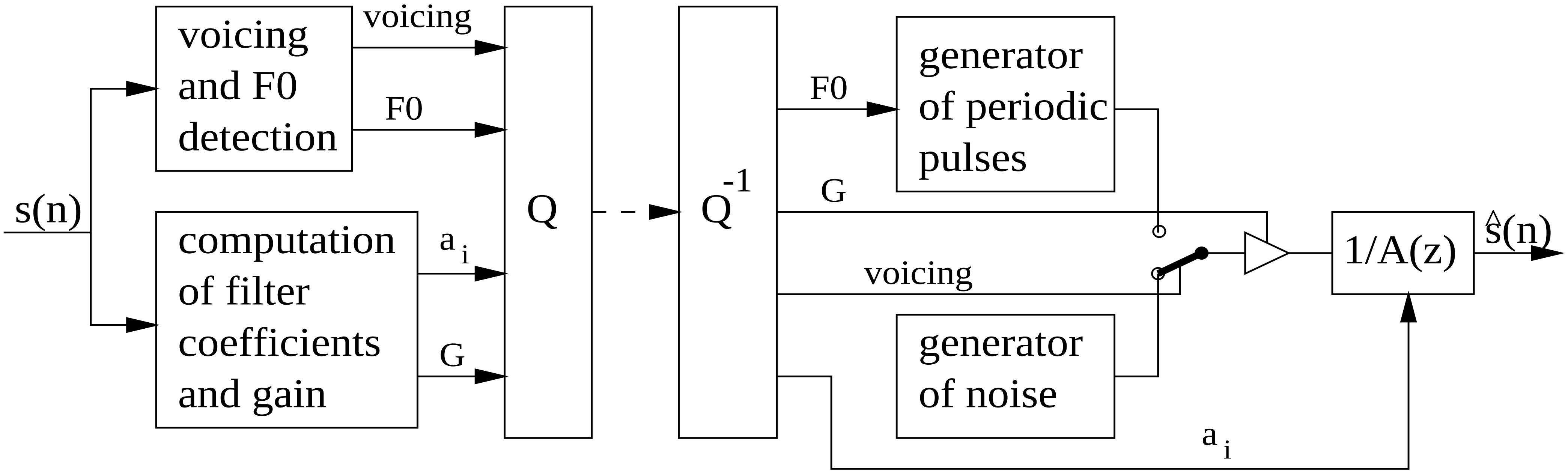
Coding based on Linear Predictive Modeling - LPC

Petr Motlíček, Mathew M-Doss

Idiap Research Institute, {motlicek,mathew}@idiap.ch

- General scheme of LPC coder
- LPC in details
- LTP
- Analysis-by-Synthesis
- Perceptual filter
- RPE-LTP – GSM full-rate
- CELP
- ACELP – GSM enhanced full rate

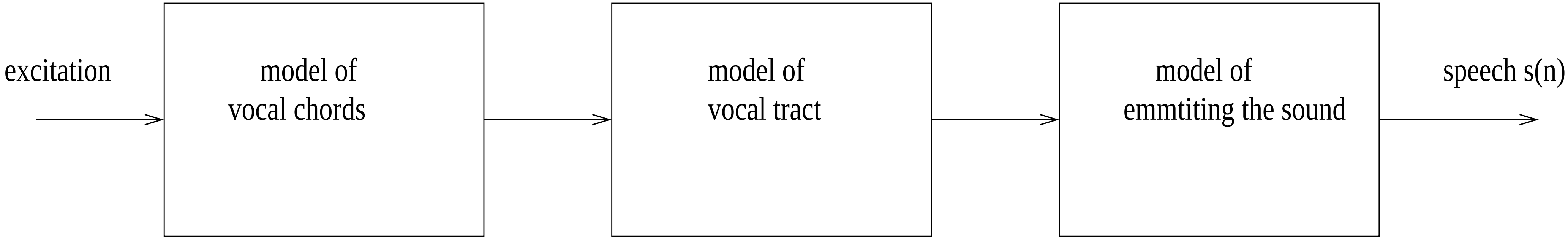
The input signal is split into segments, and a polynomial function is solved for each segment: $A(z) = 1 + \sum_{i=1}^{P} a_i z^{-i}$. A Gain G and voicing is detected. In case of voiced segment, fundamental frequency is found:



Example: US-DoD FS1015 standard: filter 1800 bps, excitation 600 bps, draw-back - very simple modeling of excitation signal $\Rightarrow$ non-natural speech. Large improvement in CELP coders.

Residual Excited Linear Prediction – RELP: for each frame, $A(z)$ are found and an error signal is estimated: $e(n)$: $E(z) = A(z)S(z)$. Such the signal is filtered by a filter: $H(z) = \frac{1}{A(z)}$. If a_i and $e(n)$ are not quantized, original $s(n)$ is obtained. But very high bit-rates.

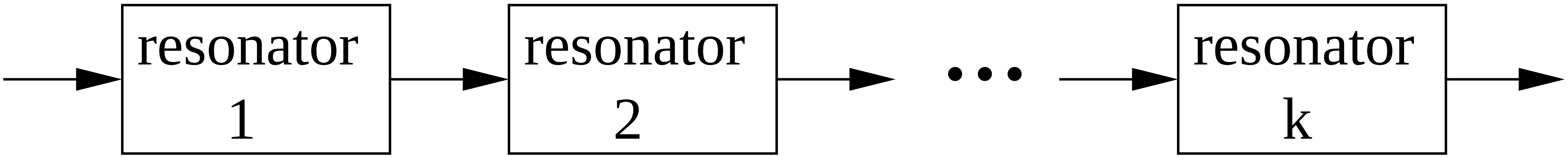
LPC



Vocal chords: low-pass filter

$$G(z) = \frac{1}{[1 - e^{-cT_s} z^{-1}]^2} \tag{1}$$

Vocal tract: cascade of 2-pole resonators:



$$V(z) = \frac{1}{\prod_{i=1}^{K} [1 - 2e^{-\alpha_i T_s} \cos \beta_i T_s z^{-1} + e^{-2\alpha_i T_s} z^{-2}]} \tag{2}$$

Model of emitting the sound: high-pass filter

$$L(z) = 1 - z^{-1} \tag{3}$$

Together:

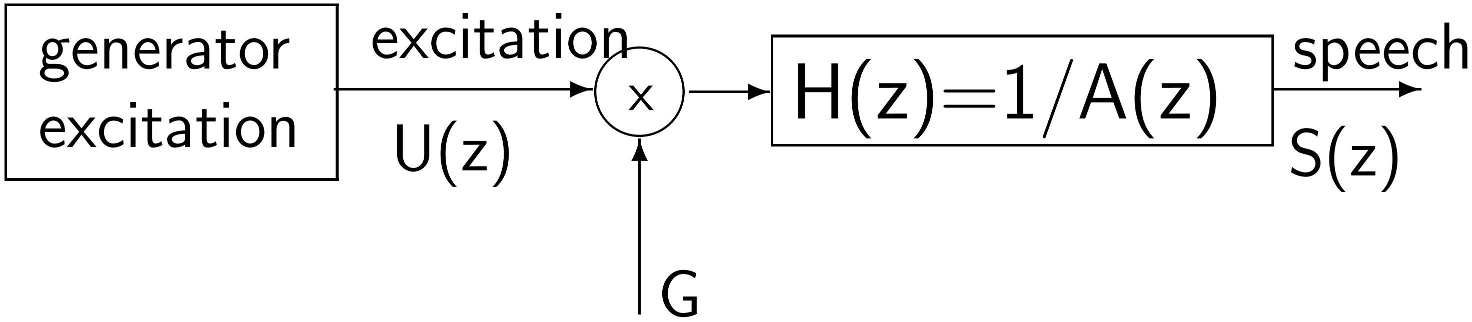
$$\begin{aligned} H(z) &= G(z)V(z)L(z) = \\ &= \frac{1 - z^{-1}}{(1 - e^{-cT_s} z^{-1})^2 \prod_{i=1}^{K} [1 - 2e^{-\alpha_i T_s} \cos \beta_i T_s z^{-1} + e^{-2\alpha_i T_s} z^{-2}]} \end{aligned} \tag{4}$$

$cT_s \to 0$ thus we can write

$$H(z) = \frac{1}{1 + \sum_{i=1}^{P} a_i z^{-i}} = \frac{1}{A(z)}, \tag{5}$$

where polynomial function $A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_P z^{-P}$ has order $P = 2k + 1$ (k - number of formants).

Speech generation using this filter

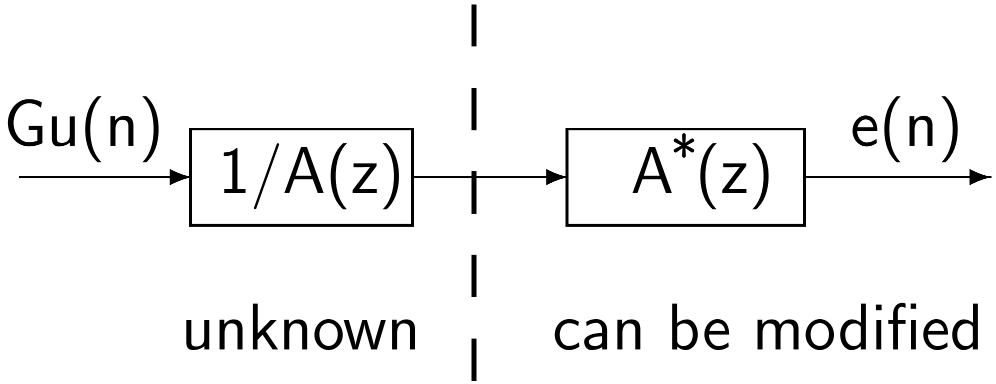


n-sample of the speech is given as:

$$s(n) = Gu(n) - \sum_{i=1}^{P} a_i s(n-i) \tag{6}$$

Estimation of filter parameters

We can construct an inverse filter $A^{\star}(z)$ with coefficients α_i:



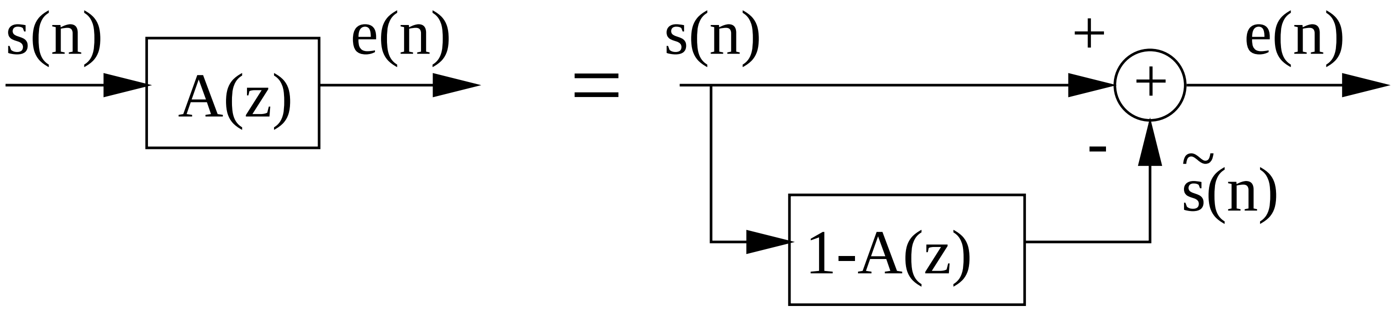
It can be shown that in case of stationary signal $s(n)$, the coefficients of a_i are identified by α_i if the energy of $e(n)$ at the output is minimized: $\mathcal{E}\{e^2(n)\}$. We tune parameters until the energy reaches the minimum ...

Why Linear Prediction

We suppose that $\mathcal{E}\{e^2(n)\}$ is already minimized, so that $A^\star(z) = A(z)$ and we can start using only a_i. $A(z)$ can be written as:

$$A(z) = 1 - [1 - A(z)] \tag{7}$$

and thus:



Signal $s(n)$ is given as a linear combination of previous samples:

$$\tilde{s}(n) = -\sum_{i=1}^{P} a_i s(n-i) \tag{8}$$

Prediction error:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - [-\sum_{i=1}^{P} a_i s(n-i)] = s(n) + \sum_{i=1}^{P} a_i s(n-i). \tag{9}$$

In z-domain:

$$E(z) = S(z)A(z) \tag{10}$$

Solution

We do not mention how many samples of the input signal is available. Un-normalized energy of the error signal:

$$E = \sum_n e^2(n) \tag{11}$$

will be minimized:

$$\begin{aligned}
\frac{\delta}{\delta a_j}\left\{\sum_n [s(n) + \sum_{i=1}^{P} a_i s(n-i)]^2\right\} &= 0 & (12)\\
\sum_n 2[s(n) + \sum_{i=1}^{P} a_i s(n-i)]s(n-j) &= 0 & (13)\\
\sum_n s(n)s(n-j) + \sum_{i=1}^{P} a_i \sum_n s(n-i)s(n-j) &= 0. & (14)\\
& & (15)
\end{aligned}$$

If:

$$\sum_{n} s(n-i)s(n-j) = \phi(i,j), \tag{16}$$

then

$$\sum_{i=1}^{P} a_i \phi(i,j) = -\phi(0,j) \quad \text{pro} \quad 1 \leq j \leq P \tag{17}$$

which leads to a set of linear equations:

$$\begin{array}{ccc} \phi(1,1)a_1 + \phi(2,1)a_2 + & \cdots & +\phi(P,1)a_P = -\phi(0,1) \\ \phi(1,2)a_1 + \phi(2,2)a_2 + & \cdots & +\phi(P,2)a_P = -\phi(0,2) \\ & \vdots & \\ \phi(1,P)a_1 + \phi(2,P)a_2 + & \cdots & +\phi(P,P)a_P = -\phi(0,P), \end{array} \tag{18}$$

In case of a correlation technique:

$$
\begin{aligned}
R(0)a_1 + R(1)a_2 + \quad &\cdots \quad +R(P-1)a_P = -R(1) \\
R(1)a_1 + R(0)a_2 + \quad &\cdots \quad +R(P-2)a_P = -R(2) \\
&\vdots \\
R(P-1)a_1 + R(P-2)a_2 + \quad &\cdots \quad +R(0)a_P = -R(P),
\end{aligned}
\tag{19}
$$

where

$$
R(k) = \sum_{n=0}^{N-1-k} s(n)s(n+k)
$$

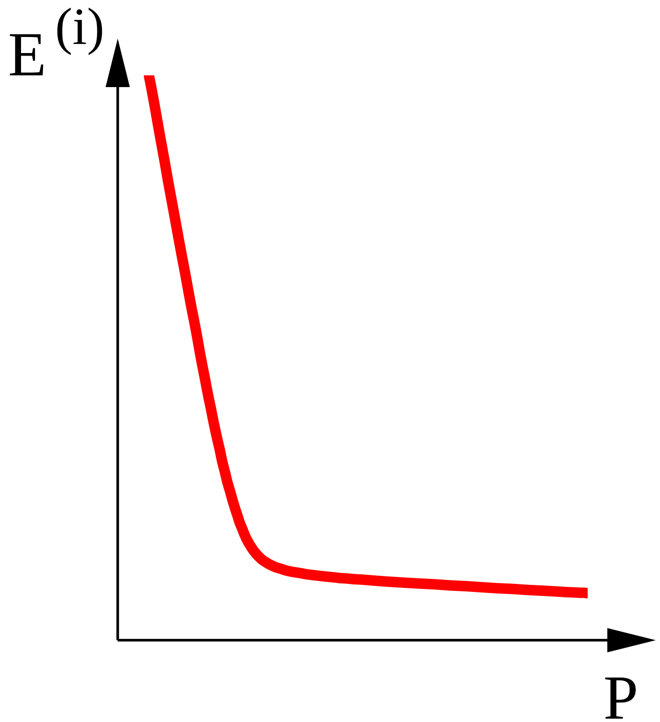
Energy of an error signal

Un-normalized energy of an error signal:

$$E = \sum_{n=0}^{N+P-1} e^2(n) = R(0) + \sum_{i=1}^{P} a_i R(i) \tag{20}$$

In order to achieve the same energy as of the input signal $s(n)$, the gain needs to be set as:

$$G^2 = \frac{E}{N} = \frac{1}{N}\left[R(0) + \sum_{i=1}^{P} a_i R(i)\right]. \tag{21}$$


E (i)
P

LSF or LSP

Line Spectrum Frequencies (LSF) or *Line Spectrum Pairs (LSP)*, are derived from two polynomials:

$$\begin{aligned} M(z) &= A(z) - z^{-(P+1)}A(z^{-1}) \\ Q(z) &= A(z) + z^{-(P+1)}A(z^{-1}). \end{aligned} \tag{22}$$

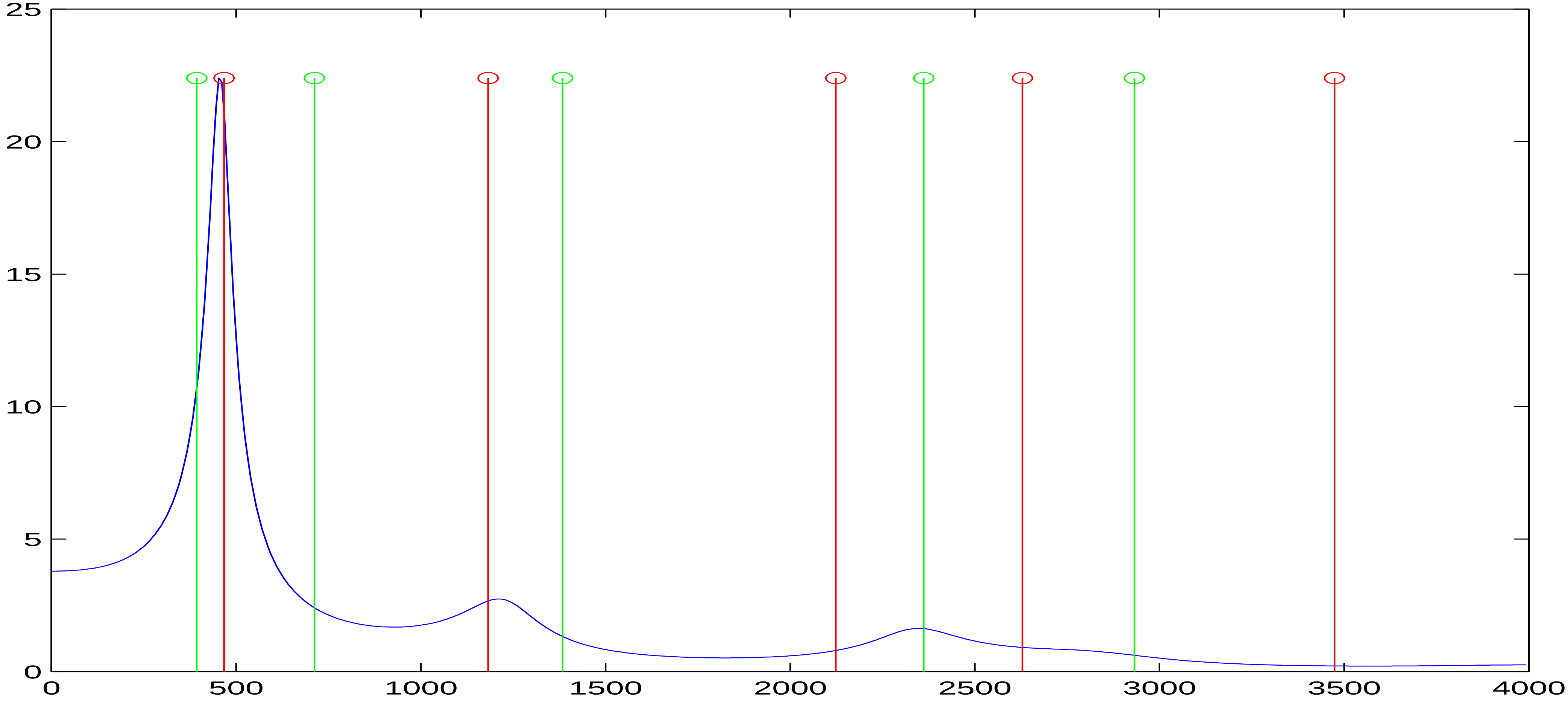
Or:

$$\begin{aligned} M(z) &= (1 - z^{-1}) \prod_{i=2,4,\ldots,P} (1 - 2z^{-1}\cos\omega_i + z^{-2}) \\ Q(z) &= (1 + z^{-1}) \prod_{i=1,3,\ldots,P-1} (1 - 2z^{-1}\cos\omega_i + z^{-2}). \end{aligned} \tag{23}$$

where ω is normalized frequency $\omega = 2\pi f$ (f is normalized "normal" frequency). Line spectral frequencies f_i are in interval (0,0.5) and are ordered from loweest to the highest:

$$0 < f_1 < f_2 < \ldots < f_{P-1} < f_P < \frac{1}{2}. \tag{24}$$

If LSFs are used for coding (quantized), we can test their "sorting" in the decoder:



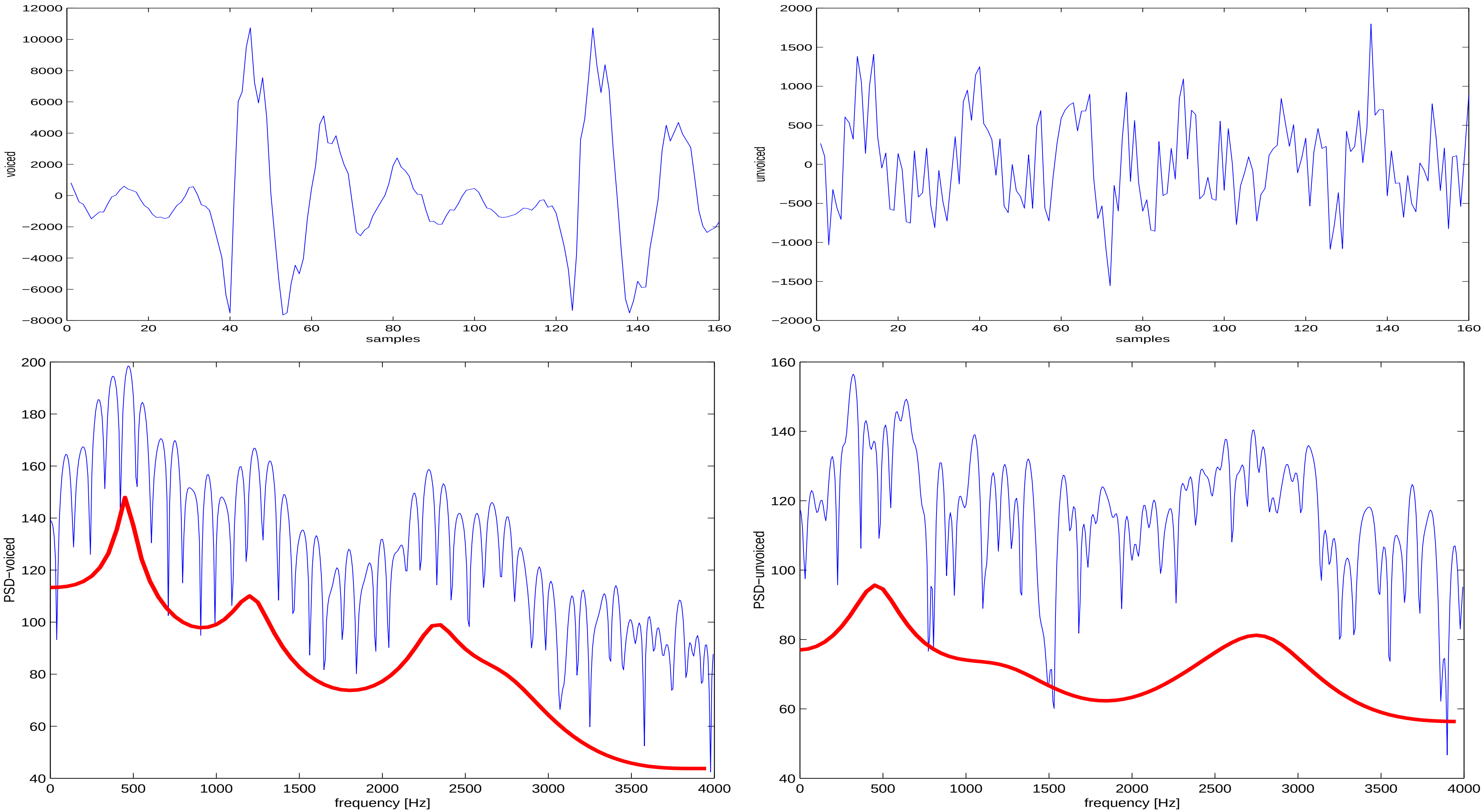
Estimation of spectral density function using LPC coefficients

$$\hat{G}_{LPC} = \left| \frac{G}{A(z)} \right|^2_{z=e^{j2\pi f}}, \tag{25}$$

where f is an frequency $f = \dfrac{F}{F_s}$. Then:

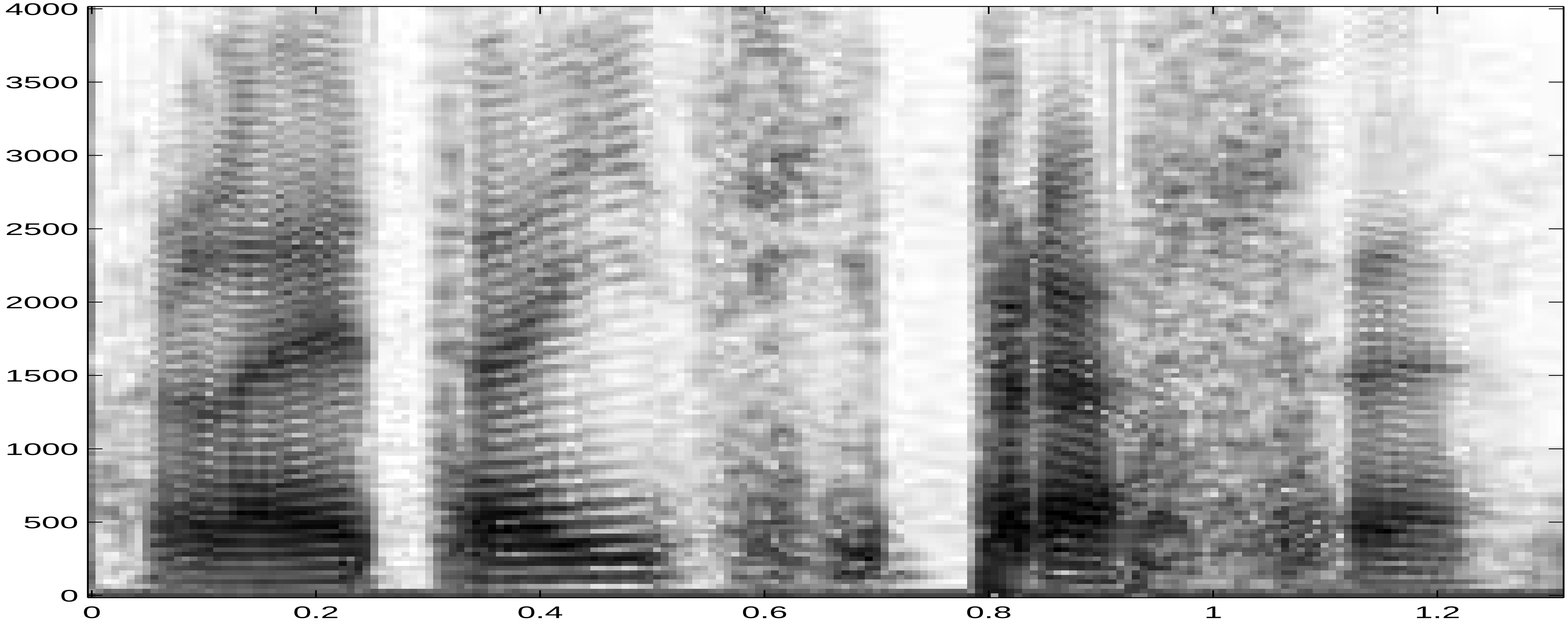
$$\hat{G}_{LPC} = \frac{G^2}{\left| 1 + \sum_{i=1}^{P} a_i e^{-j2\pi f i} \right|^2} \tag{26}$$

Example: spectral density function estimate using DFT and LPC on an voiced and un-voiced speech segment:

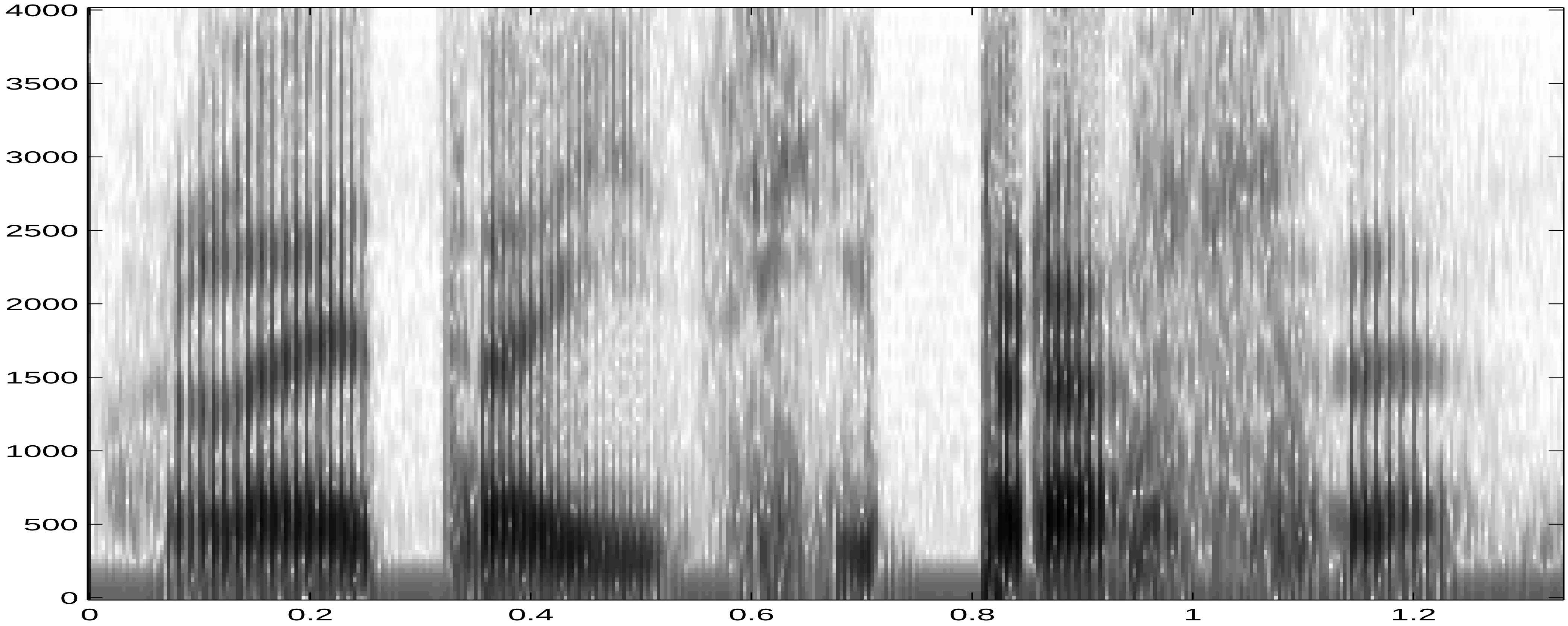


Comarison of spectrograms:

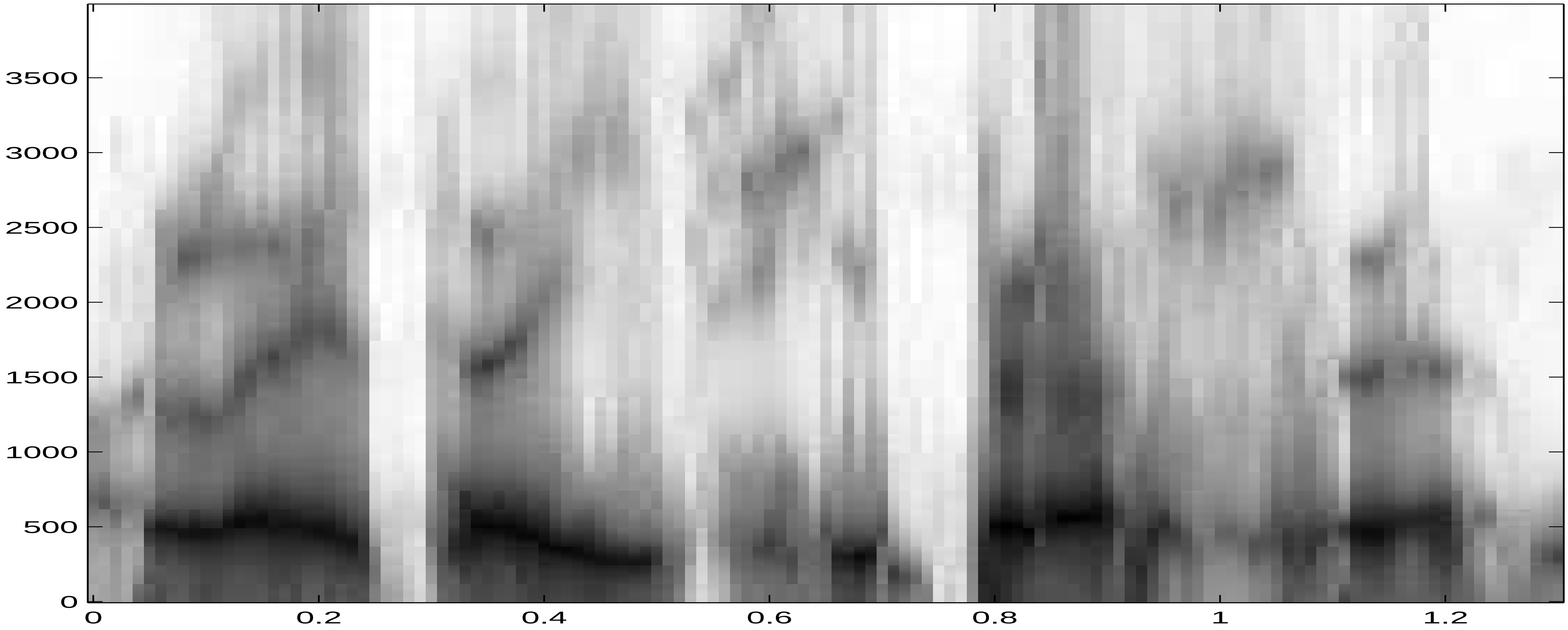
a) DFT `specgram(s,256,8000,hamming(256),200);`



b) DFT `specgram(s,256,8000,hamming(50));`



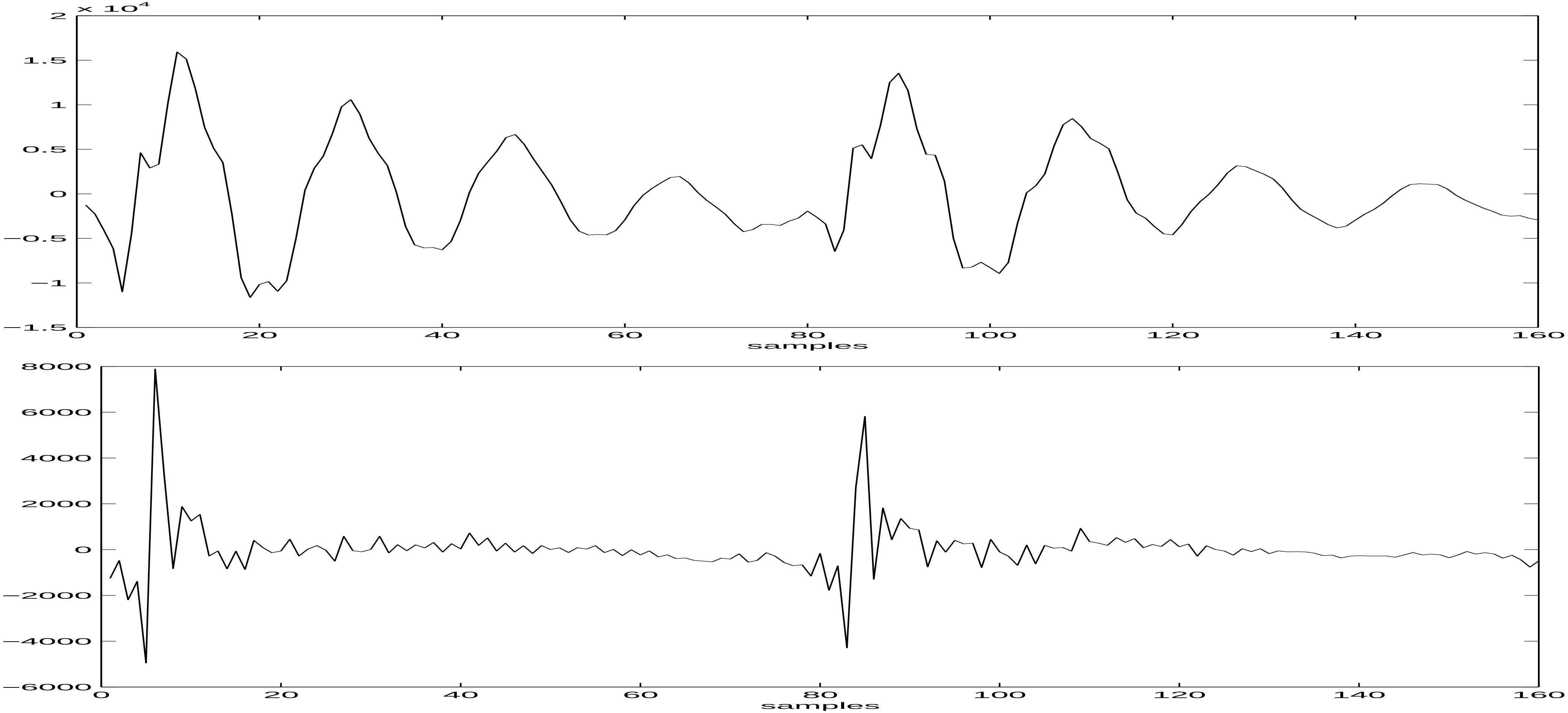
LPC spectrogram:



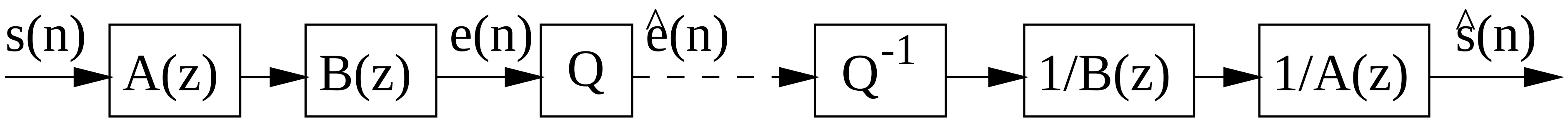
What next:

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

Long-Term Prediction - LTP

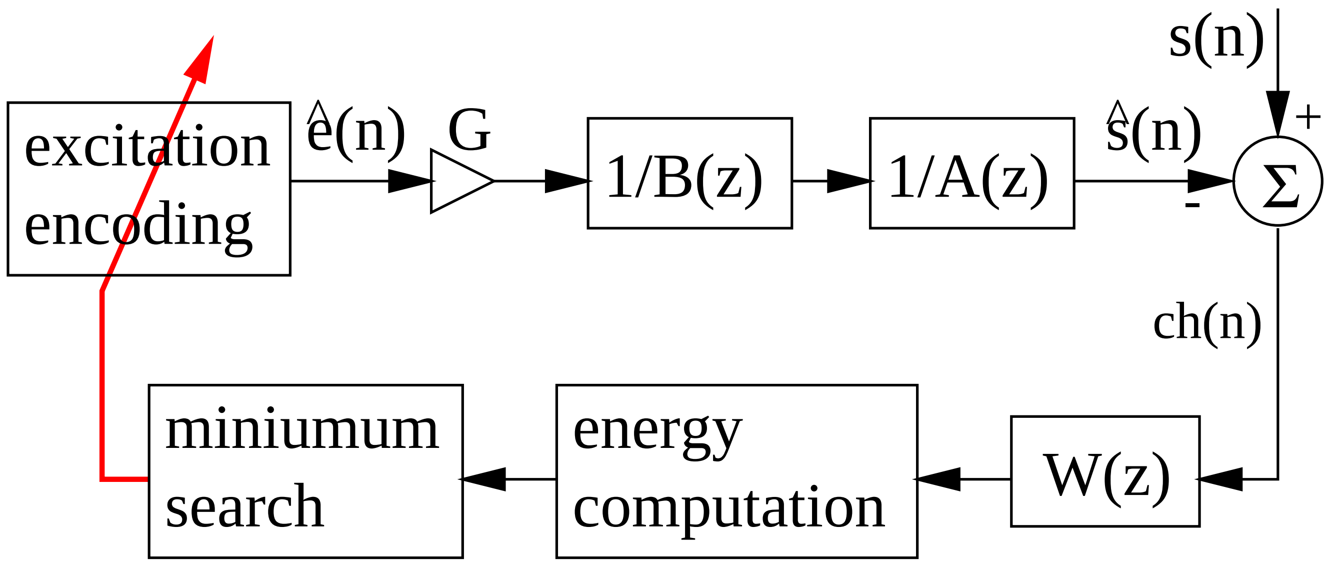


Error signal: $e(n) = s(n) - \hat{s}(n) = s(n) - [-bs(n-L)] = s(n) + bs(n-L)$,
thus $B(z) = 1 + bz^{-L}$.



Analysis-by-Synthesis

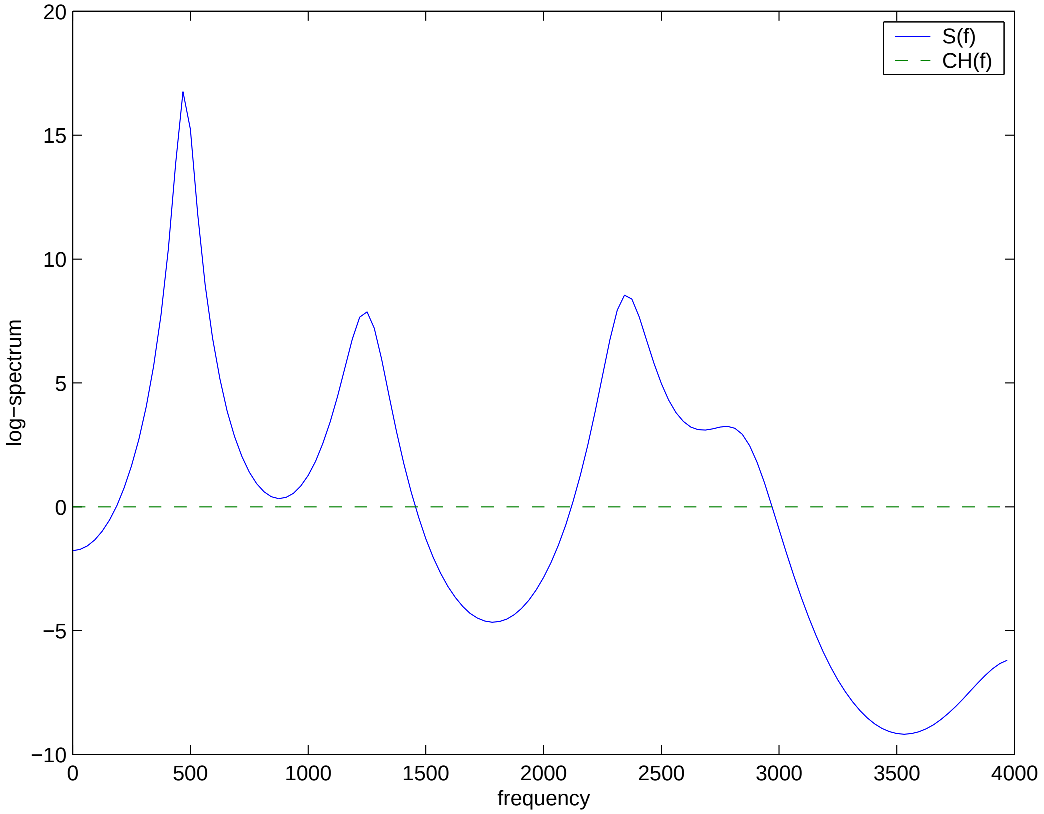
We are not able to find analytically (an equation) the optimal excitation signal $\Rightarrow$ closed loop

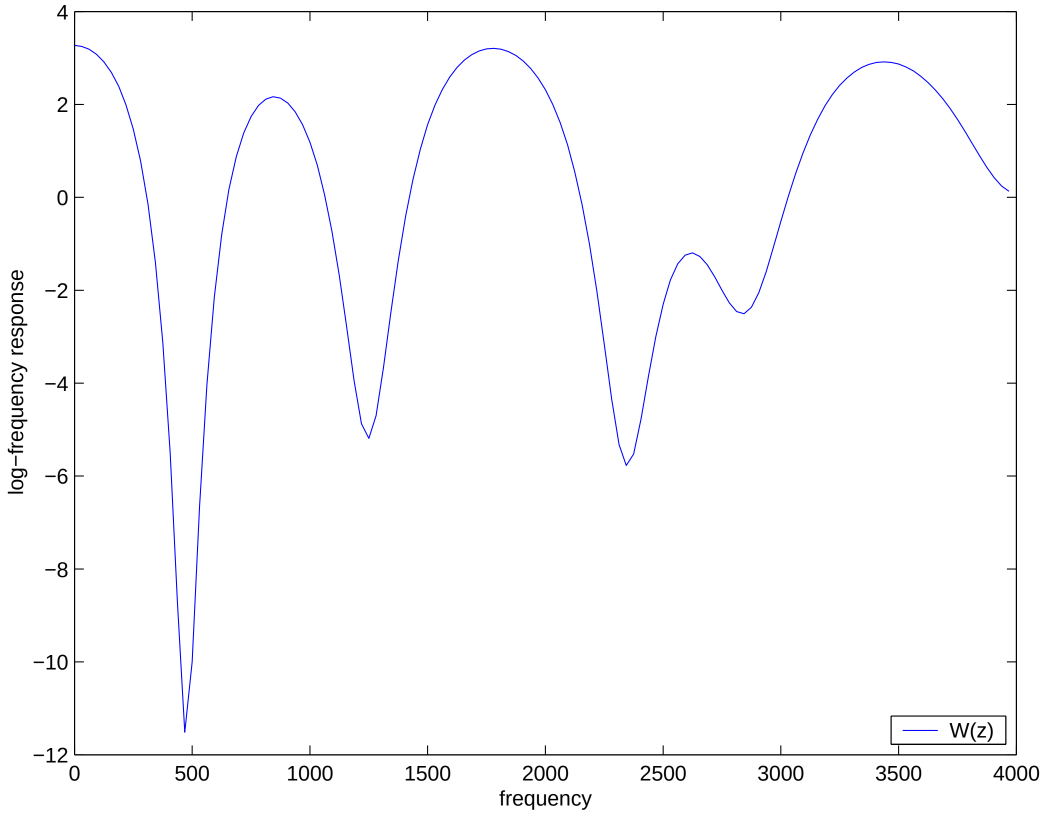


Perceptual filter

In closed-loop, a synthesized signal is compared to the original Perceptual filter - approaching the coder to the human hearing.

Comparison of spectra of the error signal and of the speech:

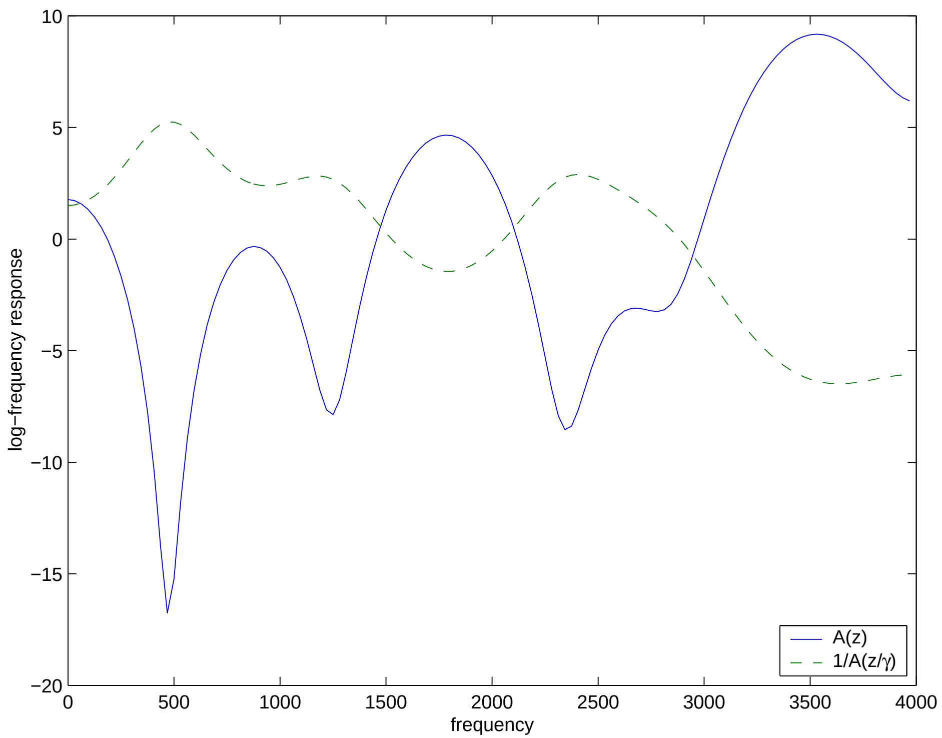


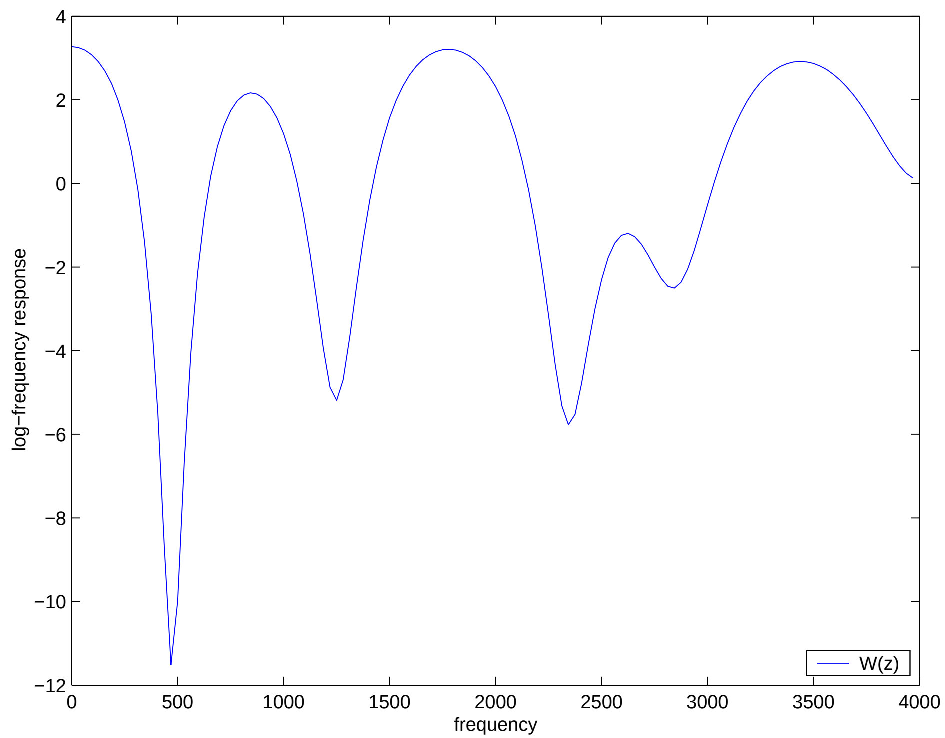


How it works:

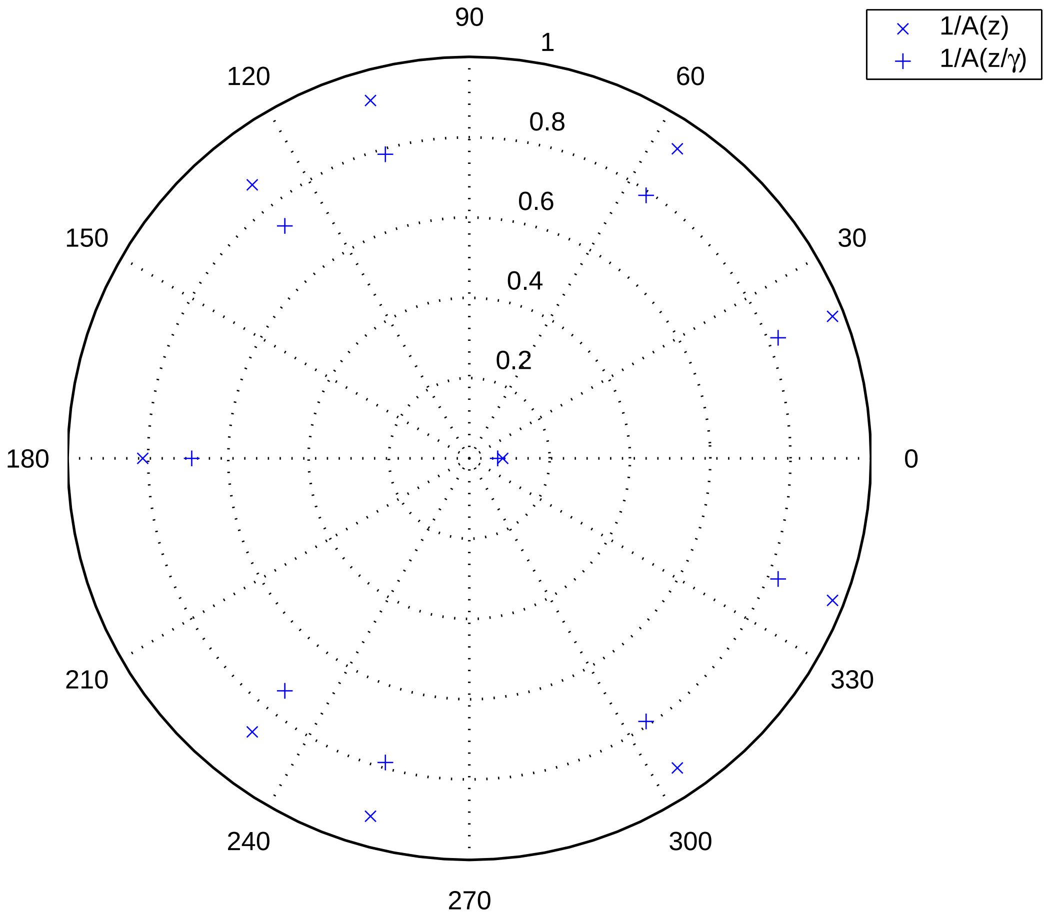
- $$W(z) = \frac{A(z)}{A(z/\gamma)}, \quad \text{where} \quad \gamma \in [0.8, 0.9] \tag{27}$$

- $A(z)$ – inverse LPC filter
- Filter $\frac{1}{A(z/\gamma)}$ is similar to $\frac{1}{A(z)}$ (speech spectral envelope), but the pole are less "peaky" due to the move of the poles to the center of the circle
- Multiplication leads to the perceptual filter.





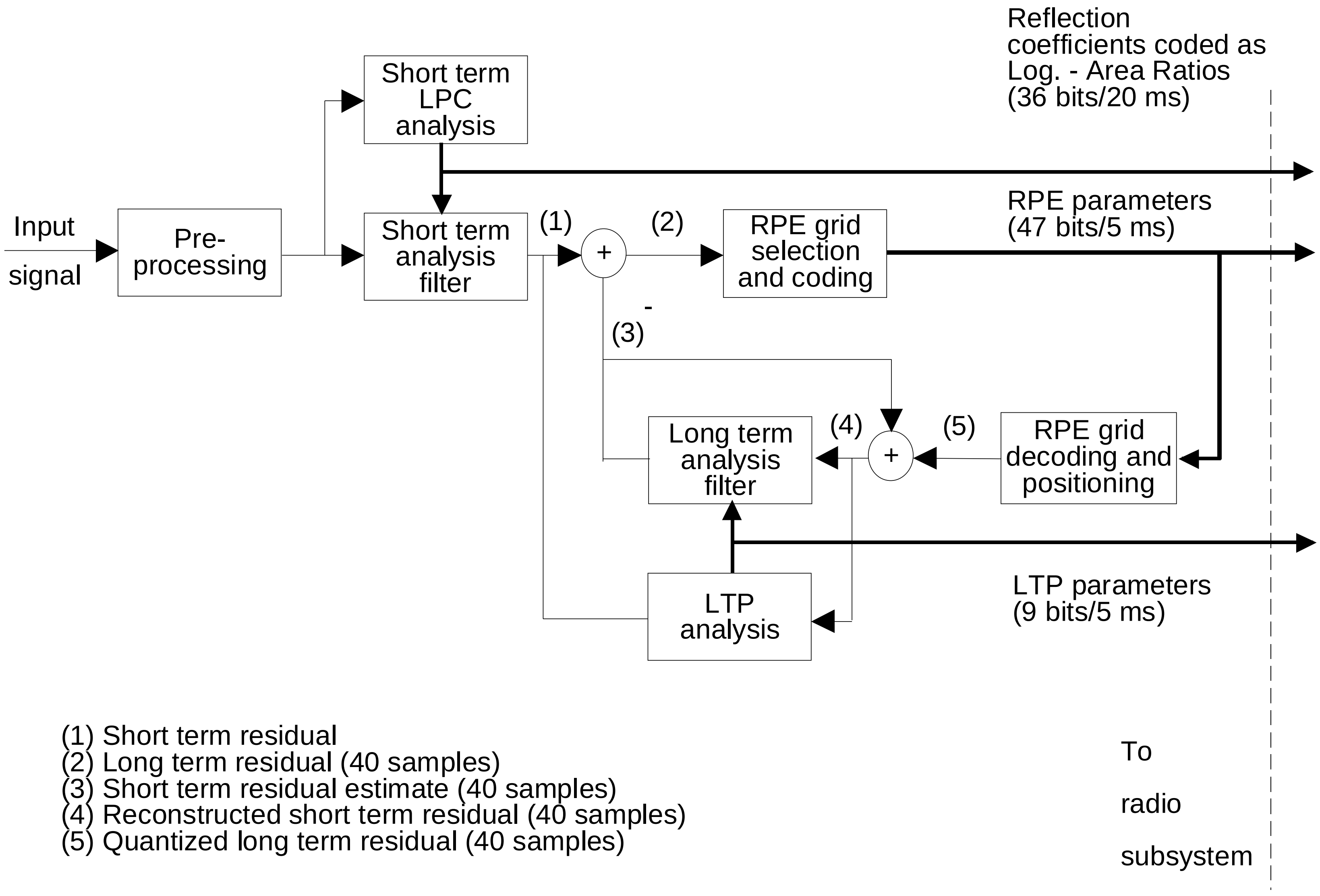
poles of the transmission function $\frac{1}{A(z)}$ and $\frac{1}{A(z/\gamma)}$:



Coding of an excitement signal in “short-frames”

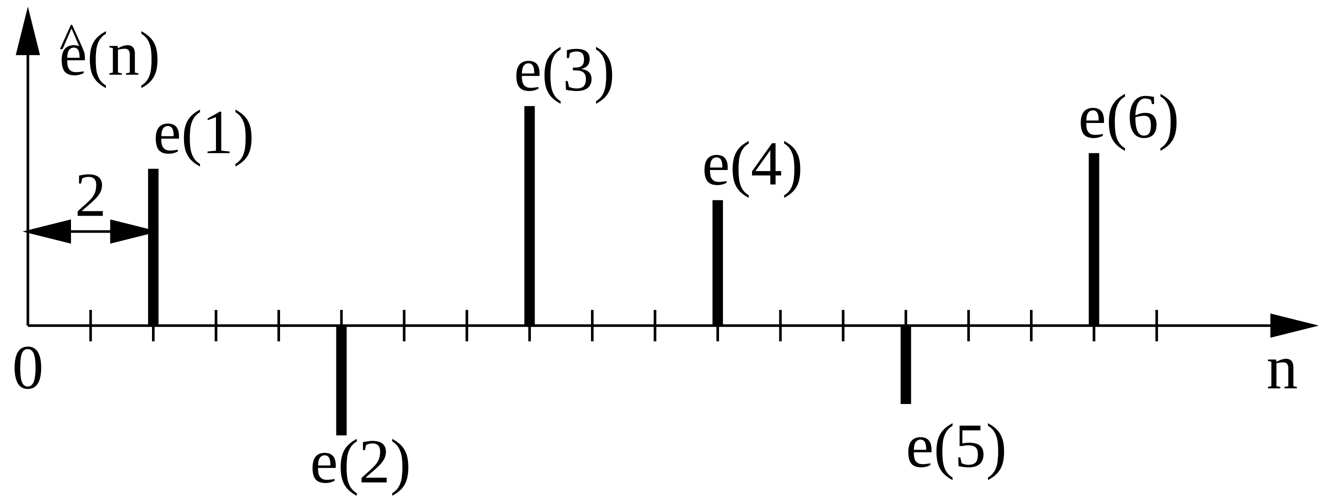
Usual length of the LPC frames is 20ms/160samples, excitement signal is encoded in shorter frames – around 40samples for 8kHz signal.

Coder 1 - 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 (20ms), coeficients of the LPC filter are transformed into 8 LAR.
- Long-term analysis (LTP) - (5 ms/40 samples) - lag a gain.
- Excitation signal is encoded in frames of 40 samples in a way that the $e(n)$ is down-sampled with factor 3 (14,13,13), and only the position of the first impulse is quantized (0,1,2,3(!))
- Sizes of the impulses are quantized using APCM.



- Results – segment of 260 bits $\times$ 50 $=$ 13 kbit/s.
- more in - standard 06.10, can be found in `http://pda.etsi.org`

Decoder RPE-LTP

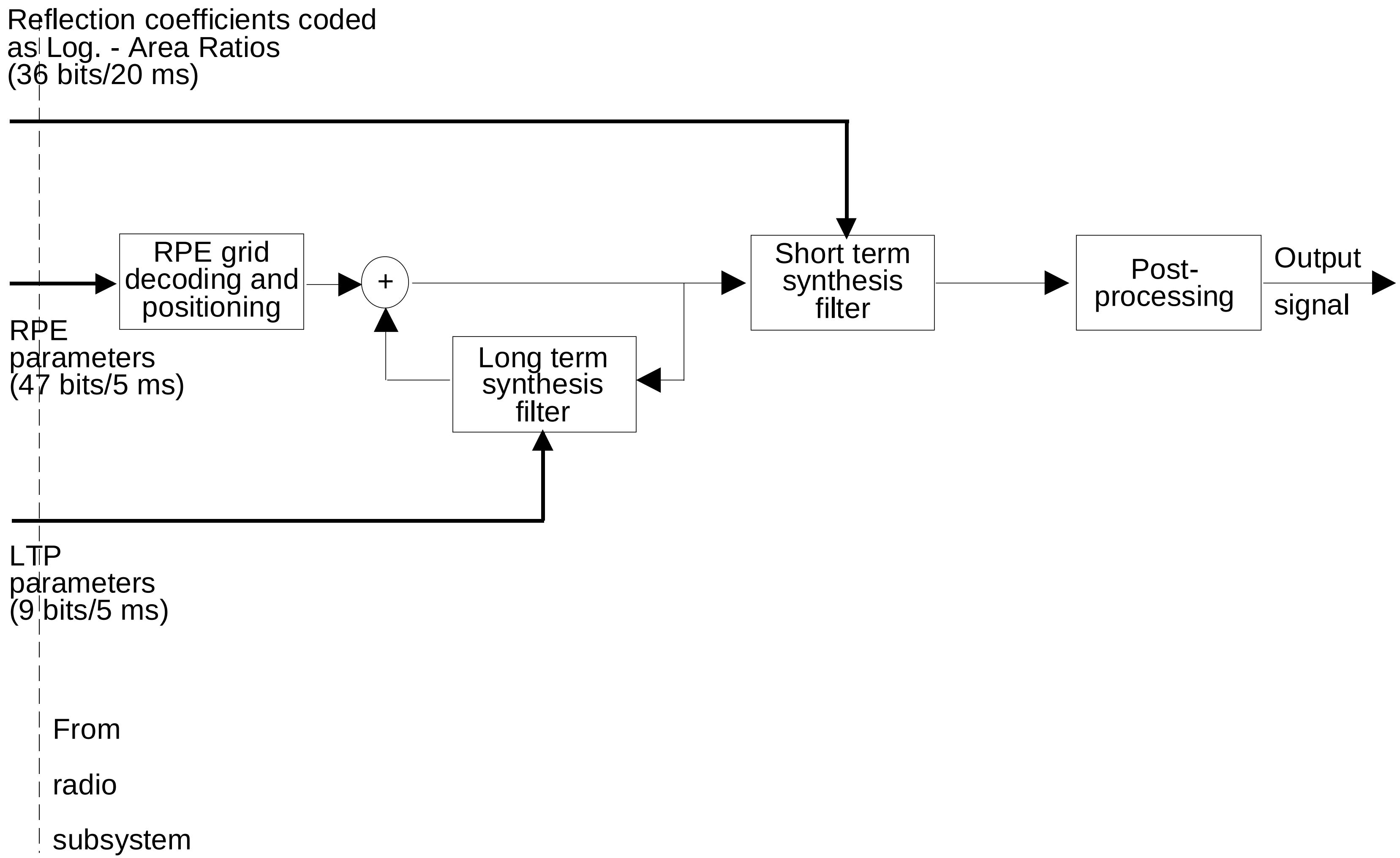
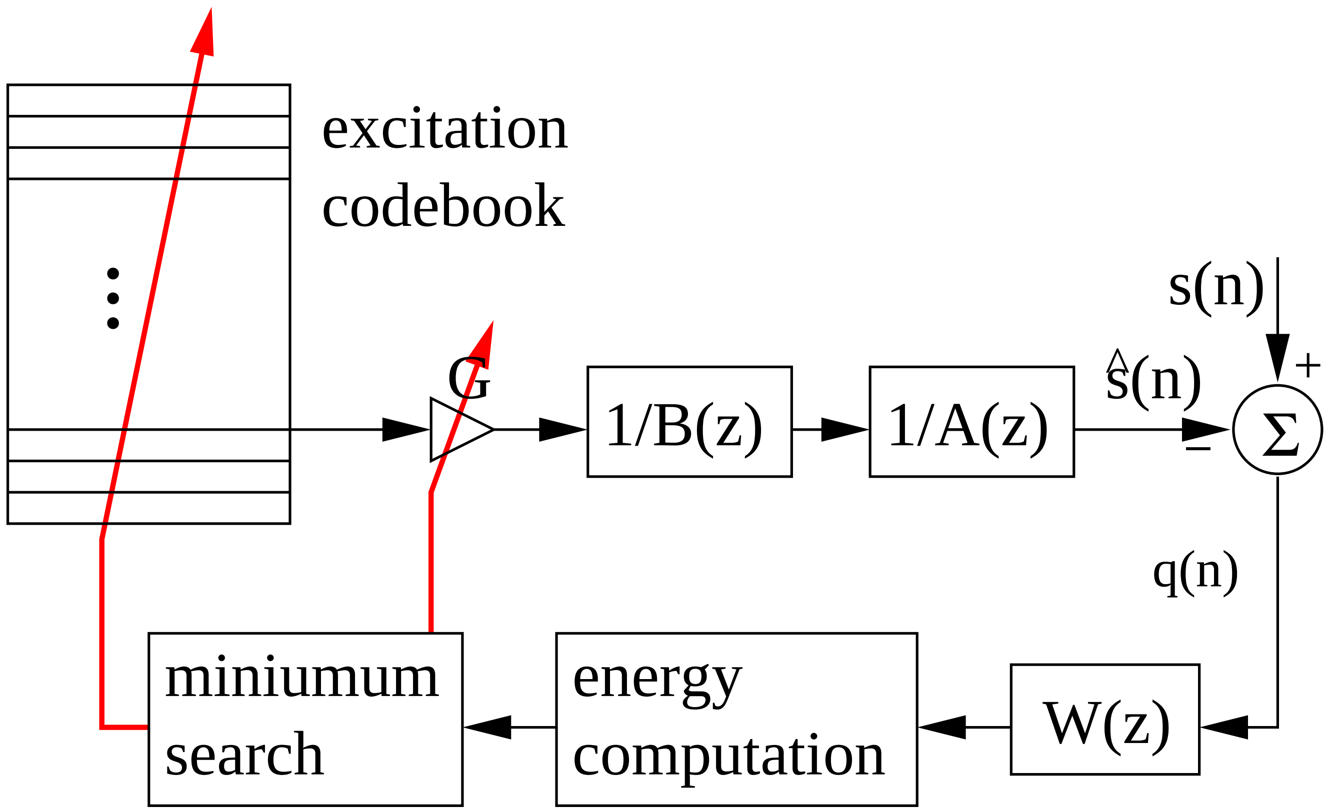


Figure 1.2: Simplified block diagram of the RPE - LTP decoder

CELP – Codebook-Excited Linear Prediction

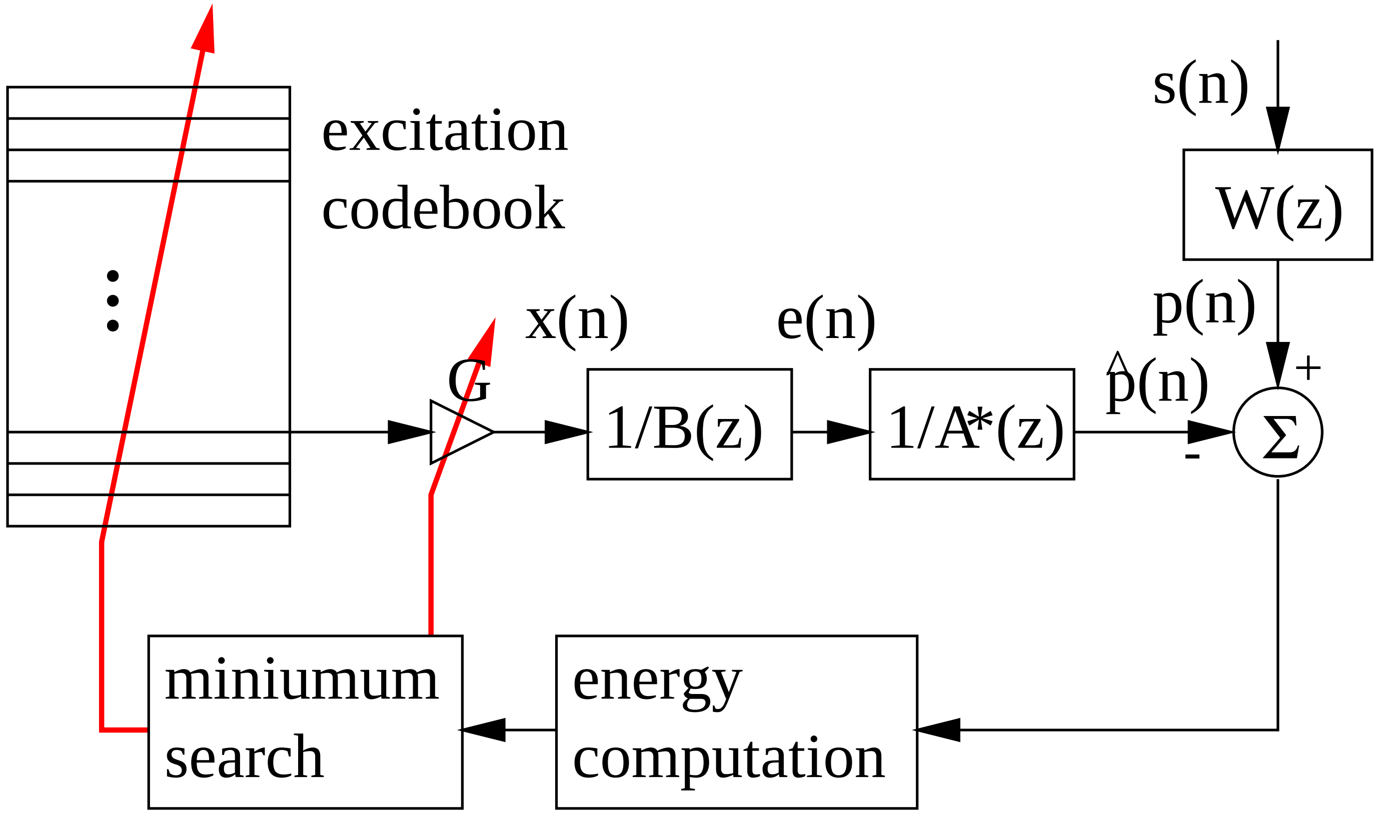
Excitement signal is encoded using a VQ - codebooks. Basic structure with a perceptual filter:



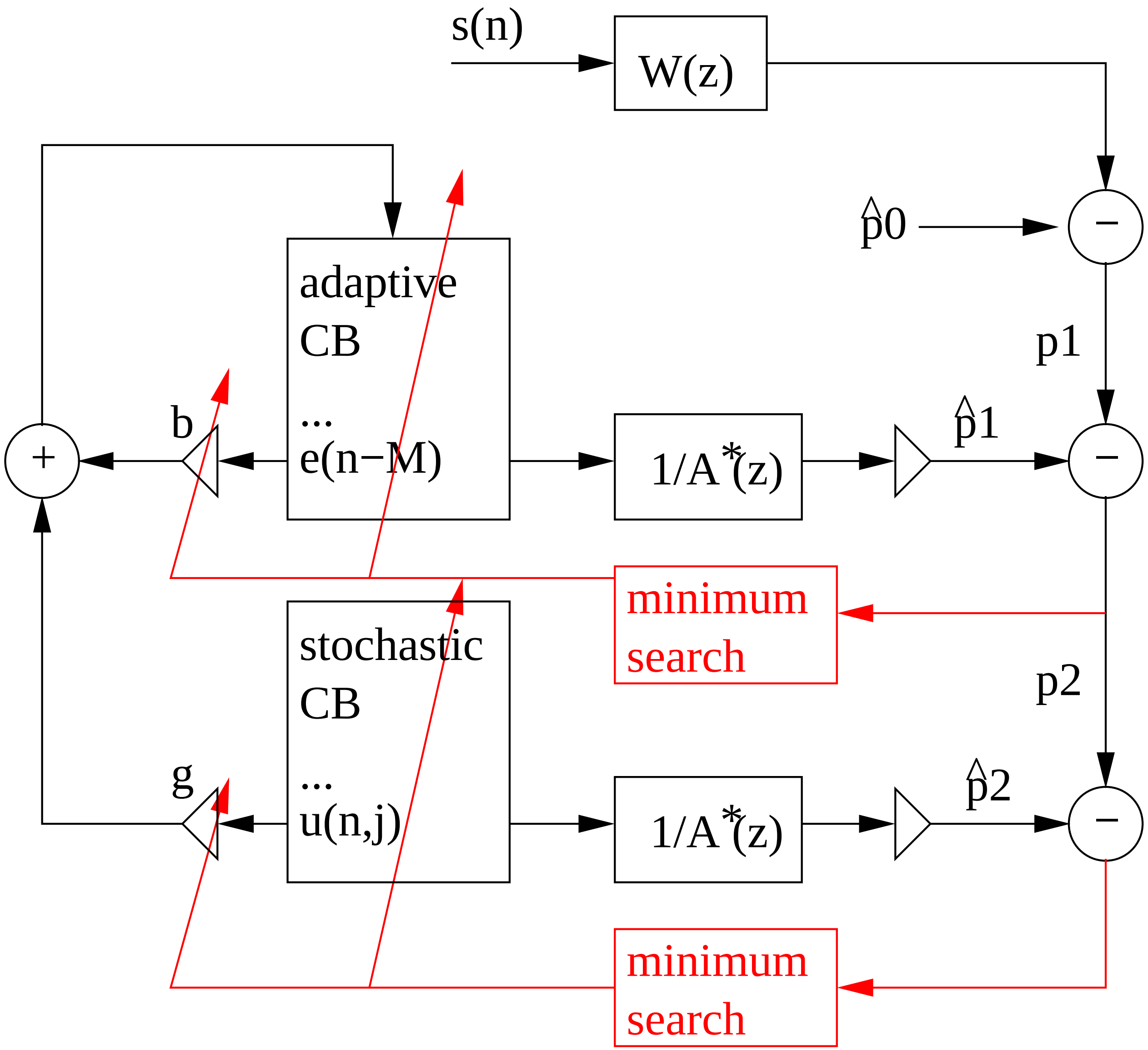
. . . each tested signal is filtered with $W(z) = \frac{A(z)}{A^\star(z)}$ — too expensive !

Perceptual filter at the input

Filtering is a linear operation $\Rightarrow W(z)$ can be moved into the both branches: into the input; and after $\frac{1}{B(z)}$–$\frac{1}{A(z)}$. We can simplify it: $\frac{1}{A(z)}\frac{A(z)}{A^\star(z)} = \frac{1}{A^\star(z)}$ — a new filter which will be used after $\frac{1}{B(z)}$.

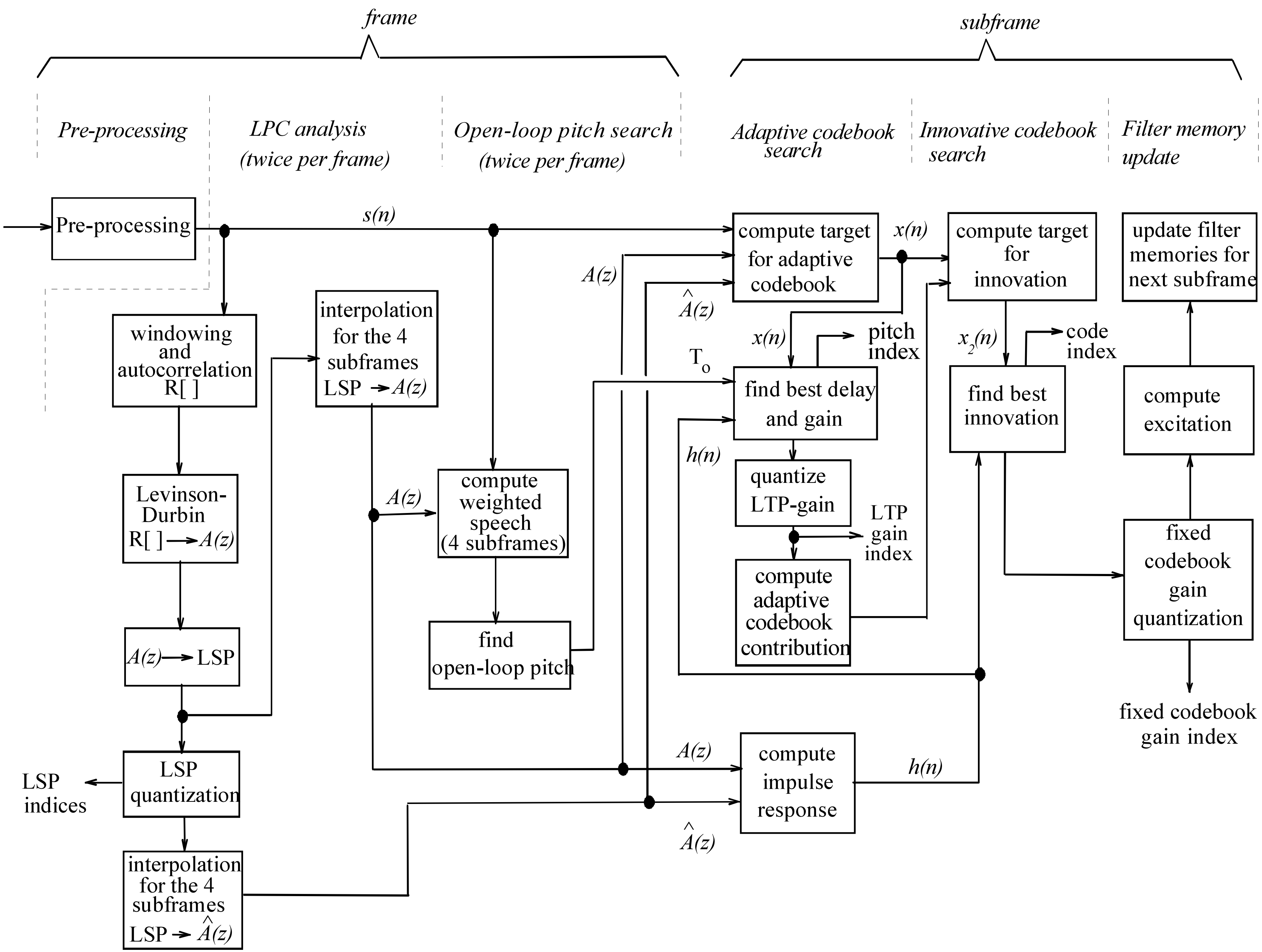


Final structure of CELP:

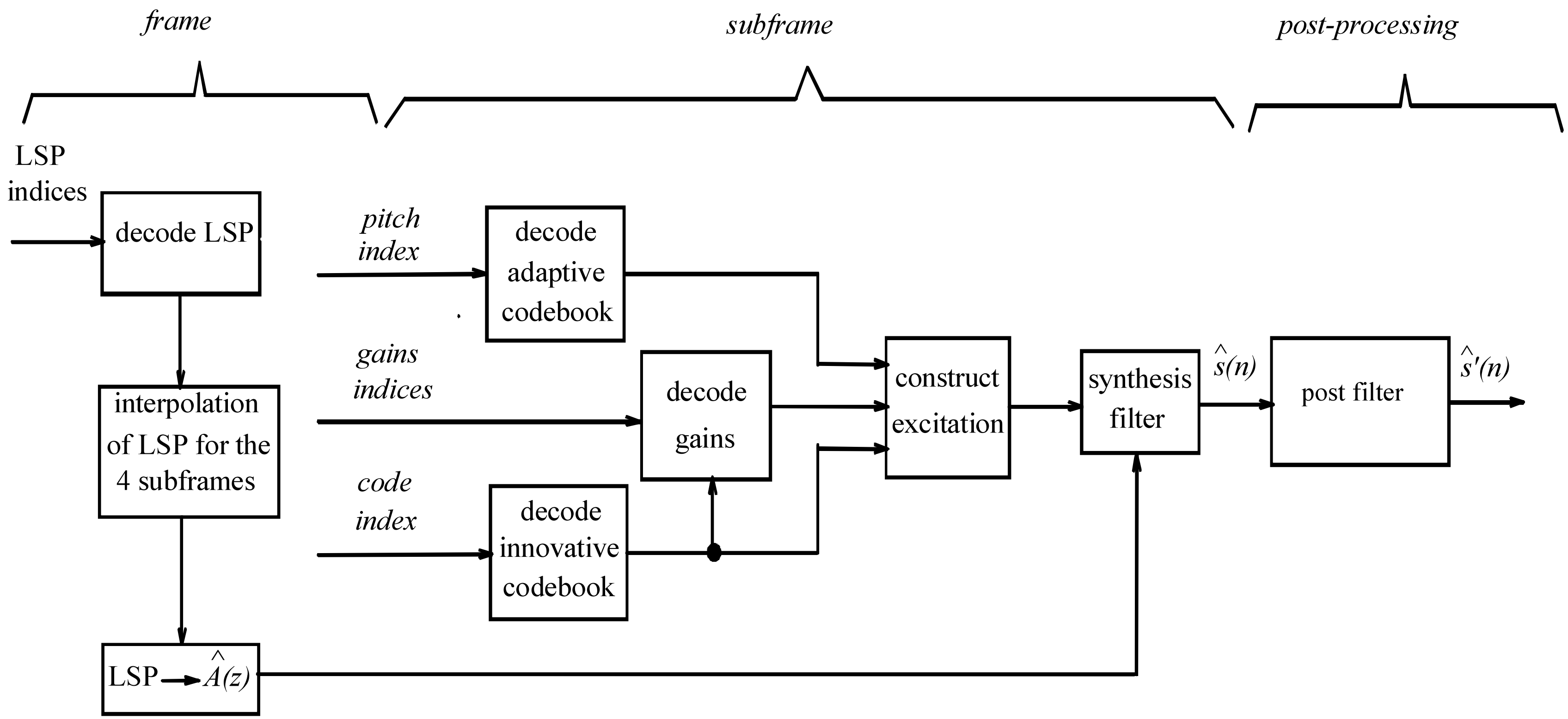


Example of CELP coder ACELP – GSM EFR

- classical CELP with an “inteligent codebook”.
- 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[ ]
interpolation for the 4 subframes LSP → A(z)
Levinson-Durbin R[ ] → A(z)
A(z) → LSP
LSP quantization
LSP indices
interpolation for the 4 subframes LSP → Â(z)
A(z)
compute weighted speech (4 subframes)
find open-loop pitch
A(z)
Â(z)
T o
compute target for adaptive codebook
x(n)
x(n)
pitch index
find best delay and gain
h(n)
quantize LTP-gain
LTP gain index
compute adaptive codebook contribution
A(z)
compute impulse response
h(n)
Â(z)
compute target for innovation
x 2 (n)
code index
find best innovation
update filter memories for next subframe
compute excitation
fixed codebook gain quantization
fixed codebook gain index

- segments of 20 ms (160 samples)
- Short-term prediction – 10 coeficients a_i in two sub-segments transformed into LSPs, from two sub-segments quantized together using split-matrix quantization (SMQ).
- 4 sub-segments of 40 samples (5 ms) for excitation.
- Estimate of lag, first for open-loop, then for closed-loop around the raw estimate, fractional pitch with resolution of $1/6$ sample.
- stochastic codebook: algebraic codebook - can contain only 10 non-zero impulses, which can only be $+1$ or -1 $\Rightarrow$ fast search (fast correlation - only addition, no multiplication), etc.
- 244 bits per segment $\times$ 50 $=$ 12.2 kbit/s.
- more in standard 06.60, can be found at `http://pda.etsi.org`
- decoder...


frame
subframe
post-processing
LSP indices
decode LSP
interpolation of LSP for the 4 subframes
LSP → Â(z)
pitch index
gains indices
code index
decode adaptive codebook
decode gains
decode innovative codebook
construct excitation
synthesis filter
ŝ(n)
post filter
ŝ'(n)

Additiona info

- Andreas Spanias (Arizona University):
`http://www.eas.asu.edu/~spanias`
in section Publications/Tutorial Papers: “Speech Coding: A Tutorial Review”, aa part was published in Proceedings of the IEEE, Oct. 1994.
- in section: Software/Tools/Demo - Matlab Speech Coding Simulations - a software for FS1015, FS1016, RPE-LTP a others.
- Standards of ETSI for cellular phones are free and available at:
`http://pda.etsi.org/pda/queryform.asp`
Possible keywords: “gsm half rate speech”. Many stanadrds are acompained with source code in C.