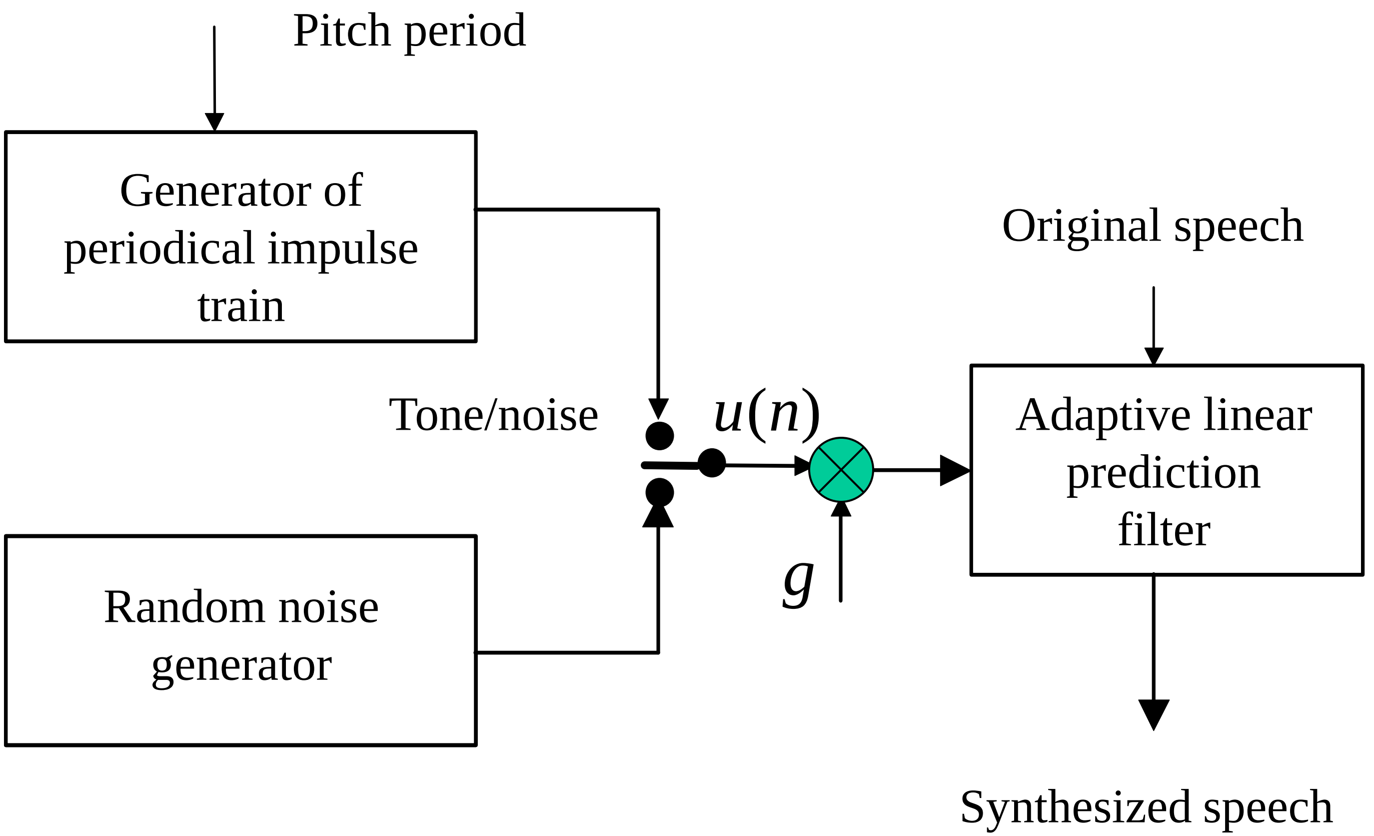
Linear prediction analysis-by-synthesis (LPAS) coders

The most popular class of speech coders for bit rates between 4.8 and 16 kb/s are model-based coders that use an LPAS method.

A linear prediction model of speech production (adaptive linear prediction filter) is excited by an appropriate excitation signal in order to model the signal over time. The parameters of both the filter and the excitation are estimated and updated at regular time intervals (frames). The compressed speech file contains these model parameters estimated for each frame.

Each sound corresponds to a set of filter coefficients. Rather often this filter is also represented by the poles of its frequency response called formant frequencies or formants.

Model of speech generation



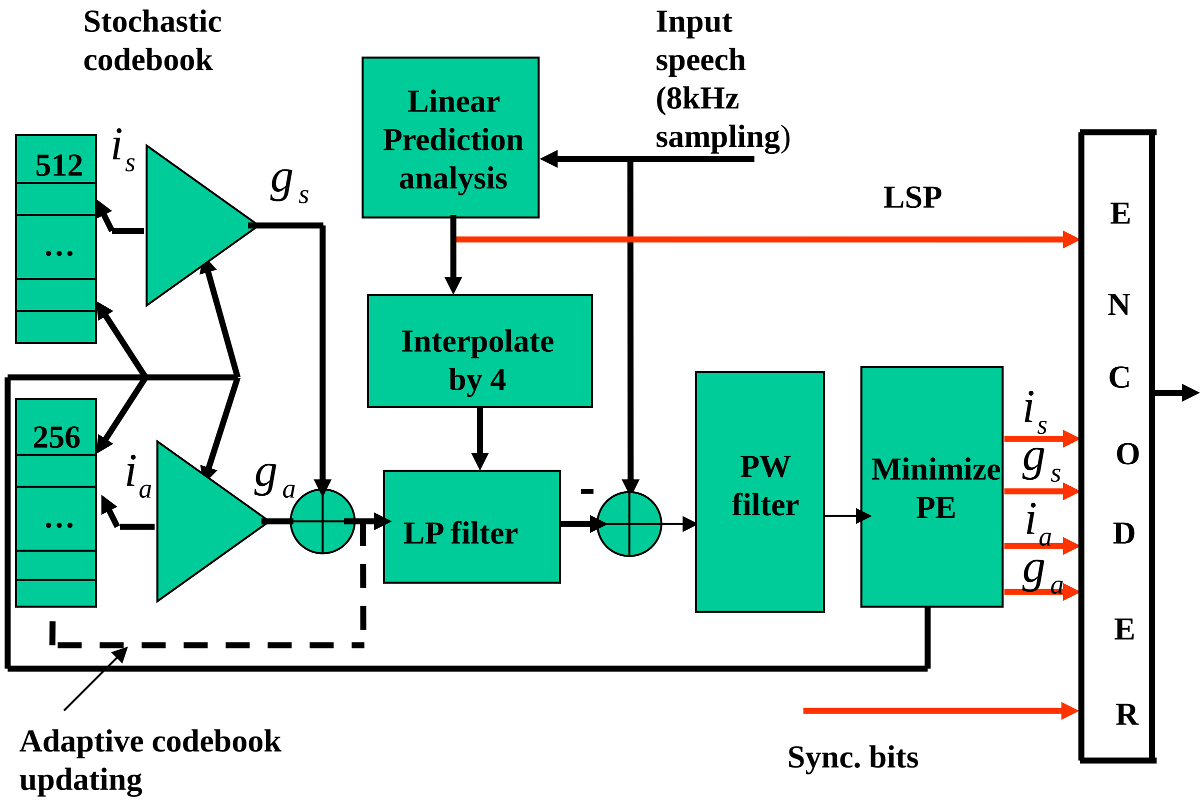
LPAS coders

The filter excitation depends on the type of the sound: voiced, unvoiced, vowel, hissing or nasal. The voiced sounds are generated by oscillation of vocal cords and represent a quasi-periodical impulse train. The unvoiced signals are generated by noise-like signals.

The period of vocal cords oscillation is called pitch period.

CELP coder is Code Excited Linear Predictive coder. It is a basis of all LPAS coders (G.729, G.723.1, G.728, IS-54, IS-96, RPE-LTP(GSM), FS-1016(CELP)).

CELP Standard



CELP Standard

Main ideas:

- A 10^{th} order LP filter is used to model the speech signal short term spectrum, or formant structure.

- Long-term signal periodicity or pitch is modeled by an adaptive codebook VQ.

- The residual from the short-term LP and pitch VQ is vector quantized using a fixed stochastic codebook.

- The optimal scaled excitation vectors from the adaptive and stochastic codebooks are selected by minimizing a time-varying, perceptually weighted distorted measure that improves subjective speech quality by exploiting masking properties of human hearing.

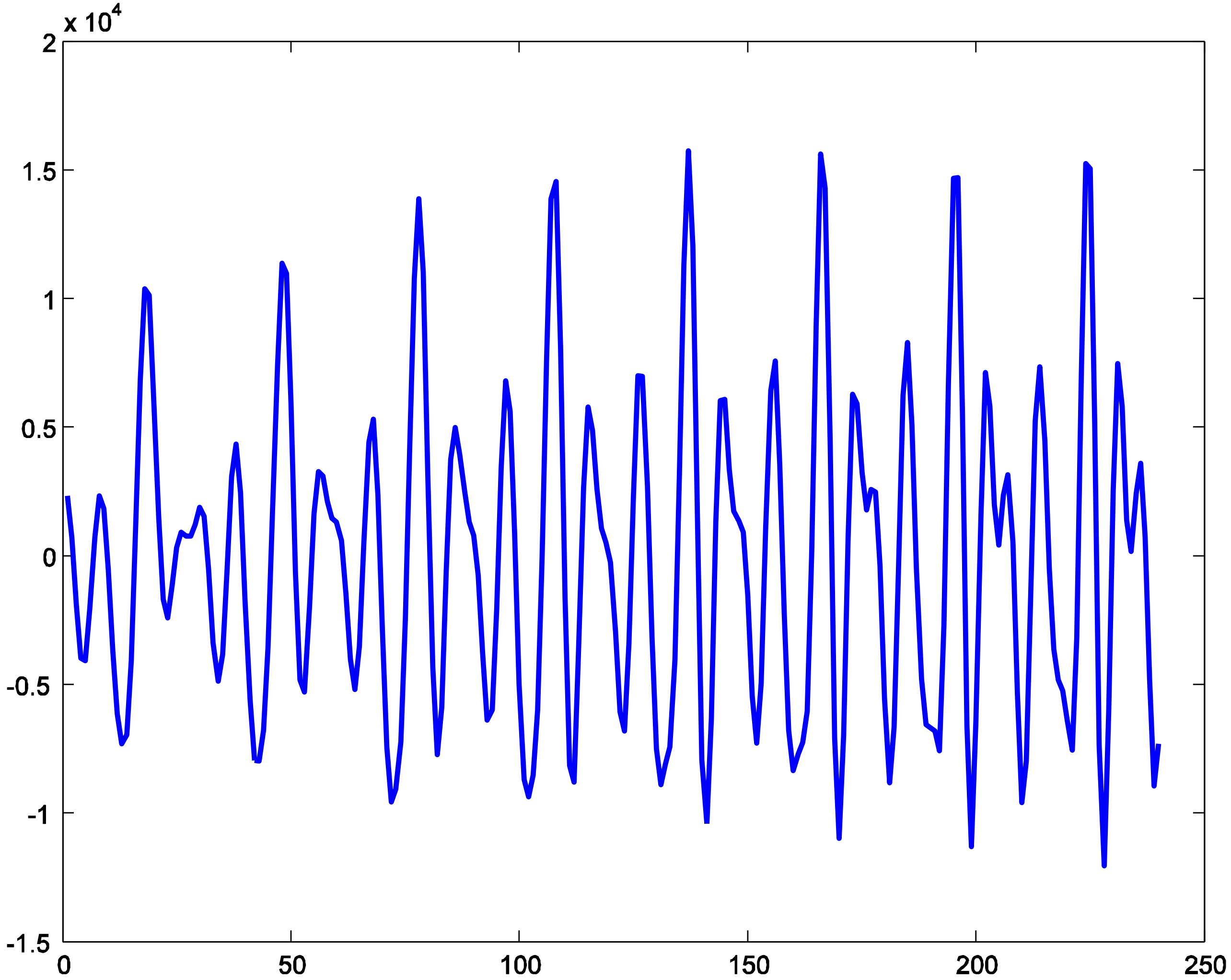
CELP Standard

CELP uses input signals at sampling rate 8 kHz and 30 ms (240 samples) frame size with 4 subframes of size 7.5 ms(60 samples.)

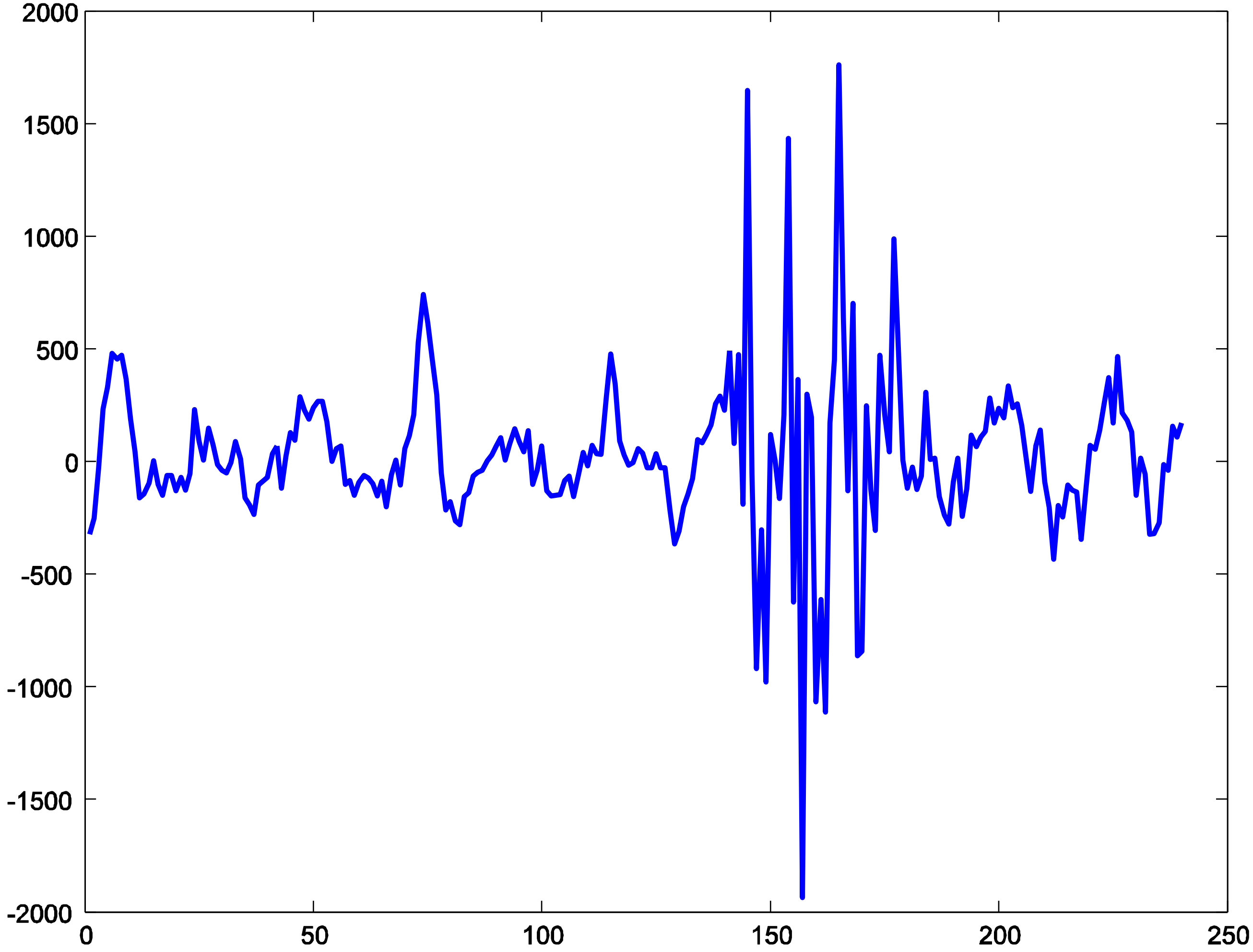
- Short-term prediction

It is performed once per frame by open-loop analysis. We construct 10^{th} order prediction filter using the autocorrelation method and the Levinson-Durbin procedure. The LP filter is represented by its linear spectral pairs (LPS) which are functions of the filter formants. The 10 LSPs are coded using 34-bit nonuniform scalar quantization. Because the LSPs are transmitted once per frame, but are needed for each subframe, they are linearly interpolated to form an intermediate set for each of the four subframes.

CELP



CELP



CELP

For the first frame we obtain :

Coefficients of the Yule-Walker equations (m=10) are:

1.0, 0.75, 0.17, -0.38, -0.65, -0.59, -0.35,-0.08, 0.17, 0.39, 0.52

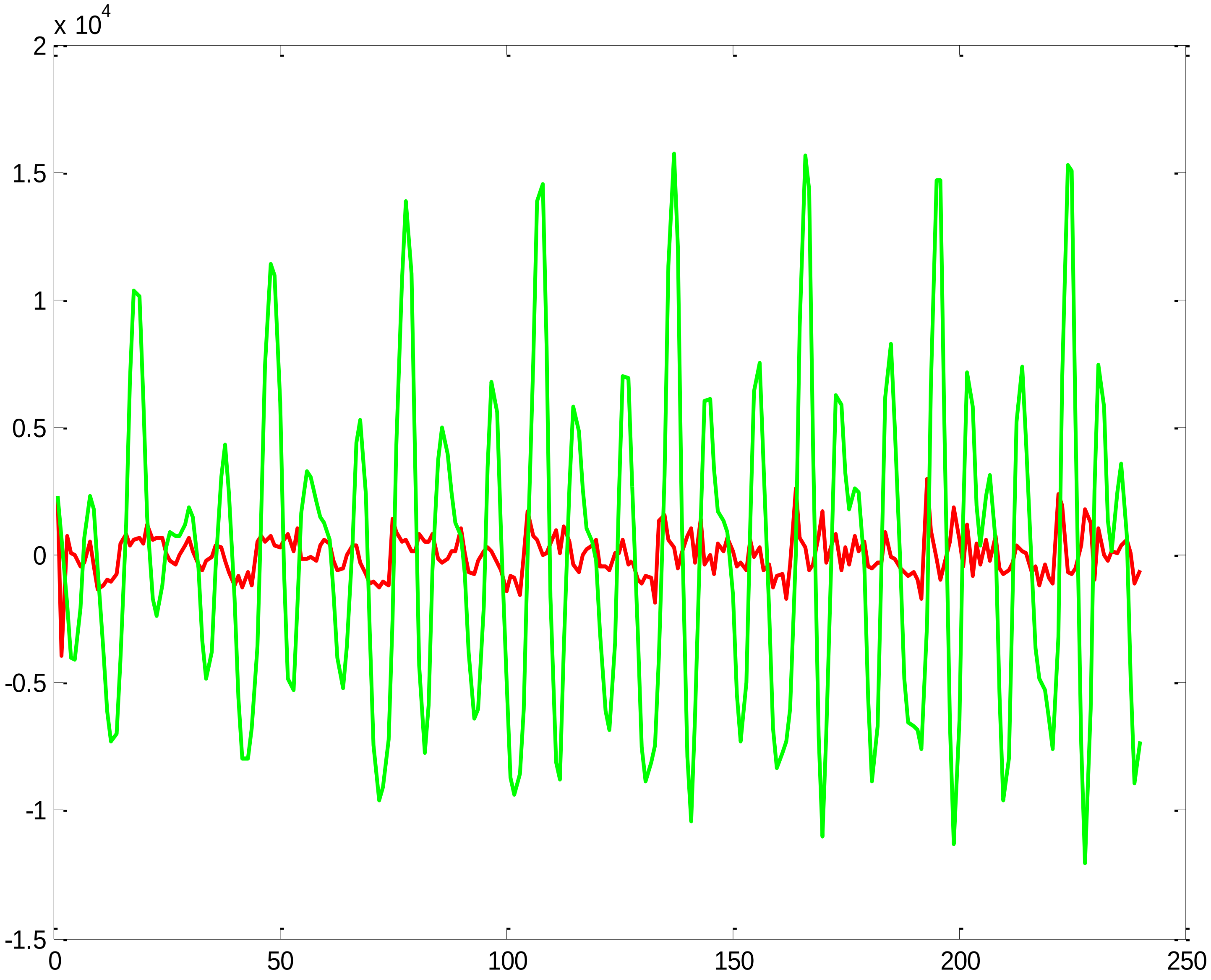
$R(0) = 3.79 \times 10^{7}$

The prediction filter coefficients are:

1.989, –1.734, 0.412, 0.096, 0.128, -0.084, -0.378, 0.303, 0.131, -0.166

The prediction error is : 1.07×10^{6}

CELP

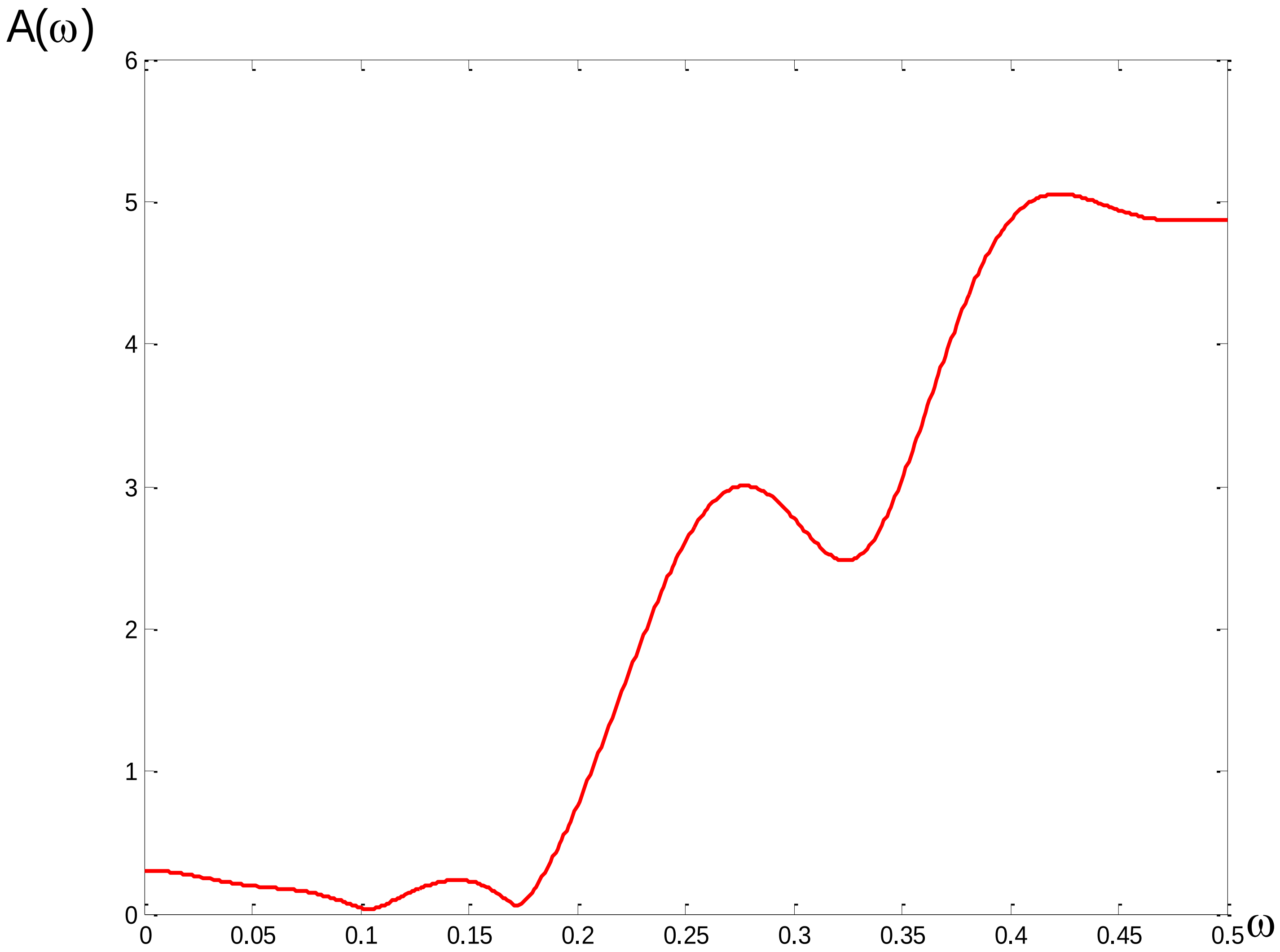


Linear spectral parameters

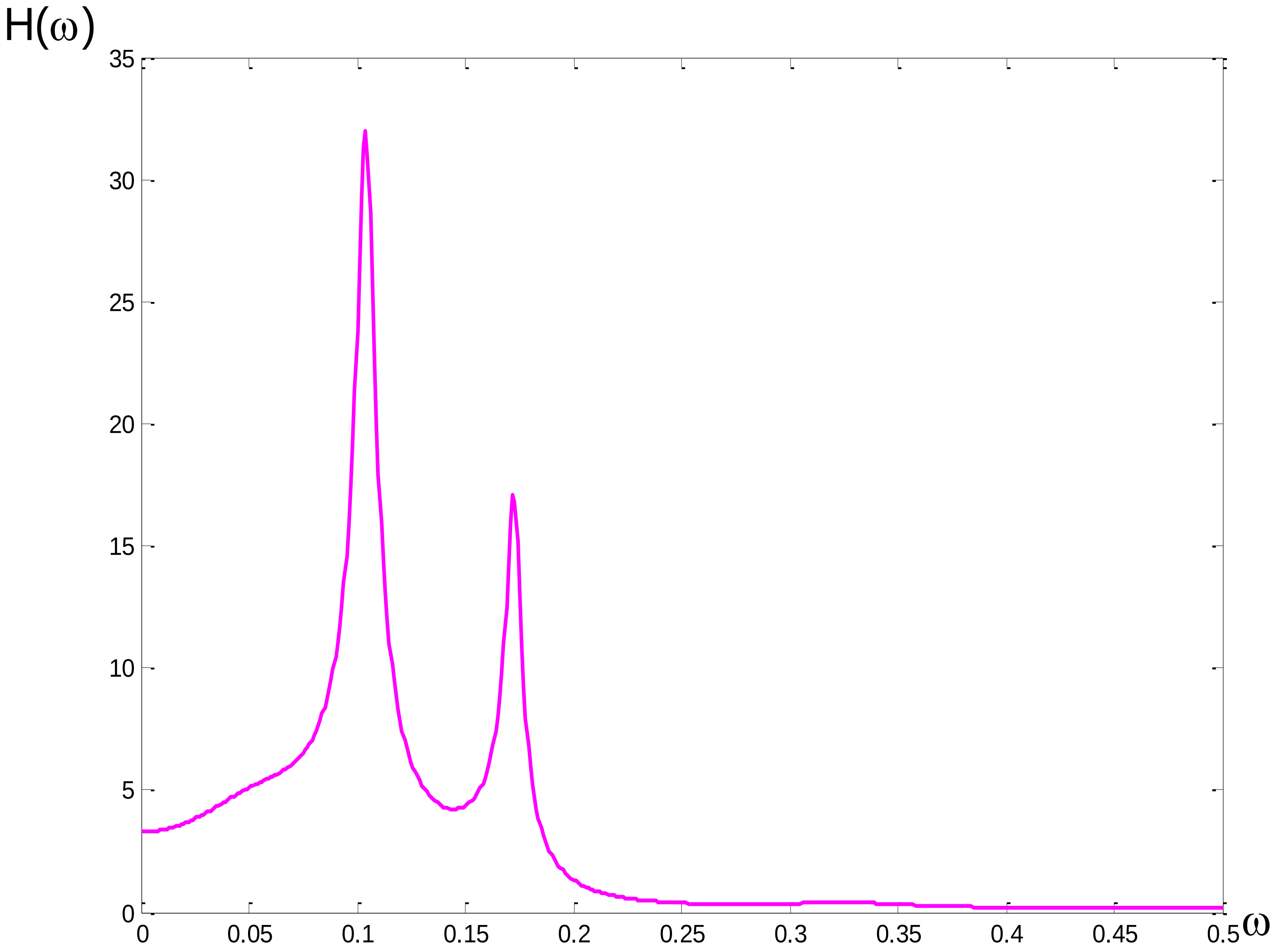
Let $H(z)$ be the transfer function of the inverse filter, that is $H(z) = 1/A(z)$, where $A(z)$ is the prediction filter. Then the frequencies $\omega_i, i = 1,2...,m$ which correspond to the poles of function $H(z)$ or zeros of $A(z)$, are called formant frequencies or formants.

LSPs are functions of formant frequencies. They can be found using the algorithm shown in Fig.10.1

Amplitude function of the prediction filter



Amplitude function of the inverse filter



LSP

Construct auxiliary polynomials $P(z), Q(z)$

↓

Reducing degree of $P(z), Q(z)$ **by 1**

↓

Construct polynomials $P^*(z), Q^*(z)$ **of degree m/2**

↓

Solve the equations $P^*(z) = 0, Q^*(z) = 0$

↓

Find $\omega_{pi} = \arccos(z_{pi}^{-1})$

$\omega_{qi} = \arccos(z_{qi}^{-1}), \quad i = 1, \ldots m/2$

$$m = 2$$

$$A(z) = A(2)z^{-2} + A(1)z^{-1} + 1$$

$$P(z) = 1 + (A(2) + A(1))z^{-1} + (A(2) + A(1))z^{-2} + z^{-3}$$

$$Q(z) = 1 + (A(1) - A(2))z^{-1} - (A(1) - A(2))z^{-2} - z^{-3}$$

$$PL(z) = P(z)/(z^{-1} + 1) = 1 + (A(1) + A(2) - 1)z^{-1} + z^{-2}$$

$$QL(z) = Q(z)/(z^{-1} - 1) = 1 + (A(1) - A(2) + 1)z^{-1} + z^{-2}$$

$$P^*(z) = (A(2) + A(1) - 1)/2 + z^{-1}$$

$$Q^*(z) = (A(1) - A(2) + 1)/2 + z^{-1}$$

$$z_p^{-1} = -(A(2) + A(1) - 1)/2$$

$$z_q^{-1} = -(A(1) - A(2) + 1)/2$$

$$\omega_p = \arccos(z_p^{-1})$$

$$\omega_q = \arccos(z_q^{-1})$$

LSP

LSP	Bits	Output Levels (Hz)
1	3	100,170,225,250,280,340,420,500
2	4	210,235,265,295,325,360,400,440, 480,520,560,610,670,740,810,880
3	4	420,460,500,540,585,640,705,775,850,950, 1050,1150,1250,1350,1450,1550
4	4	620,660,720,795,880,970,1080,1170, 1270,1370,1470,1570,1670,1770, 1870,1970
5	4	1000,1050,1130,1210,1285,1350,1430,1510,1590, 1670, 1750, 1850, 1950, 2050, 2150, 2250
6	3	1470, 1570, 1690, 1830, 2000, 2200, 2400, 2600
7	3	1800, 1880, 1960, 2100,2300, 2480, 2700,2900
10	3	3190, 3270, 3350, 3420, 3490, 3590, 3710, 3830

CELP

• Long-term signal periodicity is modeled by an adaptive codebook VQ. The adaptive codebook search is performed by closed-loop analysis using a modified minimum squared prediction error (MPSE) criterion of the perceptually weighted error signal.

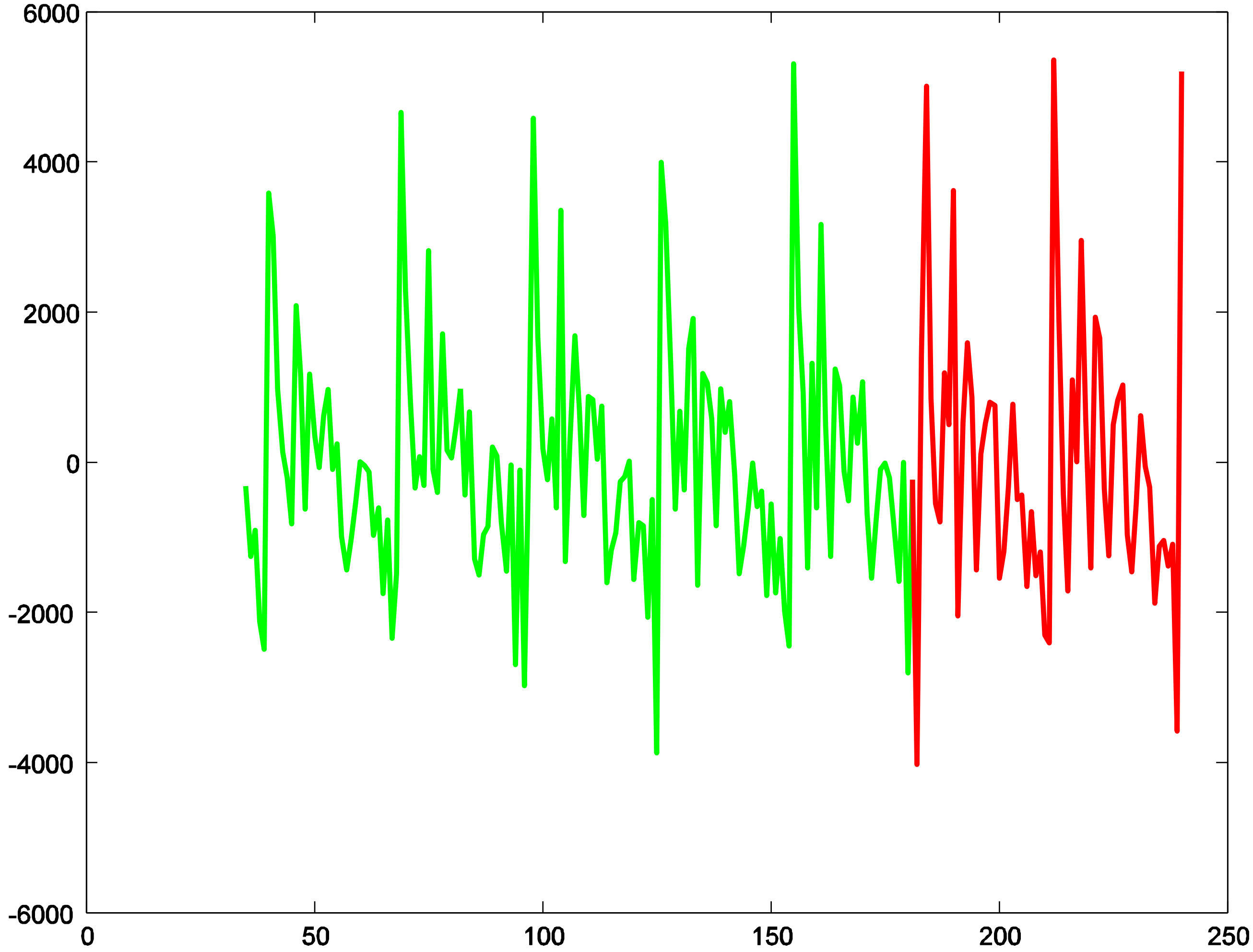
The adaptive codebook contains 256 codewords. Each codeword is constructed by the previous excitation signal of length 60 samples delayed by $20 \le M \le 147$ samples.

For delays less than the subframe length (60 samples) the codewords contain the initial M samples of the previous excitation vector. To complete the codeword to 60 elements, the short vector is replicated by periodic extension.

Adaptive codebook

Index	Delay	Adaptive CB sample numbers
255	147	-147, -146, -145,…, -89,-88
254	146	-146, -145, -144,…, -88,-87
253	145	-145, -144, -143,…, -87,-86
…	…	…
131	61	-61, -60, -59, …, -2, -1
…	…	…
1	21	-21, …,-2,-21,…,-2,-21,…, -2
0	20	-20,…, -1, -20,…-1, -20,…, -1

Adaptive codebook



Adaptive codebook

To find the best excitation in the adaptive codebook the 1st order linear prediction is used. Let $\mathbf{s}$ be the original speech vector and $\mathbf{a}_i$ be a filtered codeword $\mathbf{c}_i$ from the adaptive codebook then we search for

$$\min_i \|\mathbf{s} - g_a \mathbf{a}_i\|^2 = \min_i \left\{ \|\mathbf{s}\|^2 - 2g_a(\mathbf{s},\mathbf{a}_i) + g_a^2 \|\mathbf{a}_i\|^2 \right\}, \quad (10.1)$$

where g_a is a prediction coefficient or the adaptive codebook gain. By taking derivative with respect to g_a and setting it to zero we find the optimal gain:

$$g_a = \frac{(\mathbf{s},\mathbf{a}_i)}{\|\mathbf{a}_i\|^2} \quad (10.2)$$

Inserting (10.2) into (10.1) we obtain

$$\min_i \|\mathbf{s} - g_a \mathbf{a}_i\|^2 = \min_i \left\{ \|\mathbf{s}\|^2 - \frac{2(\mathbf{s},\mathbf{a}_i)^2}{\|\mathbf{a}_i\|^2} + \frac{(\mathbf{s},\mathbf{a}_i)^2}{\|\mathbf{a}_i\|^2} \right\} = \min_i \left\{ \|\mathbf{s}\|^2 - \frac{(\mathbf{s},\mathbf{a}_i)^2}{\|\mathbf{a}_i\|^2} \right\}$$

(10.3)

Adaptive codebook

Minimizing (10.3) over i is equivalent to maximizing the last term in (10.3) since the first term is independent of the codeword $\mathbf{a}_i$. Thus the adaptive codebook search procedure finds codeword $\mathbf{c}_i$ which maximizes the so-called match function m_i

$$m_i = \frac{(\mathbf{s}, \mathbf{a}_i)^2}{\|\mathbf{a}_i\|^2}$$

The adaptive codebook index i_a and gain g_a are transmitted four times per frame (every 7.5 ms). The gain is coded between –1 and +2 using nonuniform, scalar, 5 bit quantization.

Adaptive codebook

-.993	-.831	-.693	-.555	-.414	-.229	0.00	0.139
0.255	0.368	0.457	0.531	0.601	0.653	0.702	0.745
0.780	0.816	0.850	0.881	0.915	0.948	0.983	1.02
1.062	1.117	1.193	1.289	1.394	1.540	1.765	1.991

Gain encoding levels

Stochastic codebook

The stochastic codebook (SC) contains 512 codewords. The special form of SC represents ternary quantized

(-1,0,+1) Gaussian sequences of length 60.

The stochastic codebook search target is the original speech vector minus the filtered adaptive codebook excitation, that is, $\mathbf{u} = \mathbf{s} - g_a \mathbf{a}_{opt}$

The SC search is performed by closed-loop analysis using conventional MSPE criterion. We find such a codeword $\mathbf{x}_i$ which maximizes the following match function

$$\frac{(\mathbf{u}, \mathbf{y}_i)^2}{\|\mathbf{y}_i\|^2},$$

where $\mathbf{y}_i$ is the filtered codeword $\mathbf{x}_i$.

Stochastic codebook

The stochastic codebook index and gain are transmitted four times per frame. The gain (positive and negative) is coded using 5-bit nonuniform scalar quantization.

The weighted sum of the found optimal codeword from the adaptive codebook and the optimal codeword from the stochastic codebook are used to update the adaptive codebook. It means that 60 the most distant past samples are removed from the adaptive codebook, all codewords are shifted and the following new 60 samples are placed as the first

$$c_{i_a} g_a + x_{i_s} g_s$$

Stochastic codebook

-1330	-870	-660	-520	-418	-340	-278	-224
-178	-136	-98	-64	-35	-13	-3	-1
1	3	13	35	64	98	136	178
224	278	340	418	520	660	870	1330

Stochastic codebook gain encoding levels

CELP

The total number of bits per frame can be computed as

$$4(b_{g_s} + b_{i_s} + b_{g_a} + b_{i_a}) + b_{LSP} = 4(5+8+5+9)+34 = 142,$$

where b_{g_s}, b_{i_s} and b_{g_a}, b_{i_a} are numbers of bits for index and gain of the stochastic and adaptive codebook, respectively, b_{LSP} is the number of bits for linear spectral pairs.

Taking into account that a frame duration is 30 ms we obtain that the bit rate of the CELP coder is equal to

$$\boxed{R = \frac{142}{30 \cdot 10^{-3}} \approx 4733} \quad \text{b/s.}$$

Adding bits for synchronization and correcting errors we get that bit rate is equal to 4800 b/s.

CELP

The critical point of the standard is computational complexity of two codebook searches.

The number of multiplications for adaptive codebook search can be estimated as

$$(60 \times 10 + 60) \times 256 \approx 1.7 \times 10^5.$$

The stochastic codebook search requires

$$(60 \times 10 + 60) \times 512 \approx 3.4 \times 10^5$$

multiplications.