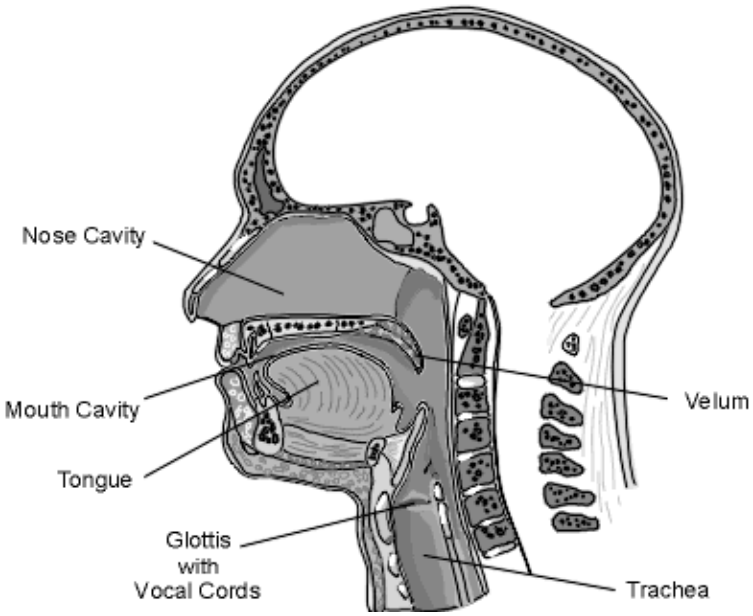
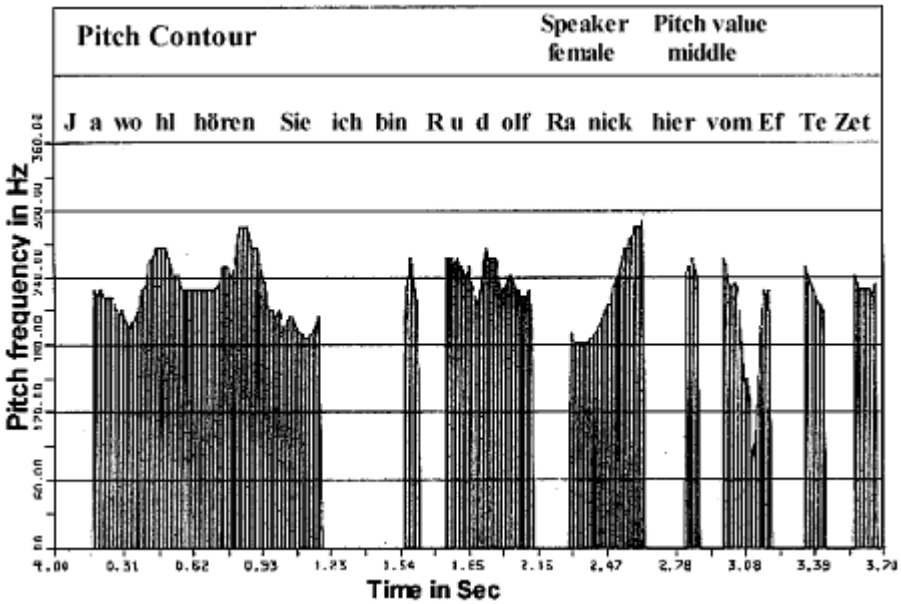
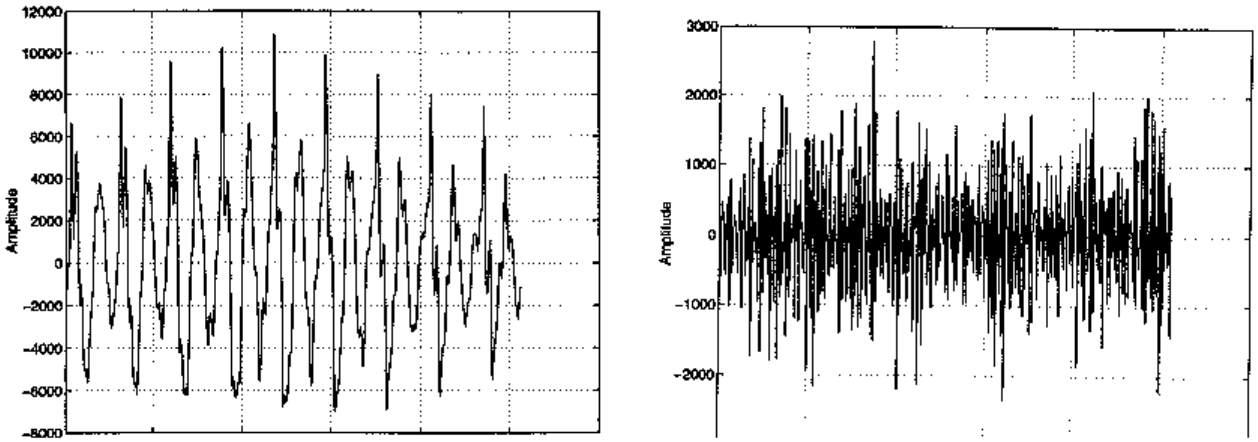
9. SPEECH PRODUCTION & LPC MODELS

Speech is produced by the interaction of the vocal tract with excitation of vocal chords in the glottis.

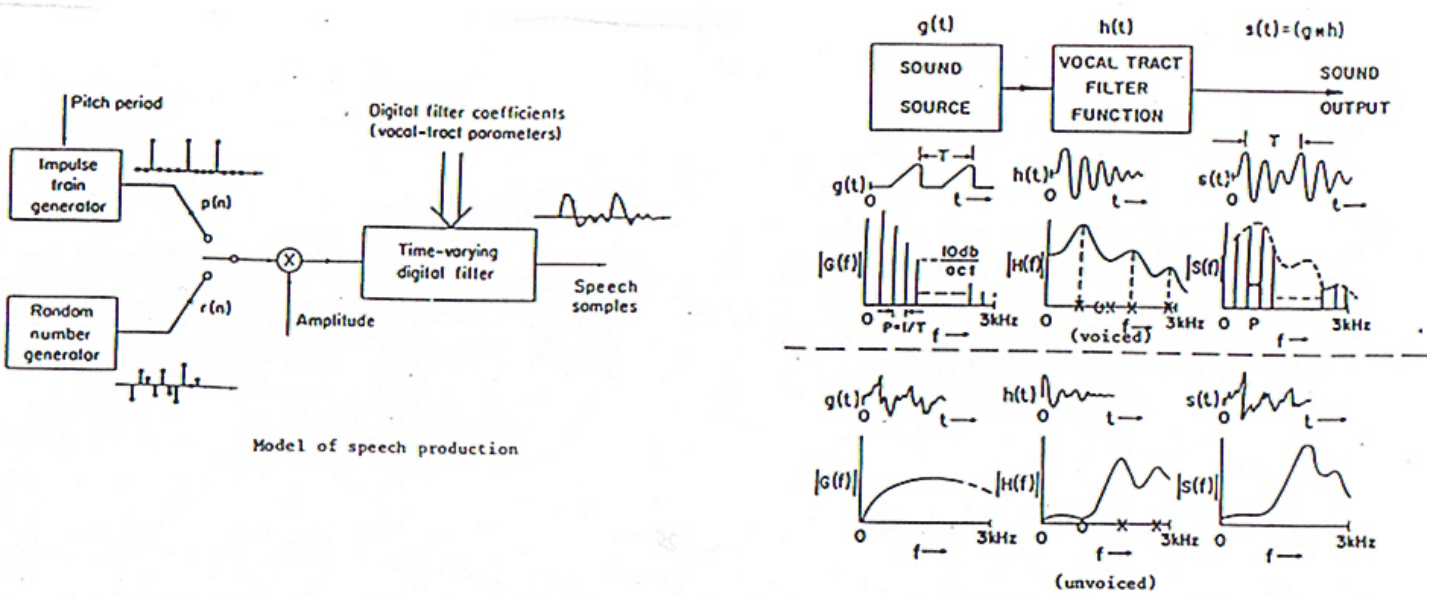




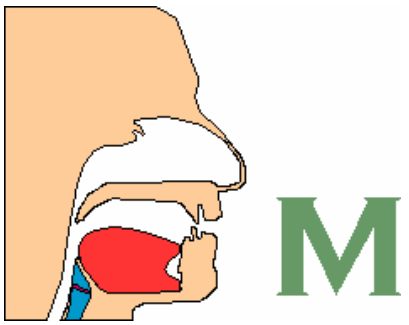
Typical impulse sequence coming from vocal chords and the pitch contour.



Two frames of 512 samples each voiced and unvoiced speech.

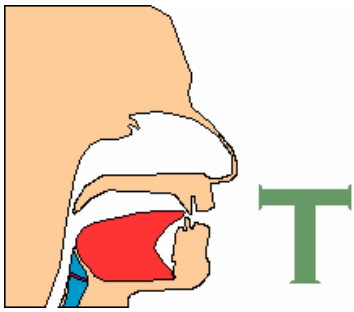


Sounds: "m" and "t"



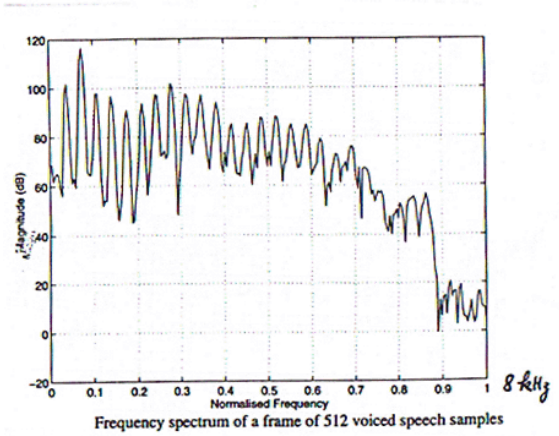
M

[Sound: m.wav]

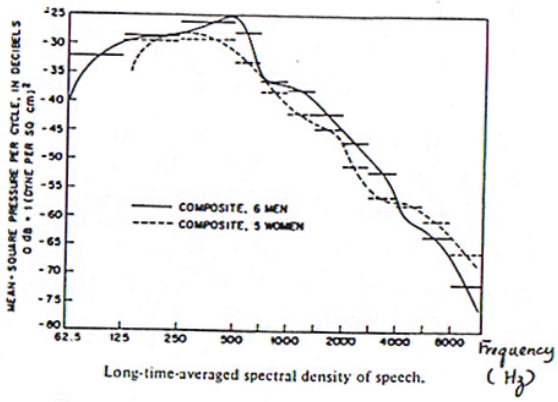


T

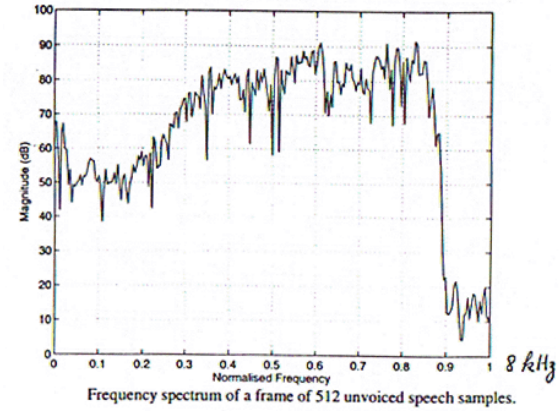
[Sound: t.wav]



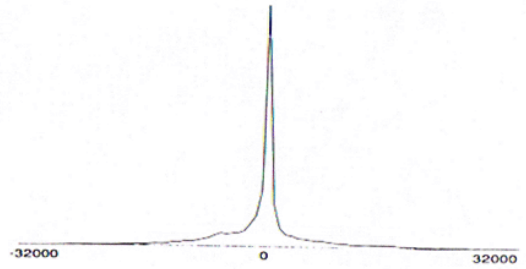
Frequency spectrum of a frame of 512 voiced speech samples



Long-time-averaged spectral density of speech.

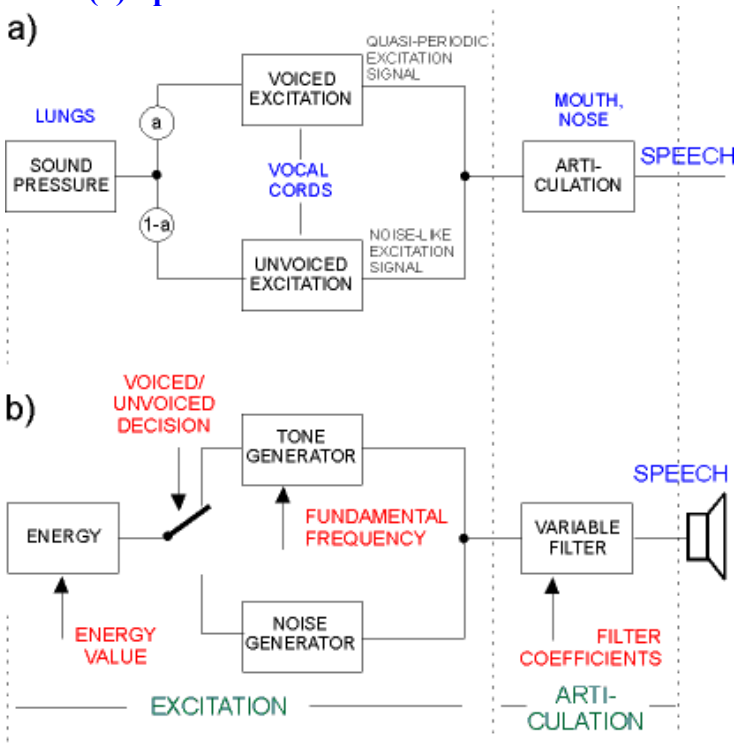


Frequency spectrum of a frame of 512 unvoiced speech samples.



A typical density function for speech. This function is the histogram of 3 seconds of speech.

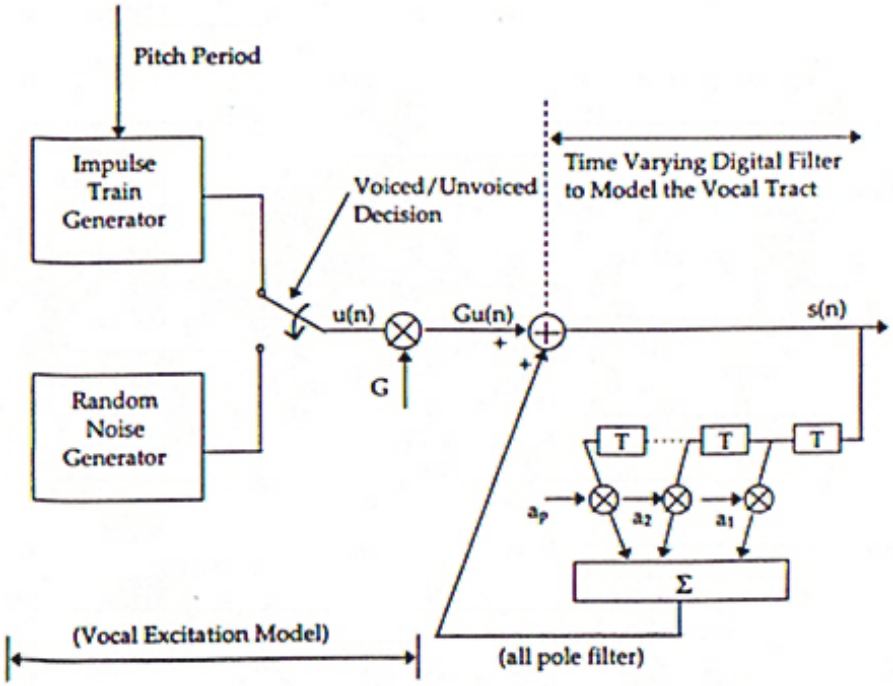
Human (a) and Synthetic (b) Speech Production:



Linear Predictive Coding (LPC) of Speech

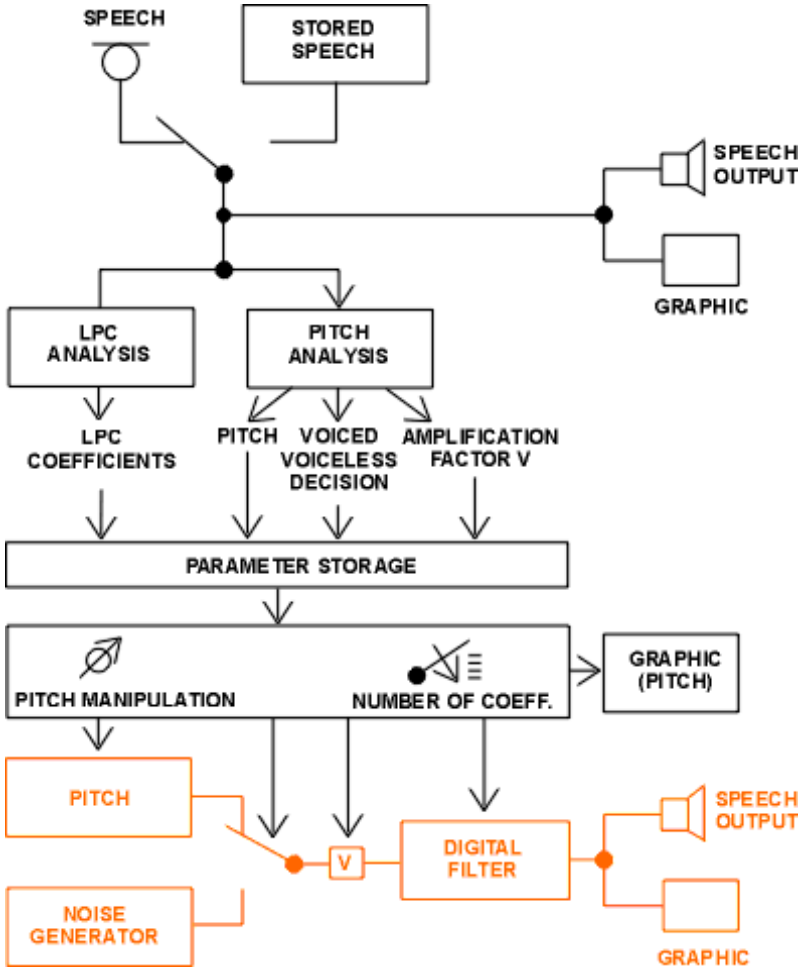
- Introduced in mid 1960's by Itakura in Japan and Atal & Schroeder in the US is the predominant technique for estimating and then representing speech in terms of:
  1. Pitch information
  2. Gain
  3. Short-term spectra
  4. Vocal tract parameters.
- Most recent sample is predicted by a linear combination of past samples.
- Analysis done by Finite-length Impulse Response (FIR) digital filter.
- Filter coefficients (LPC prediction parameters) are determined from the minimization of the sum of the squared difference between the actual speech samples and the linearly predicted ones.

LPC Model:



- Speech can be modeled from the above setup as the output of a linear, time-varying system excited by either quasi-periodic glottal pulses for voiced speech or random noise for unvoiced sounds.
- LPC method provides a robust, reliable, and accurate method for estimating the parameters that characterize the linear time-varying system.

Overall LPC Block Diagram: Analyzer (Top) and Synthesizer (Bottom)



Digital Filter to be computed and speech is generated by:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{P} a_k . z^{-k}} \tag{1}$$

$$s(n) = G.u(n) + \sum_{k=1}^{P} a_k . s(n-k) \tag{2}$$

where the first term is excitation and the latter one is the prediction and

G: Gain for each segment (20-30 ms) of speech

P: Predictor (filter) order

a_k : LPC (filter) coefficients

Problem: How do we determine the optimum set of $\{\alpha_k\}$ given a speech segment $s(n)$?

Define two terms:

$$s_P(n) \equiv \sum_{k=1}^{P} \alpha_k . s(n-k) \tag{3}$$

$$e(n) \equiv s(n) - s_P(n) = s(n) - \sum_{k=1}^{P} \alpha_k . s(n-k) \tag{4}$$

If we have: $\alpha_k = a_k$ then we get:

$$e(b) = G.u(n)$$

That is the residual (error) is equal to Gain times the excitation, glottal pulses for voiced speech and random pulses for unvoiced. The energy of $e(n)$:

$$\begin{aligned} E_e &= E\{[s(n) - \sum_{k=1}^{P} \alpha_k . s(n-k)]^2\} \\ &= E\{s(n)^2 - 2.\sum_{k=1}^{P} \alpha_k . E\{s(n)s(n-k)\} + \sum_{k=1}^{P}\sum_{j=1}^{P} \alpha_j . \alpha_k . E\{s(n-j)s(n-k)\} \\ &= E_s - 2.A^T.G + A^T.R.A \end{aligned} \tag{5}$$

where

$$A \equiv [\alpha_1, \alpha_2, \cdots, \alpha_P]^T \tag{6}$$

$$G \equiv [\psi_1, \psi_2, \cdots, \psi_P]^T \tag{7}$$

$$R = \begin{bmatrix} \psi_0 & \psi_1 & \psi_2 & \cdots & \psi_{P-1} \\ \psi_{-1} & \psi_0 & \psi_1 & \cdots & \\ \vdots & \vdots & \psi_0 & \ddots & \psi_1 \\ \psi_{-(P-1)} & \cdots & \cdots & \psi_{-1} & \psi_0 \end{bmatrix} \tag{8}$$

$$\psi_i = E\{s(n).S(n+i)\}$$

Minimize E_e: Take partial derivative with respect to each α_k and set them to zero:

$$\frac{\partial E_e}{\partial \alpha_k} = 0 \quad for\ k = 1, \cdots, P$$

which results in optimum parameter set:

$$A_{opt} = R^{-1}.G \tag{9}$$

$$A_{opt} = \begin{bmatrix} \psi_0 & \psi_1 & \psi_2 & \cdots & \psi_{P-1} \\ \psi_{-1} & \psi_0 & \psi_1 & \cdots & \\ \vdots & \vdots & \psi_0 & \ddots & \psi_1 \\ \psi_{-(P-1)} & \cdots & \cdots & \psi_{-1} & \psi_0 \end{bmatrix}^{-1} . \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_P \end{bmatrix} \tag{9a}$$

Methodologies for LPC coefficient computation:

Even though, there are basically four methods to calculate ψ_i only two are often used:

1. Autocorrelation Method
2. Covariance Method

Autocorrelation Method:

The correlation function is normally measured over a 20-30 ms segment of speech during which the characteristics assumed to be stationary. Segments could be overlapped or non-overlapped.

$$\psi_i = \sum_{r=1}^{w-i} s(r).s(r+i) \quad for\ i = 0,1,\cdots,P \tag{10}$$

where w: Number of speech samples in the current analysis window (segment).

Note: Upper limit of the sum varies as i var*ies* $0 \Rightarrow P$.

For an analysis window size of 256 samples, the correlation values are found from:

$x_0\ x_1\ x_2\ \ldots\ x_{255}\ x_{256}\ x_{257}\ \ldots\ x_{511}\ x_{512}\ x_{513}\ \ldots$

$0\ \ s(1)\ \ s(2)\ \ \ s(256)\ \ 0\ \ldots$

$$\psi_1 = \sum_{r=1}^{256-1} s(r)\, s(r+1)$$

$$\psi_0 = s(1)^2 + s(2)^2 + \ldots + s(256)^2 \quad (256 \text{ terms})$$

$$= s(1)s(2) + s(2)s(3) + \ldots + s(255)s(256) \quad (255 \text{ terms})$$

$$\psi_2 = s(1)s(3) + s(2)s(4) + \ldots + s(254)s(256) \quad (254 \text{ terms})$$

$$\psi_{10} = s(1)s(11) + \ldots + s(246)s(256)$$

$$\psi_{-1} = \sum_{r=1}^{256-(-1)} s(r)s(r-1) = \sum_{r=1}^{257} s(r)s(r-1)$$

$$= s(1)s(0) + s(2)s(1) + s(3)s(2) + \ldots$$

($s(0)$: 0)

$$+ s(256)s(255) + s(257)s(256)$$

($s(257)$: 0)

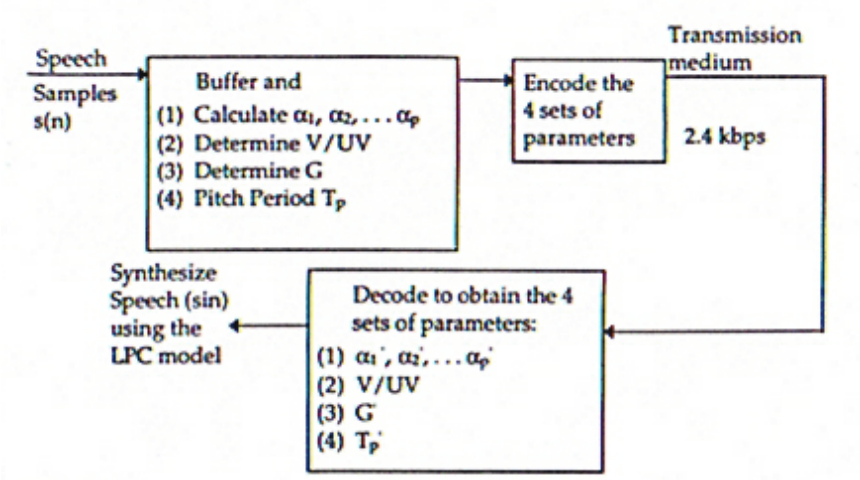
$$= s(2)s(1) + s(3)s(2) + \ldots + s(256)s(255) \quad (255 \text{ terms})$$

$$= \psi_1$$

Note: The correlation function exhibits even symmetry:

$$\psi_i = \psi_{-i} \tag{11}$$

With ψ_i's known and using (9) we can calculate the prediction coefficients $\{\alpha_k\}$, which are transmitted together with other critical parameters to the receiver as shown below.



The synthesized speech will be identical to the predicted speech (not the original though) at the transmitter if there is no quantization and channel error during the encoding and transmission, i.e., $\alpha_i' = \alpha_{i.}$

$$s'(n) = \frac{G'.u'(n)}{1-(\alpha_1.Z^{-1} + \alpha_2.Z^{-2} + \cdots + \alpha_P.Z^{-P})} \tag{12}$$

Generated by the inverse filter $H(Z)$:

$$H(Z) = \frac{1}{A(Z)} = \frac{1}{1-(\alpha_1.Z^{-1} + \alpha_2.Z^{-2} + \cdots + \alpha_P.Z^{-P})} \tag{13}$$

This inverse filter is guaranteed to be stable because $\{\alpha_k\}$ are obtained from R which is a Toeplitz form of (9a) for the autocorrelation method.

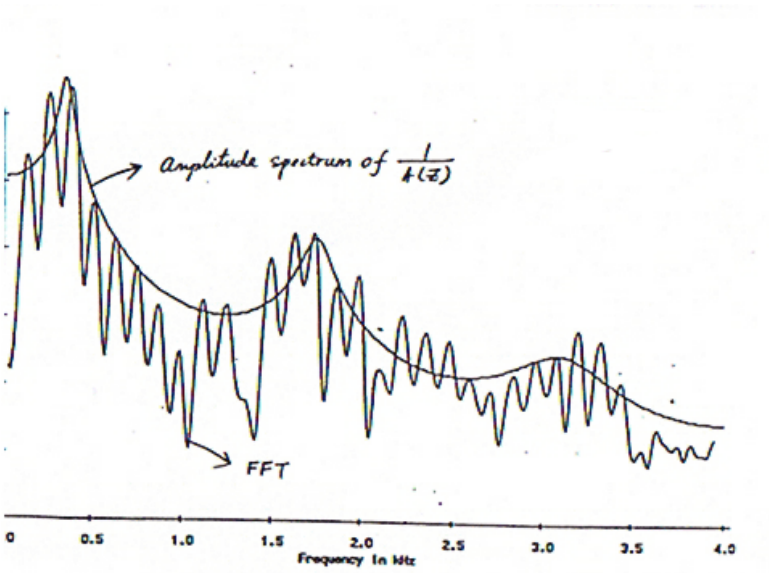
However, due to quantization errors, the LPC coefficients at the receiver are: $\alpha_1', \alpha_2', \cdots, \alpha_P'$ for LPC order P. The inverse filter $1/A(z)$ is formed by $\alpha_i's$ may have poles outside the unit circle with the potential of **instability**.

Typical LPC Coefficients: (For P=12)

α_1	1.6835	1.7193
α_2	-1.1644	-1.4482
α_3	0.7815	1.0891
α_4	-1.0581	-0.9600
α_5	0.8971	0.7701
α_6	-0.3929	-0.4010
α_7	0.5834	0.3430
α_8	-0.5470	-0.2484
α_9	-0.0495	-0.0101
α_{10}	0.2333	-0.1216
α_{11}	-0.0045	0.3203
α_{12}	-0.0816	-0.1535

The amplitude of the inverse filter:

$$\frac{1}{A(Z)} = 10\log|\frac{1}{A(e^{jwT})}|^2 \text{ as } w \text{ varies from } 0 \to w_S/2 \qquad (14)$$



Spectrum of the actual speech data and the LPC model spectrum.

Covariance Method:

The covariance function is measured similarly over a 20-30 ms segment of speech during which the characteristics assumed to be stationary. Segments again could be overlapped or non-overlapped. Here the matrix R and the vector G are defined by:

$$R = \begin{bmatrix} \phi_{0,0} & \phi_{0,1} & \phi_{0,2} & \cdots & \phi_{0,p-1} \\ \phi_{1,0} & \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,p-1} \\ \phi_{2,0} & \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,p-1} \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \cdot & & & & \cdot \\ \phi_{p-1,1} & \cdot & \cdot & \cdot & \phi_{p-1,P-1} \end{bmatrix} \quad G = \begin{bmatrix} \phi_{1,0} \\ \phi_{2,0} \\ \cdot \\ \cdot \\ \cdot \\ \phi_{p,0} \end{bmatrix} \qquad (15)$$

where we use:

$$\phi_{i,j} = \sum_{n=1}^{W} s(n-i).s(n-j) \qquad (16)$$

Note: Upper limit of the sum is fixed at W, the window size. In other words, we need to borrow samples from neighboring analysis windows.

$$x_0\ x_1 \ldots x_{255},\ x_{256},\ x_{257} \ \ldots\ x_{511}\ x_{512} \ \ldots$$

$$s(-1)\ s(0)\ s(1)\ s(2) \ \ldots \qquad s(256)$$

$$\phi_{0,0} = \sum_{n=1}^{256} s(n-0)s(n-0)$$

$$= s(1)^2 + s(2)^2 + \ldots + s(256)^2$$

$\longleftarrow$ 256 terms $\longrightarrow$

$$\phi_{1,1} = \sum_{n=1}^{256} s(n-1)s(n-1)$$

$$= s(0)^2 + s(1)^2 + \ldots + s(255)^2$$

$\longleftarrow$ 256 terms $\longrightarrow$

$$\phi_{1,0} = \sum_{n=1}^{256} s(n-1)s(n-0)$$

$$= s(0)s(1) + s(1)s(2) + \ldots + s(255)s(256)$$

$\longleftarrow$ 256 terms $\longrightarrow$

$$\phi_{0,1} = \sum_{n=1}^{256} s(n-0)s(n-1)$$

$$= s(1)s(0) + s(2)s(1) + \ldots + s(256)s(255)$$

$\Rightarrow \phi_{i,j} = \phi_{j,i}$ Even function.

In (15) **R** is a symmetric matrix but the diagonal elements $\phi_{0,0}, \phi_{1,1}, \cdots \phi_{p-1,p-1}$ may not be identical and **R** may not be a Toeplitz matrix. As more terms are used to calculate the covariance function in comparison with the auto method, the LPC coefficients obtained this way are more accurate. Unfortunately, the inverse filter formed from these coefficients:

$$H(Z) = \frac{1}{A(Z)}$$

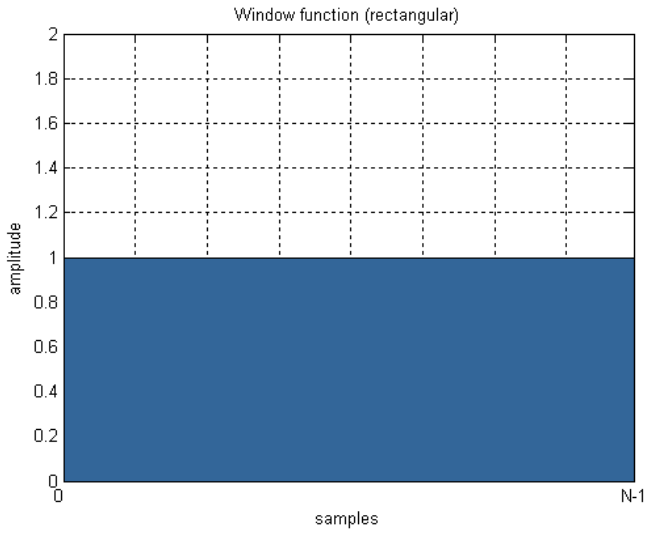
may not be stable because the *R* matrix is not Toeplitz.

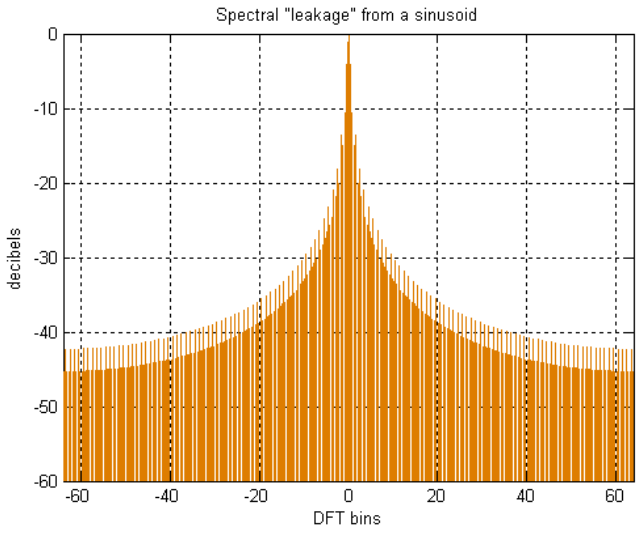
Windows:

The type of window selected for the segment of speech has also an impact on the spectrum due to the end-point transients in spectrum computation.

If no special window is chosen, we cay say that we have a rectangular window with:

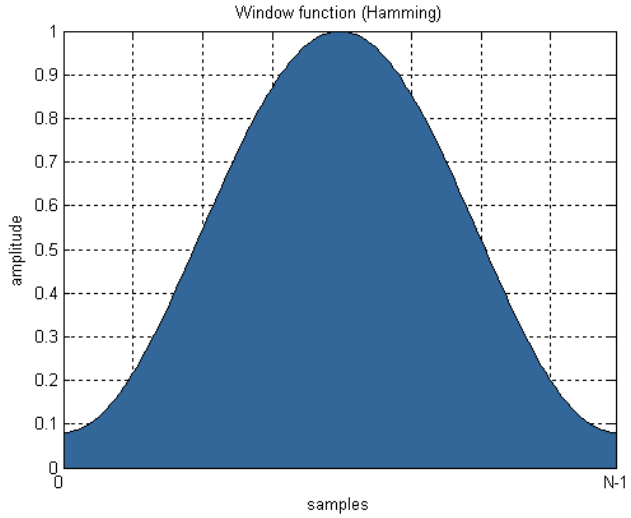
$$h_R(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & Otherwise \end{cases}$$

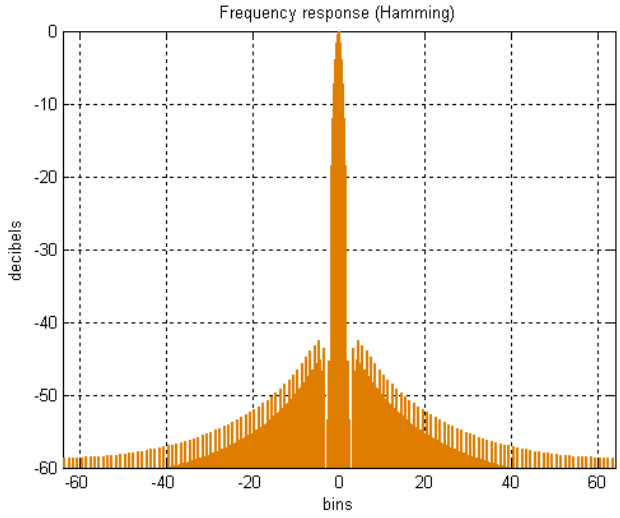




In practice, we normally use Hamming window to minimize the end-point effects.

$$h_H(n) = \begin{cases} 0.54 - 0.46Cos(\frac{2\pi n}{N-1}) & 0 \le n \le N-1 \\ 0 & Otherwise \end{cases}$$





Levinson-Durbin Recursion to solve Autocorrelation Method:

$$\begin{bmatrix} \Psi_0 & \Psi_1 & \Psi_2 & \cdots & \Psi_{p-1} \\ \Psi_1 & \Psi_0 & \Psi_1 & \cdots & \Psi_{p-2} \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \cdot & & & \cdot & \cdot \\ \Psi_{p-1} & \cdot & \cdot & \cdot & \Psi_0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \cdot \\ \cdot \\ \cdot \\ \Psi_p \end{bmatrix} \qquad (17)$$

$$\sum_{k=1}^{P} \alpha_k . \psi_{|i-k|} \quad for\ 1 \le i \le P$$

Recursion:

(1) Set $E^{(0)} = \psi_0$ (18)

(2) Calculate: $k_i = [\Psi_i - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} . \Psi_{|i-j|}] / E^{(i-1)} \qquad for\ 1 \le i \le P$ (19)

That is: P=1 $\qquad k_1 = [\Psi_1 - \sum_{j=1}^{0} \alpha_j^{(0)} . \Psi_{|1-j|}] / E^{(0)} = \frac{\Psi_1}{\Psi_{01}}$

(3) $\alpha_i^{(i)} = k_i$
$\alpha_1^{(1)} = k_1$ (20)

(4) $\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)}$
for $1 \le j \le i-1$ (21)

(5) $E^{(i)} = (1-k_i^2)E^{(i-1)}$
$E^{(1)} = (1-k_1^2)E^{(0)}$ (22)

If P=3, then carry on:

(2) $k_3 = \left\{ \psi_2 - \sum_{j=1}^{3-1} \alpha_j^{(3-1)} \psi_{|3-j|} \right\} / E^{(3-1)}$

(3) $\alpha_3^{(3)} = k_3$

(4) $\alpha_2^{(3)} = \alpha_2^{(2)} - k_3 \alpha_1^{(2)}$
$\alpha_1^{(3)} = \alpha_1^{(2)} - k_3 \alpha_2^{(2)}$

(5) $E^{(3)} = (1-k_3^2)E^{(2)}$
If P = 3
$\alpha_1 = \alpha_1^{(3)} \quad \alpha_2 = \alpha_2^{(3)}$
$\alpha_3 = \alpha_3^{(3)}$

Computation of the Gain (*G*) for the LPC Model:

$$G.u(n) = s(n) - \sum_{k=1}^{P} a_k .s(n-k)$$

At the transmitter (encoder), we first calculate k_i, s which give us α_i, s, we can calculate the prediction error $e(n)$ defined as:

$$e(n) = s(n) - \sum_{k=1}^{P} \alpha_k .s(n-k) \quad (23)$$

Assumptions and Computation:

1. $u(n) = \delta(n)$ (24)
2. $G.\delta(n) = s(n) - \sum_{k=1}^{P} \alpha_k .s(n-k)$ (25)
3. $\psi_m = \sum_n s(n).s(n+m)$

Then:

$$\psi_0 = \sum_n s(n).s(n)$$

$$= \sum_n s(n).[G.\delta(n) + \sum_{k=1}^{P} \alpha_k .s(n-k)] = \sum_n s(n).G.\delta(n) + \sum_{k=1}^{P} \alpha_k .s(n).s(n-k)$$

$$= \sum_n s(n).G.\delta(n) + \sum_{k=1}^{P} \alpha_k .\psi_k = \sum_n [G.\delta(n) + \sum_{k=1}^{P} \alpha_k .s(n-k)].G.\delta(n) + \sum_{k=1}^{P} \alpha_k .\psi_k$$

$$= \sum_n G^2 .\delta^2(n) + \sum_{k=1}^{P} \alpha_k . \sum_n G.\delta(n).s(n-k) + \sum_{k=1}^{P} \alpha_k .\psi_k$$

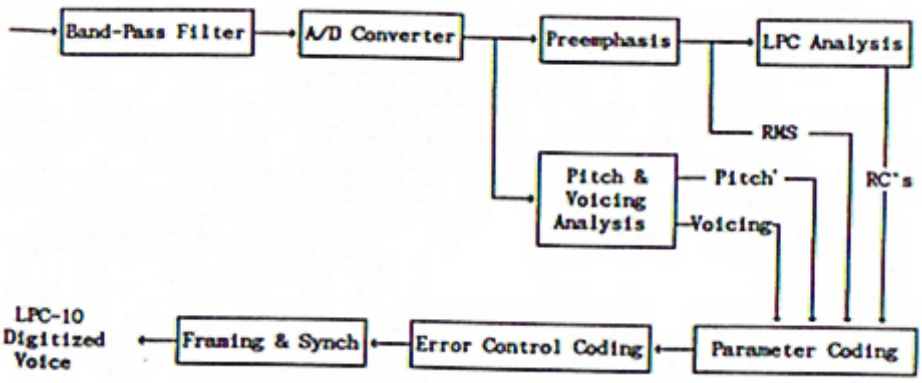
$$= G^2 + 0 + \sum_{k=1}^{P} \alpha_k .\psi_k$$

Therefore, we have:

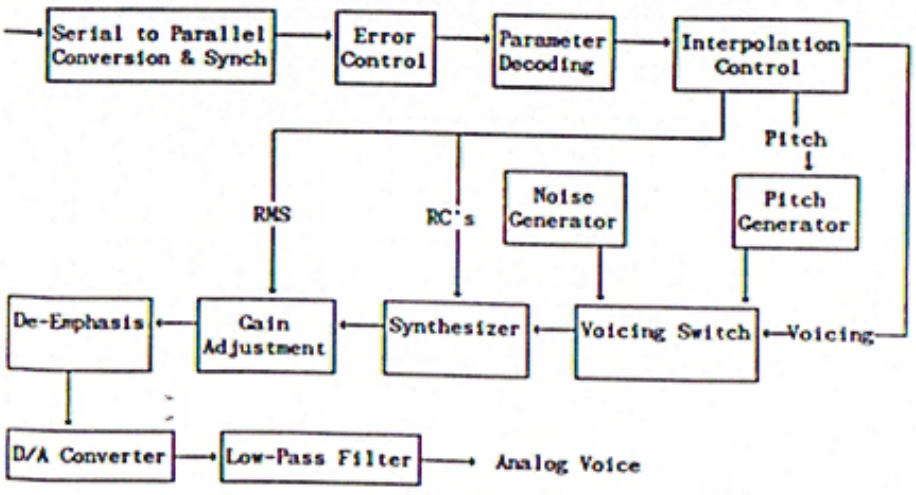
$$G^2 = \psi_0 - \sum_{k=1}^{P} \alpha_k .\psi_k \qquad (26)$$

This gain (G) is then quantized and transmitted to the receiver.

Typical LPC-based Transmitter:



Typical LPC-based Receiver:



Some key points for LPC-10 based FED-STD 1015:

1. Input Signal conditioning with a BP filter (100-3600 Hz):
   - Less than 1.0 dB of in-band ripple
   - 3.0 dB attenuation at 100 Hz. and 3600 Hz.
   - 23 dB attenuation at 4,000 Hz; 46 dB at 4,400 Hz.
2. Sampling: $F_S = 8.0 \pm 0.1\%\ kHz$
3. A/D Conversion: 12-bits uniform quantizer
4. Pre-emphasis: First-order digital filter with: $1 - 0.9375 * Z^{-1}$
5. Frame length: 22.5 ms=180 samples
6. Number of bits for encoding per frame: 54 bits/frame
7. Bit rate: 54/22.5 ms= 2400 bits/second.

In most systems, the prediction coefficients $\alpha_k's$ are transformed to equivalent sets through some 1:1 transformations, such as LSP pairs recently or "*Reflection Coefficients*" (RC) in earlier LPC systems or sine transforms.

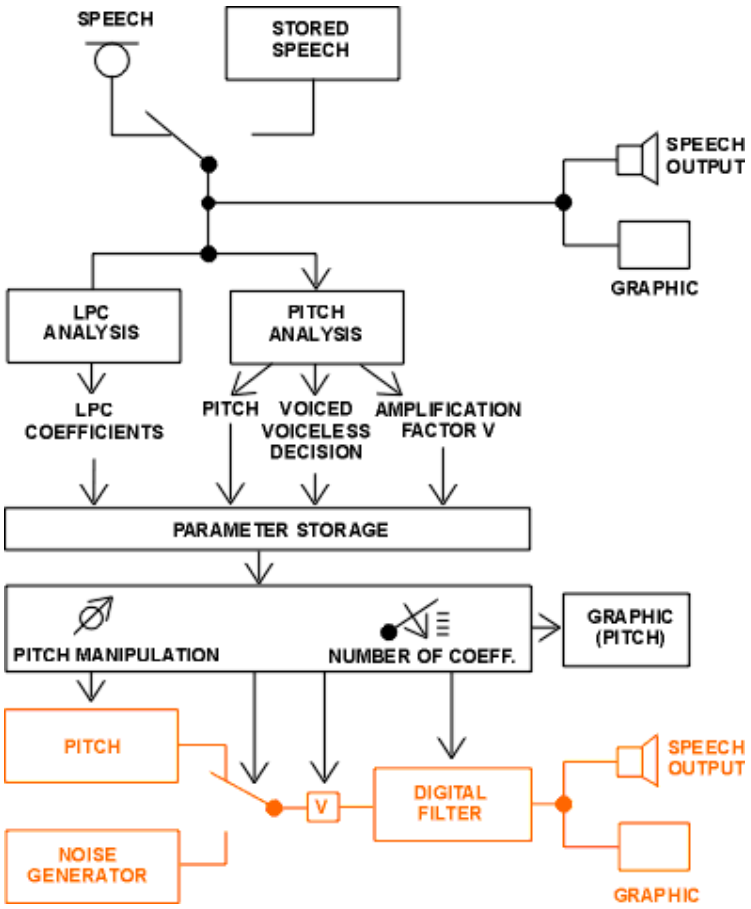
Bit Assignment for FED-STD 1015:

Bit assignment is the next critical step in linear prediction based encoding or communication systems. The Federal Standard-1015 is the standard established in mid 1970's and used throughout the world for military and civilian applications.

	Voiced	Nonvoiced
Pitch & Voicing	7	7
RMS Amplitude	5	5
RC(1)	5	5
RC(2)	5	5
RC(3)	5	5
RC(4)	5	5
RC(5)	4	
RC(6)	4	
RC(7)	4	
RC(8)	4	
RC(9)	3	
RC(10)	2	
Error Control		20
Synchronization	1	1
Unused		1
Total	54	54

Complete Example of a Fed-Std 1015 2.4kbps LPC Codec Implementation

- Block diagram of a 2.4 kbps LPC Codec:



- LPC coefficients are represented as ***line spectrum pair*** (LSP) parameters.
- LSP are mathematically equivalent (one-to-one) to LPC.
- LSP are more amenable to quantization and calculated as follows:

$$\begin{aligned} P(z) &= 1+(a_1-a_{10})z^{-1}+(a_2-a_9)z^{-2}+\ldots+(a_{10}-a_1)z^{-10}-z^{-11} \\ Q(z) &= 1+(a_1+a_{10})z^{-1}+(a_2+a_9)z^{-2}+\ldots+(a_{10}+a_1)z^{-10}+z^{-11} \end{aligned}$$

- Factoring the above equations, we get:

$$\begin{aligned} P(z) &= (1-z^{-1})\prod_{k=2,4,\ldots,10}(1-2\cos w_k z^{-1}+z^{-2}) \\ Q(z) &= (1+z^{-1})\prod_{k=1,3,\ldots,9}(1-2\cos w_k z^{-1}+z^{-2}) \end{aligned}$$

$\{w_k\,;\,k=1,2,\cdots,10\}$ are called the LSP parameters.

- LSP are *ordered* and *bounded*:

$$0<w_1<w_2<\cdots<w_{10}<\pi$$

- LSP are more correlated from one frame to the next than LPC.
- The frame size is 20 msec. There are 50 frames/sec. 2400 bps is equivalent to 48 bits/frame. These bits are allocated as follows:

Parameter Name	Parameter Notation	Rate (bits/frame)
LPC or LSP	$\{a_k\,;\,k=1,2,\cdots,10\}$ or $\{w_k\,;\,k=1,2,\cdots,10\}$	34 or 34
Gain	***G***	7
Voiced/Unvoiced/Pitch	V/UV/T	7
Total		**48**

- The 34 bits for the LSP are allocated as follows:

LSP Index	No. of Bits assigned
w_1	3
w_2, w_3, w_4, w_5	4
$w_6, w_7, w_8, w_9, w_{10}$	3
Total	**34**

- The gain, ***G***, is encoded using a 7-bit non-uniform scalar quantizer (a 1-dimensional vector quantizer).

- For voiced speech, values of ***T*** ranges from 20 to 146. ***V/UV,T*** are jointly encoded as follows:

V/UV	T	Encoded Value
UV	-	0
V	20	1
V	21	2
V	22	3
V	23	4
:	:	:
V	146	127

LPC10 (2400 bps) Demo