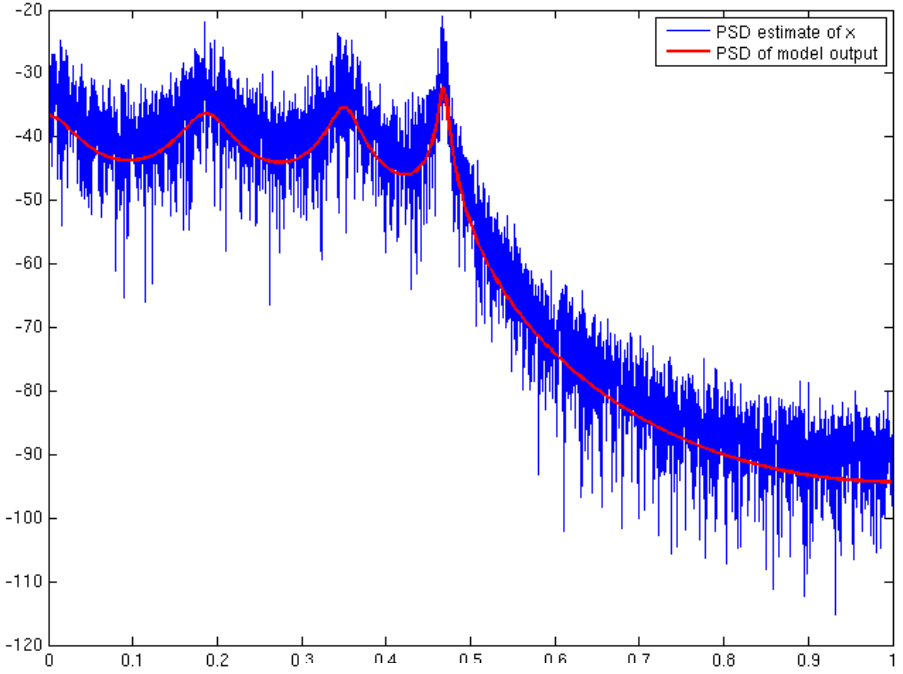
Linear Prediction Coding



Nimrod Peleg

Update: March 2009

Linear Prediction and Speech Coding

- The earliest papers on applying LPC to speech:
  - Atal 1968, 1970, 1971
  - Markel 1971, 1972
  - Makhoul 1975
- This is a family of methods which is widely used: from standard telephony (toll quality), to military communication (low quality).
- Typical rates: 800-16Kbps

General Overview

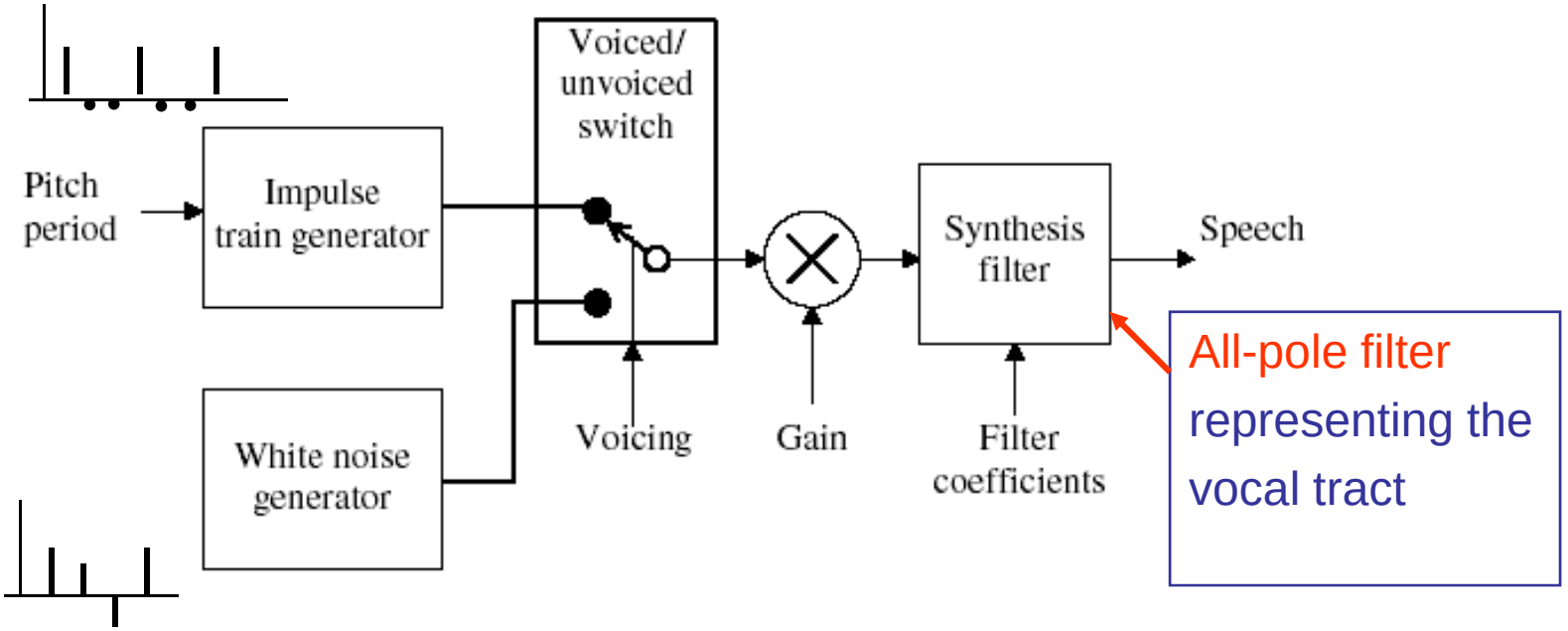
- LPC is a model for speech signal production: based on the assumption that the speech signal is produced by a very specific model.
- The basic idea is very simple, but there are many different ways of looking at it.
- There are many variants over the basic scheme: LPC-10, CELP, MELP, RELP, VSELP, ASELP, LD-CELP…

The Model and Its Variants

- All LPC variants are based on the same simple model : an excitation signal and a filter.
- The most simple case - the excitation is an impulse train OR white noise

- Most variants consists of a more advanced excitation signal (***hybrid*** *coders*)

Speech Production Model



The Classical model of speech production: Linear Prediction Coefficients (LPC)

- Voicing: Voiced or unvoiced speech frame
- Gain: Energy level of the frame
- Filter Coefficients: Synthesis filter response
- Pitch period: Time duration between consecutive excitation pulses (voiced)

What is it Good For ?

- A coding scheme which is closely related to the *model* of signal production can lead to an **efficient representation** of the signal.
- That's why LPC is efficient in exploiting the *parametric redundancy*.

- This is true ONLY for *monophonic* signals like speech (as opposed to audio)

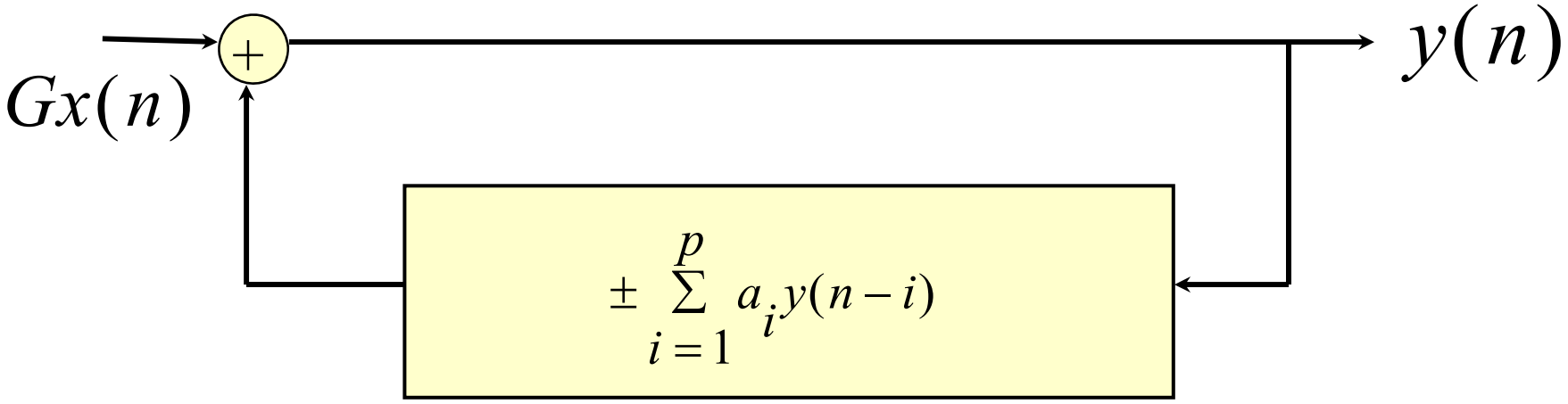
A Mathematical Representation

- This model is equivalent to a signal produced by a *difference equation*:

$$y(n) = \sum_{i=1}^{p} a_i y(n-i) \pm G x(n)$$

  - $x(n)$ is the excitation signal
  - $y(n)$ is the speech signal
  - G is "history parameter"
  - a_i are the filter coefficients
- the +/- sign is arbitrary (as long as it is consistent)

Another Representation Option



- Consider the block to be a *predictor*, which tries to predict the current output as a linear combination of previous outputs (hence *LPC*)
- The predictor's input is the *prediction error* (innovation, residual...)

Parameter Estimation Process

- The parameter estimation process is repeated for each frame, with the results representing information on the frame.
- Thus, instead of transmitting the PCM samples, **parameters** of the model are sent.
- By carefully allocating bits for each parameter so as to minimize distortion, an impressive compression ratio can be achieved – up to 50-60 times !
  - The cost: loss of quality (communication applications)

Speech Model Parameters

- Estimating the parameters is the responsibility of the encoder.
- The decoder takes the estimated parameters and uses the speech production model to synthesize speech.
- the **output waveform** (synthesized speech), is completely different from the original !
- The point is that the power spectral density of the original speech is captured by the synthesis filter.
- Therefore, the PSD of the synthetic speech is close to the original (due to the **flat spectrum** of the input excitation)

PSD: power spectral density

Phase Information

- The approach throws away all **phase** information of the original waveform, preserving only the magnitude of the frequency spectrum.
- The synthetic waveform sounds like the original because, **for a human listener**, phase has a relatively lower rank than magnitude information.
- This phenomenon is the reason why signal-to-noise ratio (SNR) is a poor, and sometimes, **senseless** measure of speech quality.

The Synthesis Filter

- The synthesis filter shapes the flat spectrum of the noise input so that the output imitates the envelope of the original spectrum.
- It is important to note that this is true only for **noise excitation** in the unvoiced case !
- for the voiced case, however, the input is an impulse train sequence of regularly spaced impulses.
- This violation of the basic model for voiced signal is one of the fundamental limitations of the LPC model for speech production !

Excitation by Impulse Train

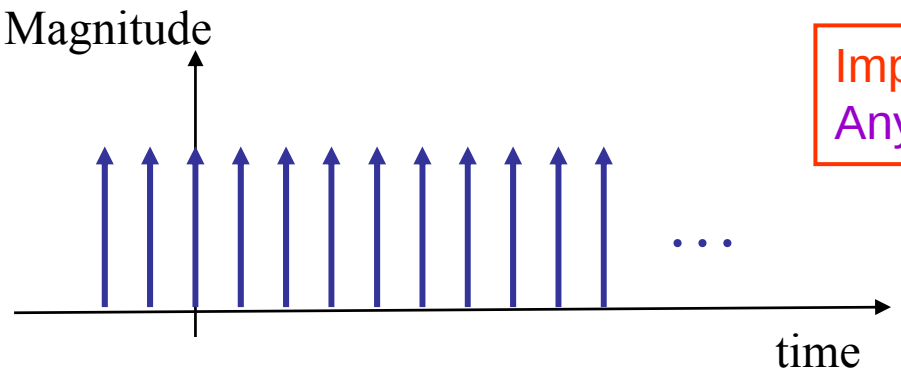
The impulse train for excitation is given by:

(T: a positive constant, being the period)

$$\sum_{i=-\infty}^{\infty} \delta_{(n-iT)}$$

where $\delta_{(n)} \begin{cases} 1, & \text{if n=0} \\ 0, & \text{Otherwise} \end{cases}$

- The use of a periodic impulse train is to create periodicity in the output waveform, so that the resulting signal possesses a PSD that resembles voiced signals.

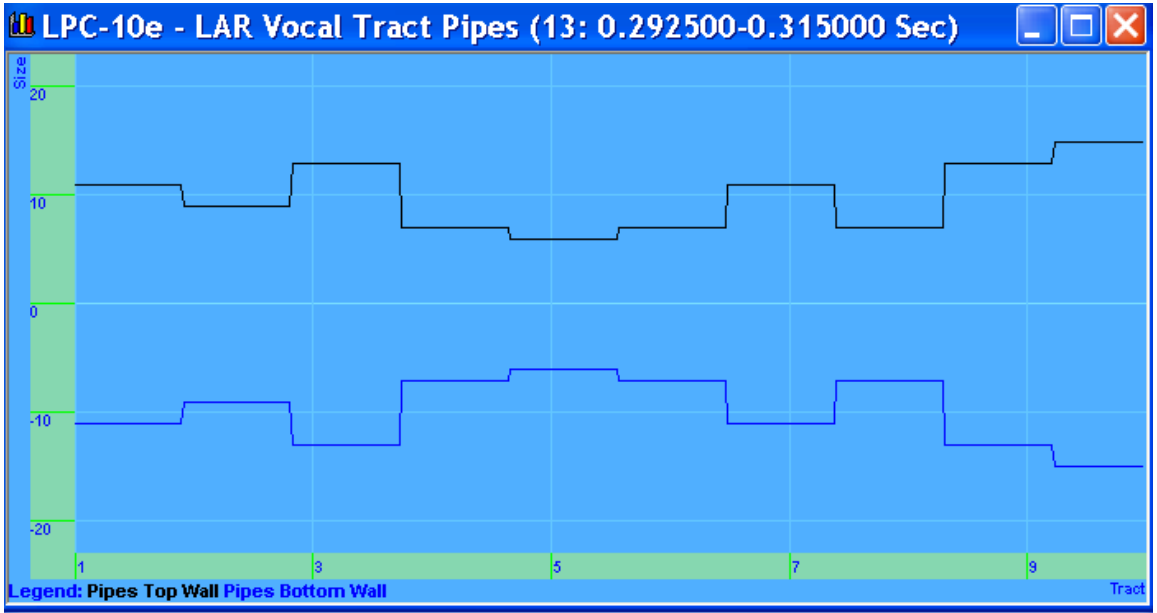


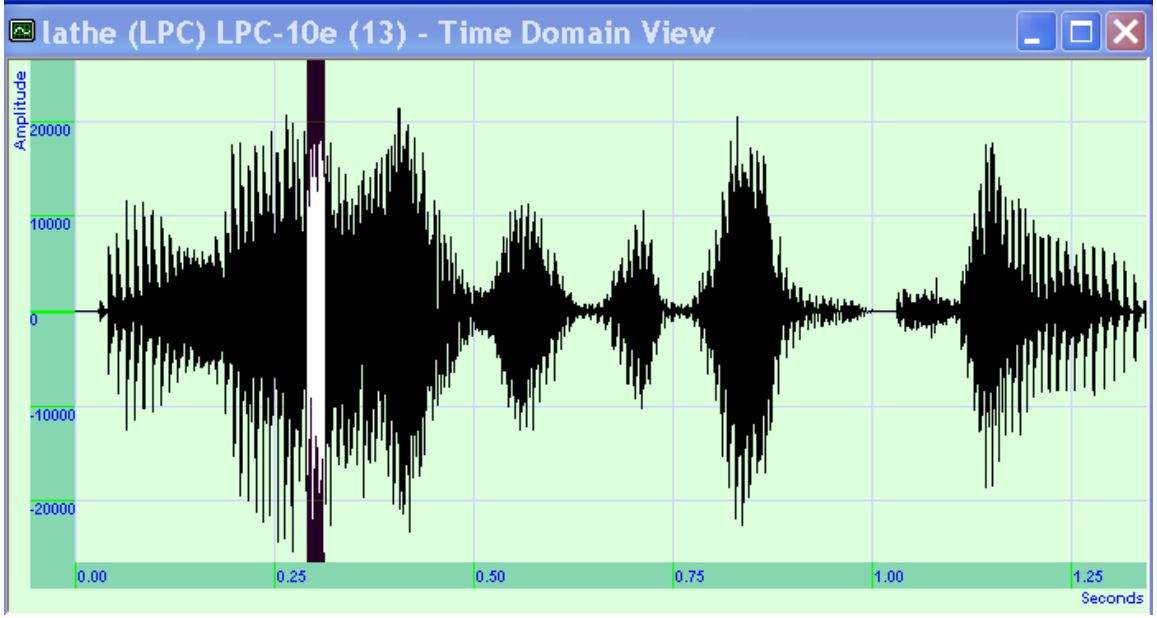
Impulse Response completely **defines**
Any Linear-Time-Invariant (LTI) system !

The Filter Coefficients

- Since the coefficients of the synthesis filter must be quantized and transmitted, only a few of them are calculated , to maintain low bit-rate.
- A prediction order of ten is in general enough to capture the spectrum envelope for unvoiced frames
- For voiced frames, a much higher order is required due to correlation of distant samples.
- The LPC coder solves this by using an **impulse train input**: if the period of the input excitation matches the original pitch, periodicity is introduced to the synthetic speech with a PSD that is similar to the original.
- In this manner, high prediction order is avoided, thus achieving the low bit-rate objective.

What do the Filter Represent ?



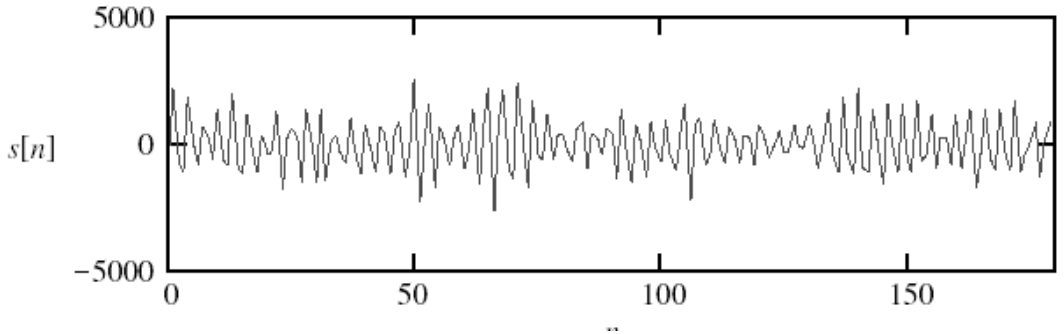


Remember the "pipelines model" ?

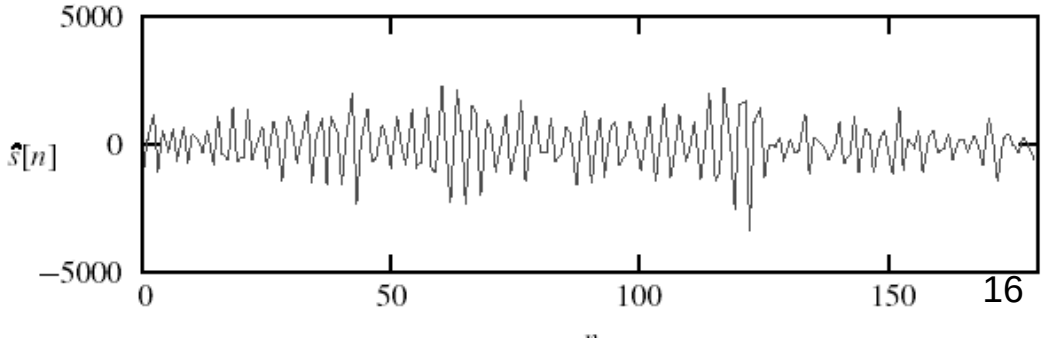
From: SPDemo 3.0

LPC Coding: Unvoiced Example

Unvoiced frame
having 180 samples:

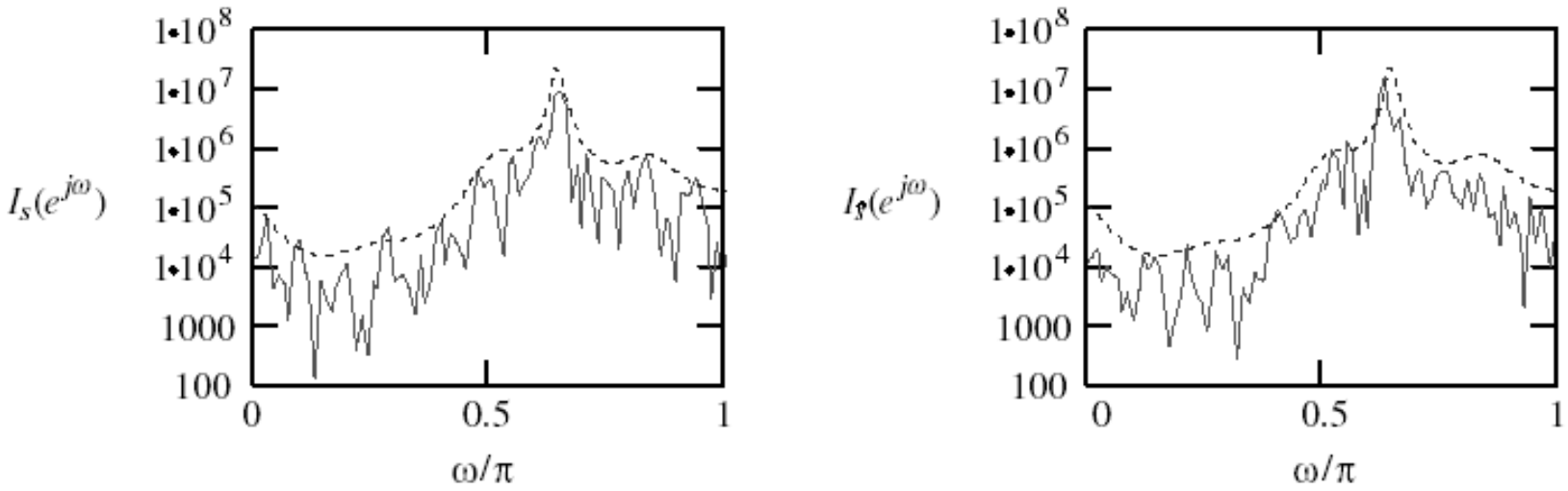


Since the frame is unvoiced, white noise with uniform distribution is used as excitation. The generated white noise has unit variance, with a gain (energy level). As we can see, in the time domain the two waveforms are completely different. They sound similar because the power spectral densities have similar shapes.



Unvoiced Example: Cont'd

Power Spectrum Envelope



Left: Original Right: synthetic

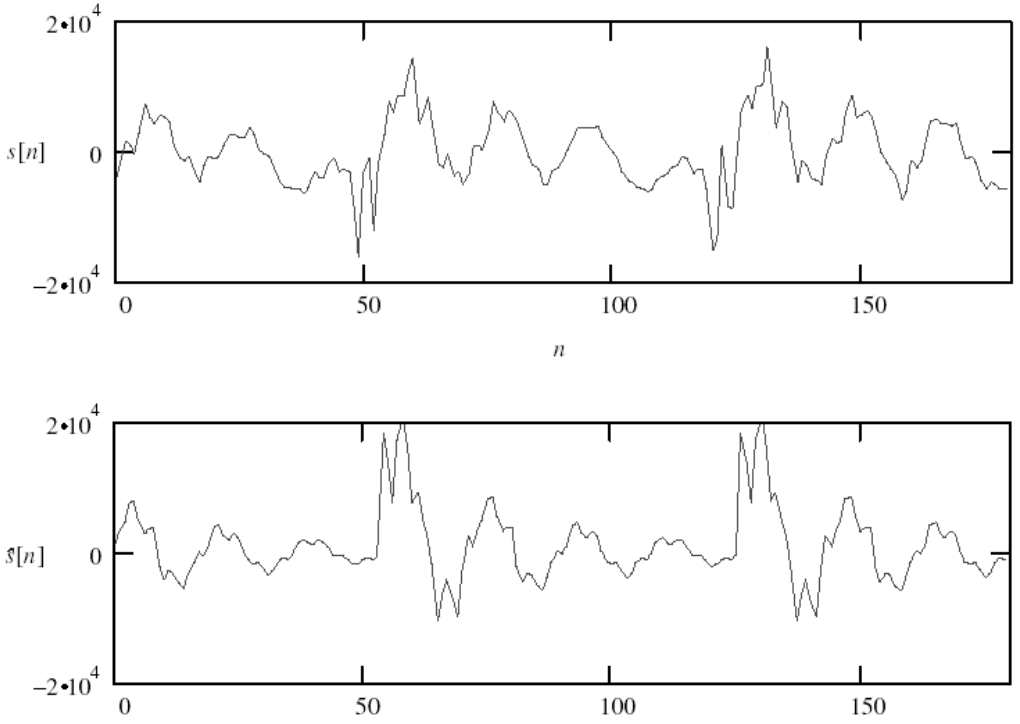
Plots of **power spectrum** for an unvoiced frame

Dotted line : The PSD using the estimated LPC

LPC Coding: Voiced Example

Voiced Frame:
180 samples, Pitch=75

Synthetic speech is generated using a train of impulses with unit amplitude, scaled by a gain term so that the energy level of the synthetic speech matches the original.



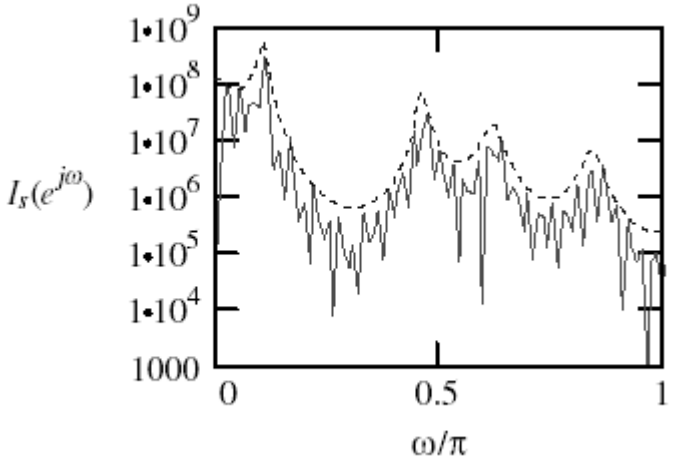
Voiced frames plots

Top: Original bottom: synthetic.

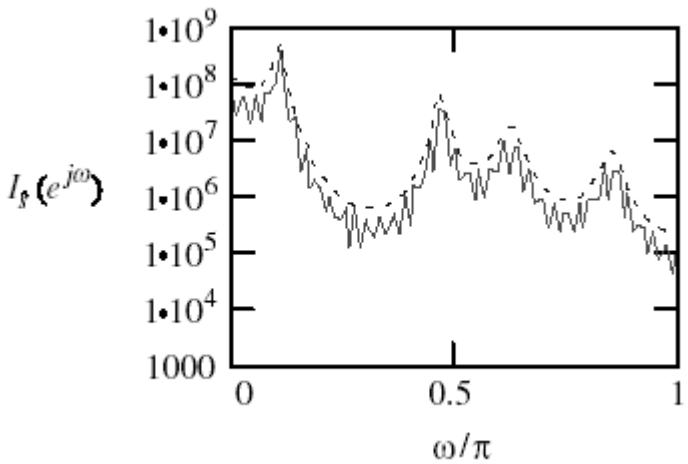
Voiced Example Cont'd

The periodograms of the two frames have similar appearance. Presence of harmonic components due to periodicity of the frame is also evident from the regularly separated peaks.

For the original signal, the structure of the harmonics looks more irregular and randomized, while **for the synthetic case**, a more regular structure appears.

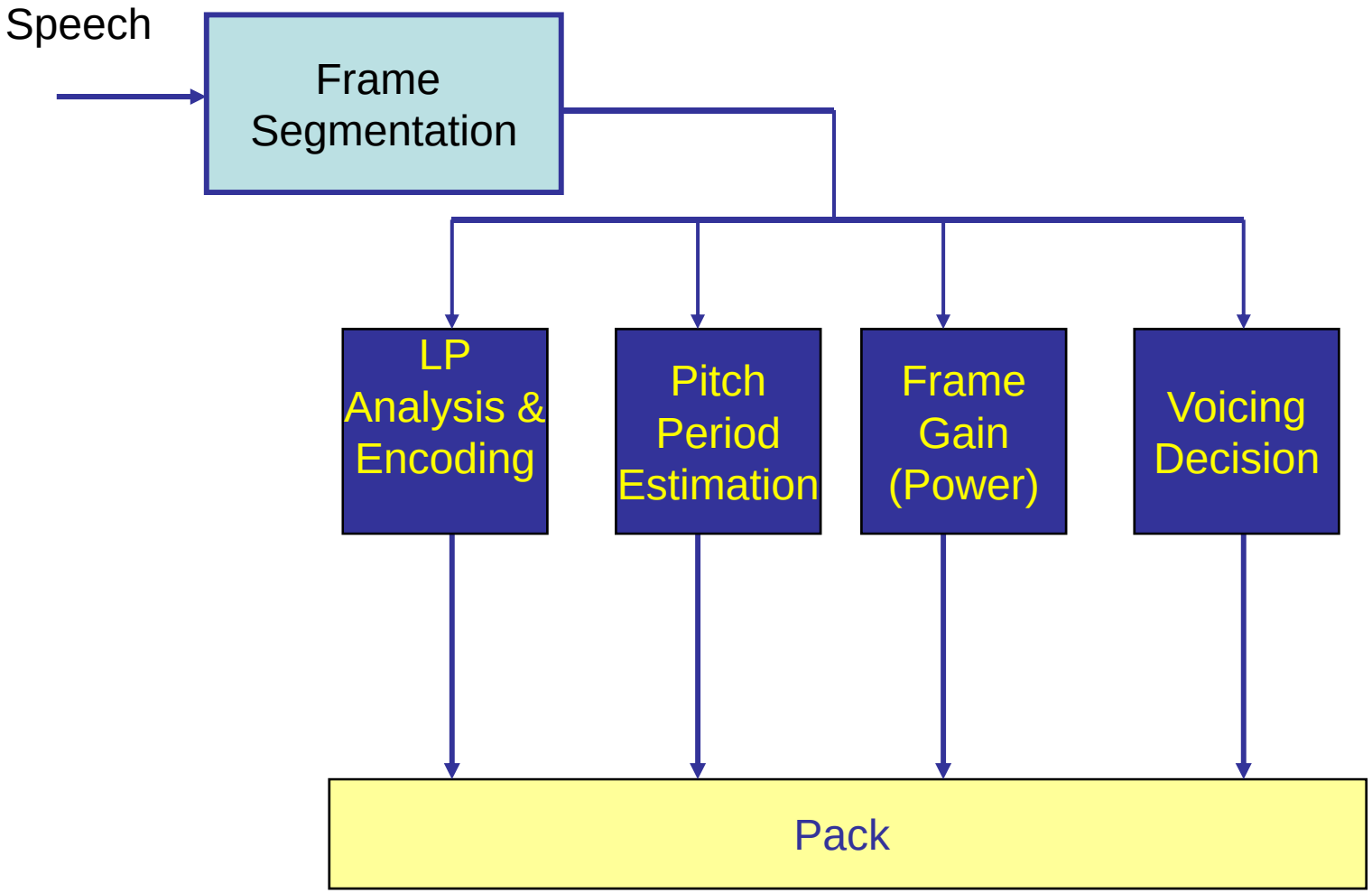


Original

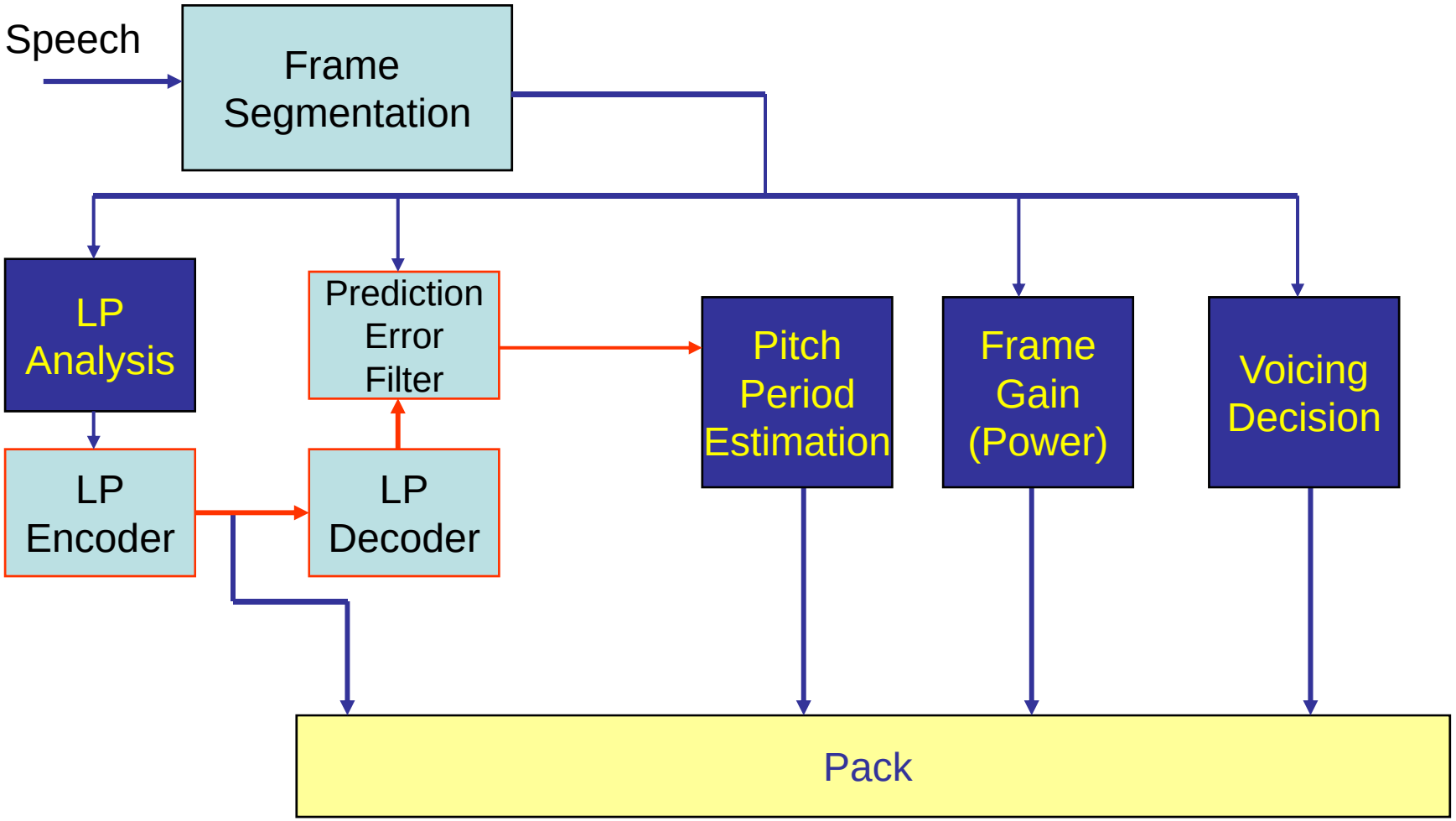


Synthetic

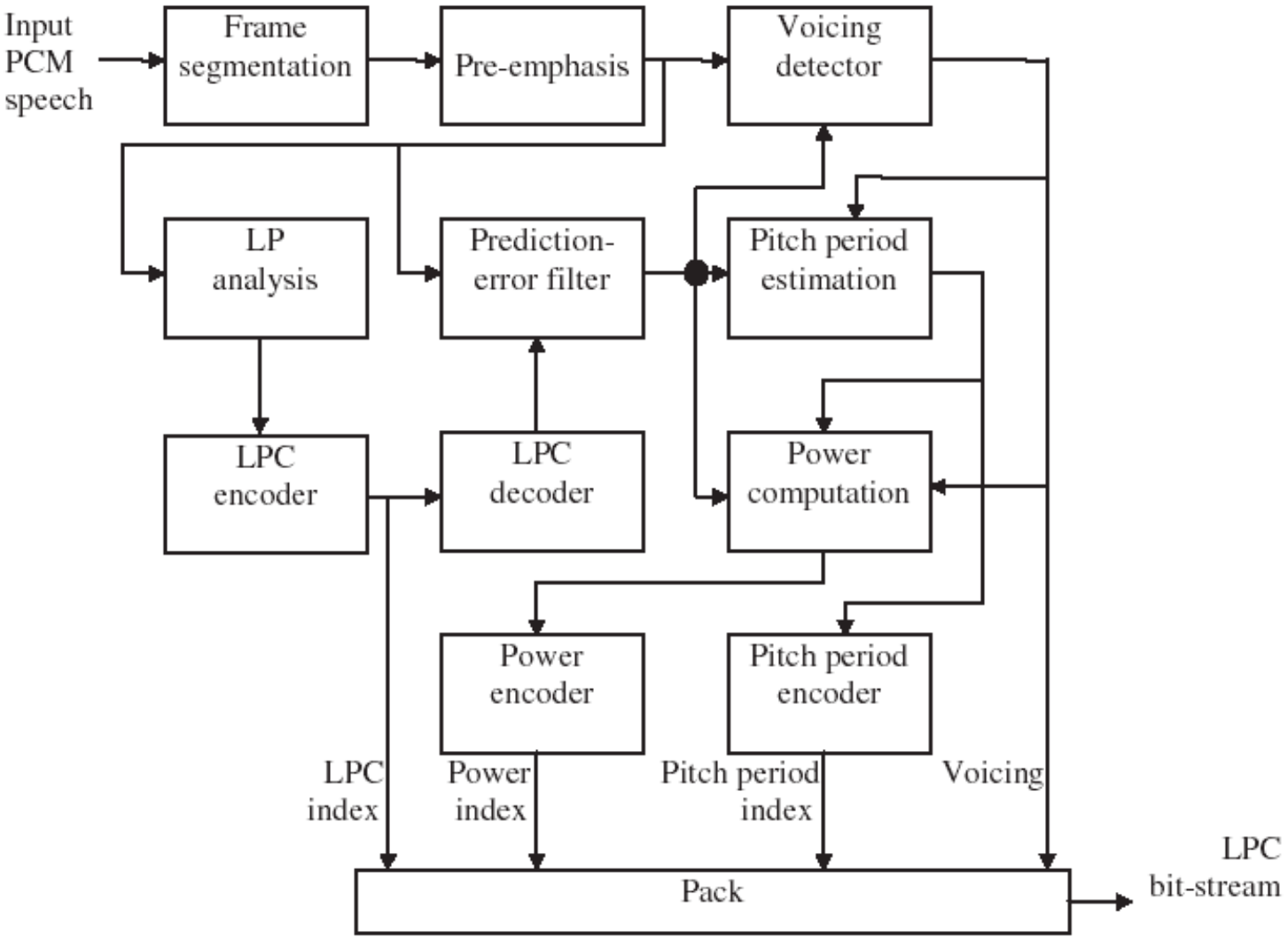
A Simple LPC Encoder



Using Prediction Error Filter



Complete LPC Encoder *

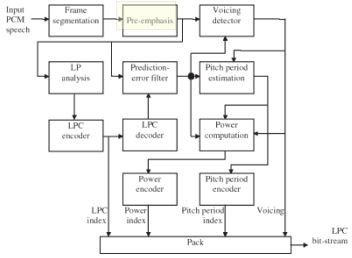


* Based on FS-1015 (LPC-10)

LPC Encoder Structure

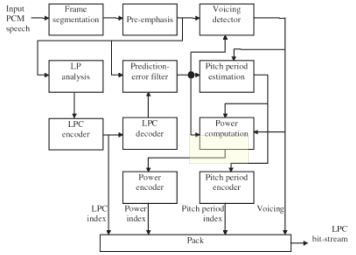
1. The input speech is segmented into non-overlapping frames.
2. A pre-emphasis filter is used to adjust the spectrum of the signal.
3. The voicing detector classifies the current frame as voiced or unvoiced and outputs one bit indicating the voicing state.
4. The pre-emphasized signal is used for LP analysis, where ten LPC coefficients are derived.
5. These coefficients are quantized with the indices transmitted as information of the frame.
6. The quantized LPCs are used to build the prediction-error filter, which filters the pre-emphasized speech to obtain the **prediction-error signal at its output.**
7. Pitch period is estimated from the prediction-error signal if the frame is voiced.

   By using the prediction-error signal as input to the pitch period estimation algorithm, a more accurate estimate can be obtained since the formant structure (spectrum envelope) due to the vocal tract is removed.



Pre-Emphasis

- Typical spectral envelope of speech signal has a high frequency attenuation due to radiation effects of the sound from the lips: high-frequency components have relatively low amplitude - this increases the dynamic range of the speech spectrum.
- As a result, LP analysis requires high precision to capture the features at the high end of the spectrum.
  - More importantly, when these features are very small, the correlation matrix can become ill-conditioned and even singular, leading to computational problems.
- One simple solution is to process the speech signal using the filter which is high-pass in nature.
  - The purpose is to increase the energy of the high-frequency spectrum. The effect of the filter can also be thought of as a flattening process, where the spectrum is ‘‘whitened.’’



Power Calculation

- Power of the prediction-error sequence is different for voiced and unvoiced frames.
- For the unvoiced case, denoting the prediction-error sequence as:
  - with N being the length of the frame.

$$P = \frac{1}{N}\sum_{n=0}^{N-1} e^2[n]$$

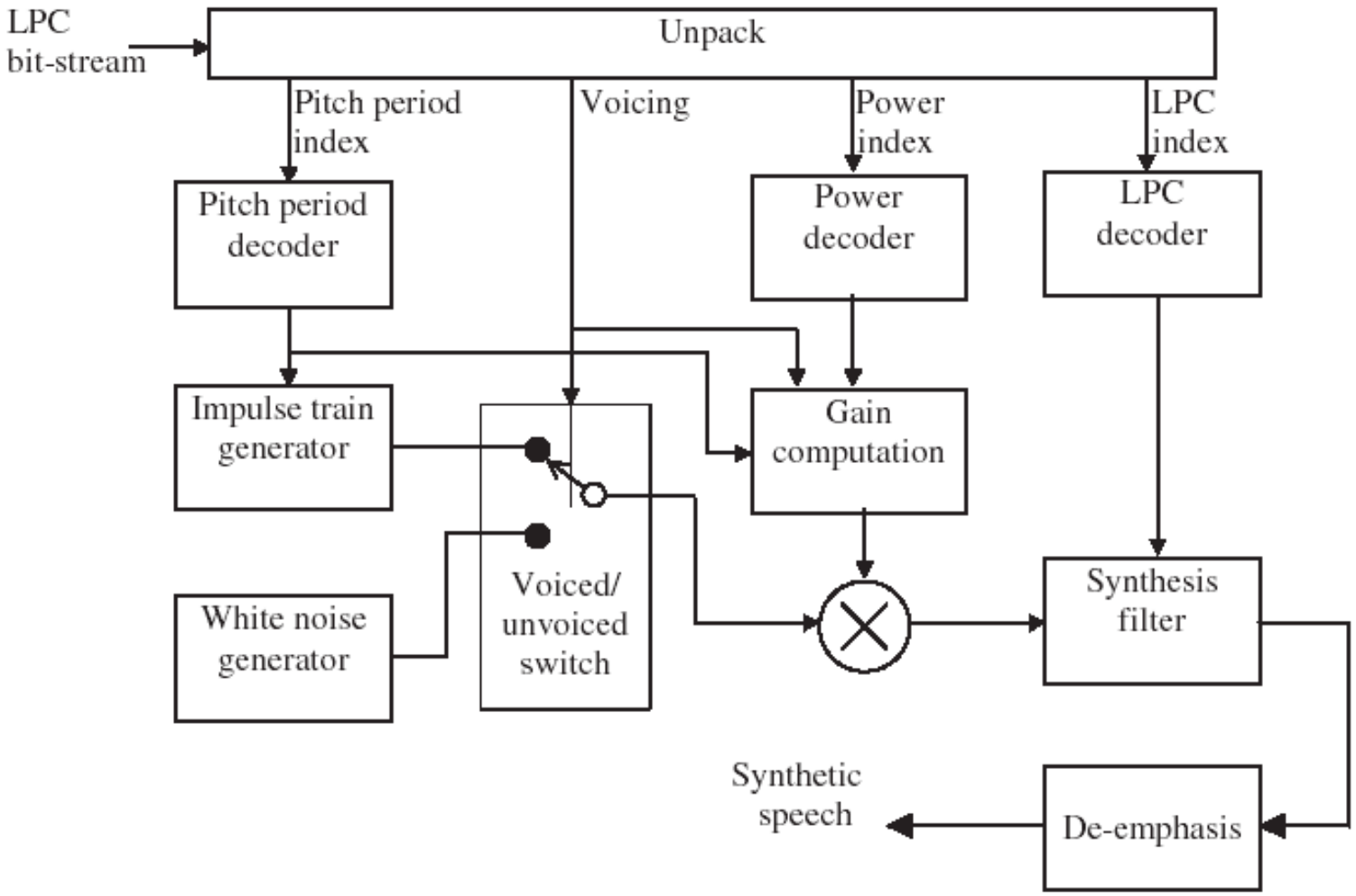
- For the voiced case, power is calculated using an integer number of pitch periods:

$$P = \frac{1}{\lfloor N/T_p \rfloor T_p}\sum_{n=0}^{\lfloor N/T \rfloor T-1} e^2[n]$$

e(n) , n ∈ [0,N-1] ; N: Frame Length
Tp: Pitch period

It is assumed that N > Tp, and hence use of the floor function ensures that the summation is always performed within the frame's boundaries ! (for pitch period synchronization purpose)

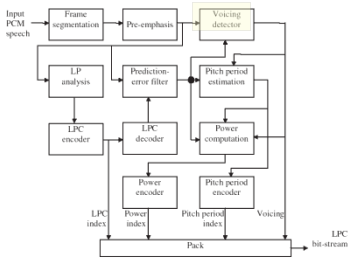
LPC Decoder *



* Based on FS-1015 (LPC-10)

LPC Decoder Structure

- The decoder is essentially the LPC model of speech production with parameters controlled by the bit-stream.
- Gain computation is performed separately for Voiced and Unvoiced frames, according to the energy of each.
- Finally, the output of the synthesis filter is de-emphasized to yield the synthetic speech.



Voicing Detector

- Purpose: classify a given frame as voiced or unvoiced.
- The boundary between V/UV is not always clear: this happens for transition frames, where the signal goes from voiced to unvoiced or vice versa.
- The necessity to perform a strict V/UV classification is one of the limitations of the LPC model.
- It is a critical component, since misclassification of voicing states can have disastrous consequences on the quality of the synthetic speech.

Voicing Detector: Energy

- Typically, voiced sounds are several order of magnitude higher in energy than unvoiced.
- For a frame (of length N) ending at instant *m*, the energy is given by:

$$E_{[m]} = \sum_{n=m-N+1}^{m} s^2_{[n]}$$

- The Magnitude Sum Function serves a similar purpose:

$$MSF_{[m]} = \sum_{n=m-N+1}^{m} \left| s_{[n]} \right|$$

- Since voiced speech has energy concentrated in the low-frequency region, better discrimination can be obtained by low-pass filtering the speech signal prior to energy calculation.
- A bandwidth of 800 Hz is adequate for the purpose since the highest pitch frequency is around 500 Hz.

Voicing Detector: Zero Crossing Rate

- The zero crossing rate of the frame ending at time instant m is defined by:

$$SC_{[m]} = \frac{1}{2} \sum_{n=m-N+1}^{m} \left| \operatorname{sgn}(s_{[n]}) - \operatorname{sgn} s_{[n-1]} \right|$$

  - the *sgn* function returning 1 depending on the sign of the operand.

- For voiced speech, the zero crossing rate is relatively low due to the presence of the pitch frequency component (of low-frequency nature)
- For unvoiced speech, the zero crossing rate is high due to the noise-like appearance of the signal with a large portion of energy in the high-frequency region.

Voicing Detector: Prediction Gain

- Defined as the ratio between the energy of the signal and the energy of the prediction error:

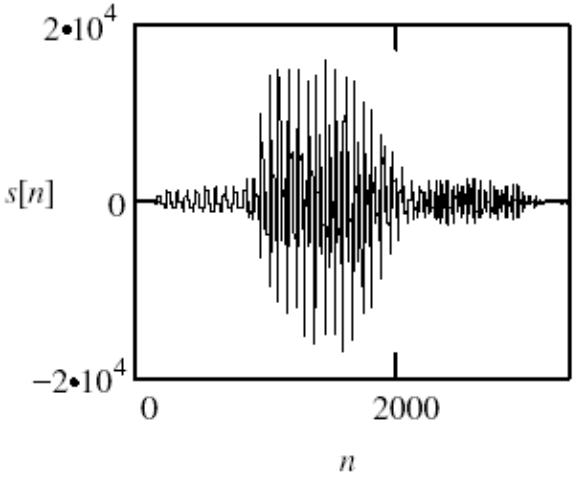
$$PG_{[m]} = 10\log_{10}\left(\frac{\sum_{n=m-N+1}^{m} s^2[n]}{\sum_{n=m-N+1}^{m} e^2[n]}\right)$$

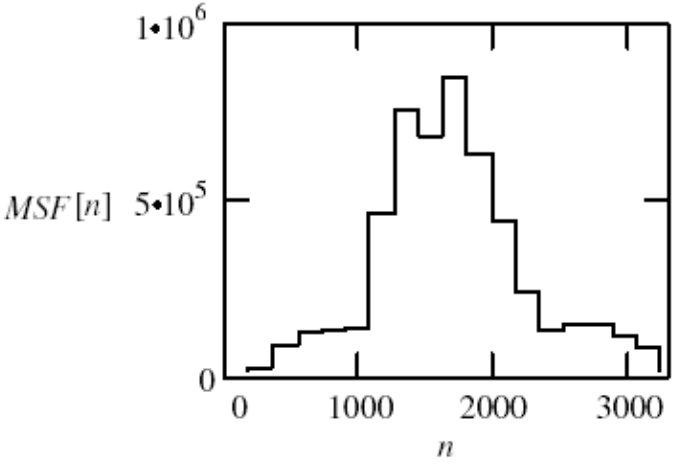
• Voiced frames on average achieve 3 dB or more in prediction gain than unvoiced frames, mainly due to the fact that periodicity implies higher correlation among samples, and thus easier to predict.

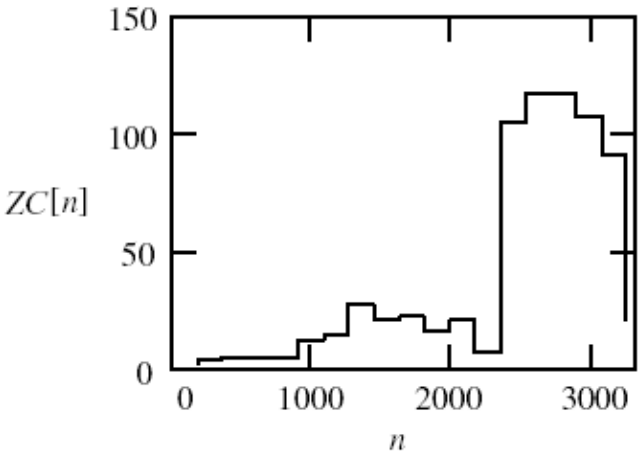
• Unvoiced frames, are more random and therefore less predictable.

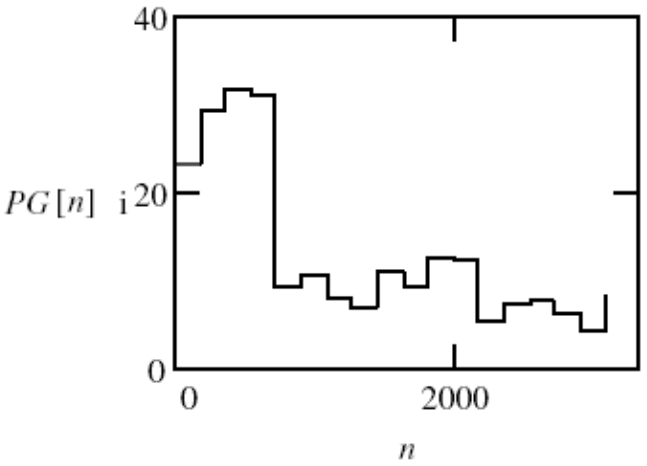
  • For very low-amplitude frames, prediction gain is normally not calculated to avoid numerical problems; in this case, the frame can be assigned as unvoiced just by verifying the energy level.

Voicing Detector Example









Calculated in frames of 180 samples, where the parameters assumed to be constant

Voicing Detector Example cont'd

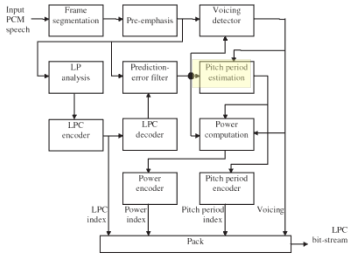
- Roughly speaking, the signal is voiced for n < 2200, and unvoiced beyond that limit.

- **For n < 1000**, the signal has low amplitude but is periodic.

The calculated parameters reflect this property of the signal: for n < 1000, the magnitude sum and zero crossing rate are low, with high prediction gain, typical of a low-amplitude voiced frame.

- **For 1000 < n < 2200**, the energy is high, and the zero crossing rate is low with medium prediction gain, common for most voiced frames.

- **For n > 2200**, energy and prediction gain are low with high zero crossing rate, typical characteristics of unvoiced frames.



Pitch Period Estimation

- The time between successive vocal cord openings is called the fundamental period, or pitch period.
- **For men**, the possible pitch frequency range is usually found somewhere between 50 and 250Hz, while **for women** the range usually falls between 120 and 500 Hz.
- In terms of period, the range for a **male** is 4 to 20 ms, while for a **female** it is 2 to 8 ms.

Pitch Estimation Cont'd

- Design of a pitch period estimation algorithm is a complex undertaking due to:
  - **Lack** of perfect periodicity
  - **Interference** with formants of the vocal tract
  - **Uncertainty** of the starting instance of a voiced segment
  - **Other real-world** elements such as noise and echo

- In practice, pitch period estimation is implemented as a trade-off between computational complexity and performance.

- Many techniques have been proposed for the estimation of pitch period and only a few will be reviewed here.

Pitch Estimation I:
The Autocorrelation Method

- The autocorrelation value reflects the similarity between the frame $s_{[n]}$ and the time-shifted version $s_{[n-l]}$
  - $n = [m-N+1, m]$
  - l is a positive integer representing a time lag.

$$R_{[l,m]} = \sum_{n=m-N+1}^{m} S[n]S[n-l]$$

The range of lag is selected so that it covers a wide range of pitch period values.
For instance, for l=20 to 147 (2.5 to 18.3 ms), the possible pitch frequency values range from 54.4 to 400 Hz at 8kHz sampling rate.
• This range of l is applicable for most speakers and can be encoded using 7 bits, since there are 2^7=128 values of pitch period.

The Autocorrelation Method Cont'd

- By calculating the autocorrelation values for the entire range of lag, it is possible to find the value of **lag** associated with the highest autocorrelation representing the pitch period estimate.
- in theory, autocorrelation is maximized when the lag is equal to the pitch period.

The Autocorrelation Method Cont'd

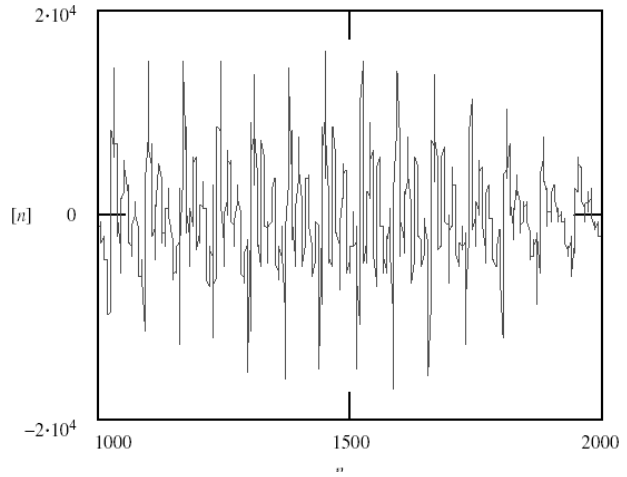
- The method is summarized with the following pseudo-code:

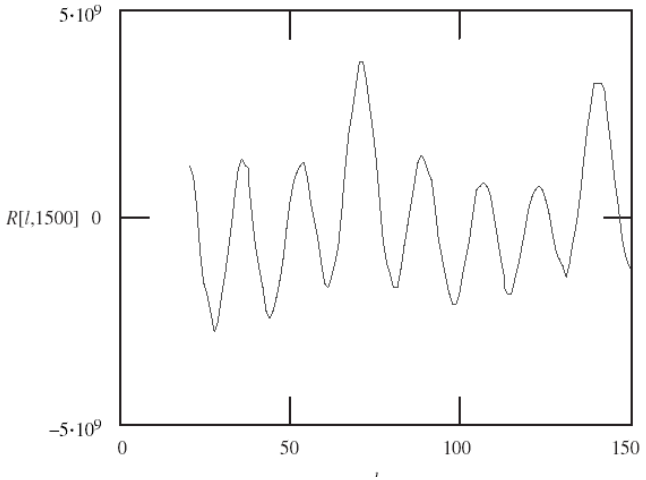
```
PITCH(m, N)
1.  peak  ← 0
2.  for l ← 20 to 150
3.          autoc ← 0
4.          for n ← m-N+ 1 to m
5.                  autoc ← autoc + s[n] s[n- l]
6.  if  autoc > peak
7.          peak ← autoc
8.          lag ← l
9.  return lag
```

It is important to mention that, in practice, the speech signal is often lowpass filtered before being used as input for pitch period estimation. Since the fundamental frequency associated with voicing is located in the low-frequency region (<500Hz), lowpass filtering eliminates the interfering high-frequency components as well as out-of-band noise, leading to a more accurate estimate.

The Autocorrelation Method: Example

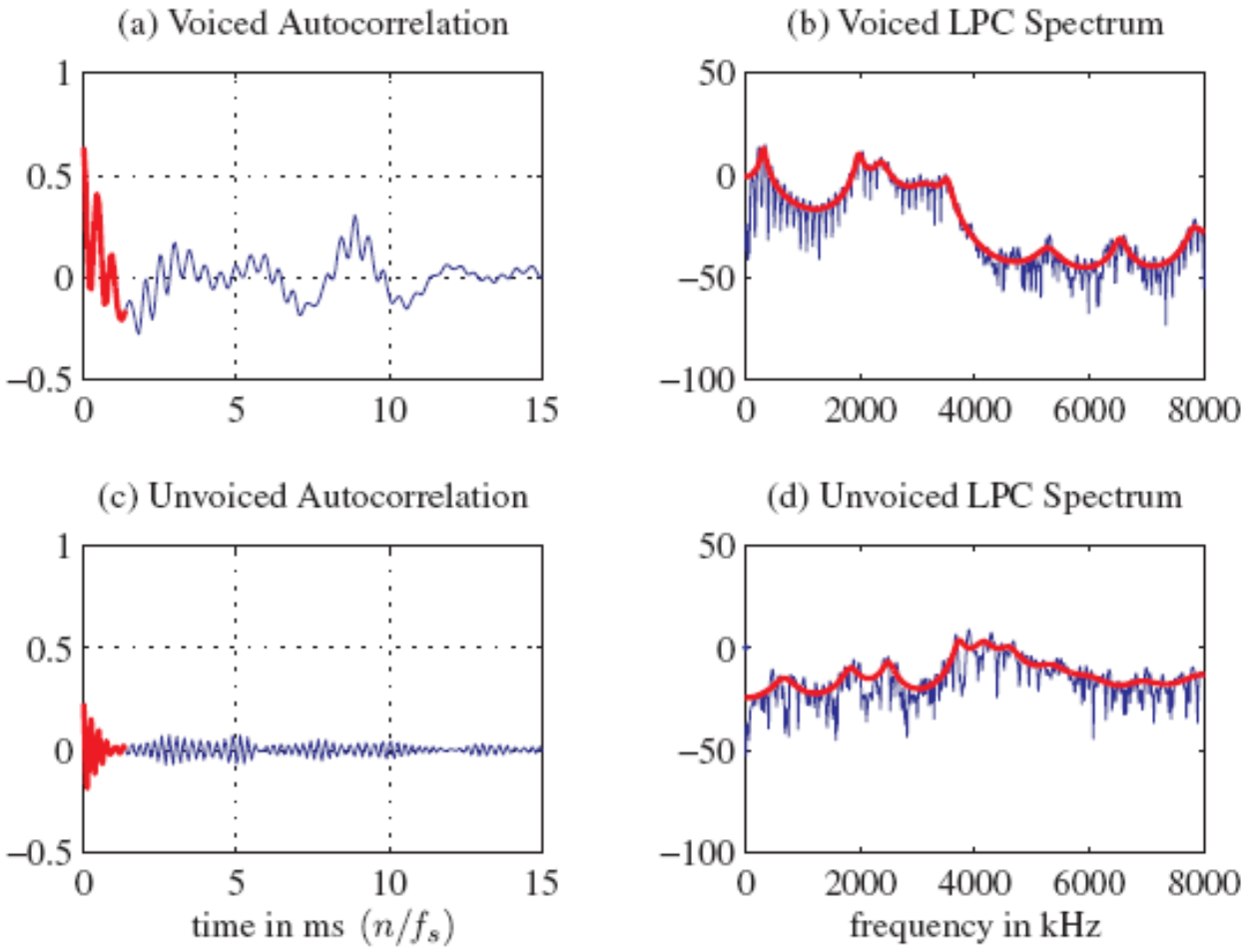
- Computing the autocorrelation of the given voiced speech sample, for l=20 to 150 gives the following plot:





Two strong peaks are obtained together with minor peaks.
The lag corresponding to the highest peak is 71 and is the pitch period estimate (for m =1500 and N =180)
This estimate is close to the period of the signal in time domain.
The next strong peak is located at a lag of 140, roughly doubling our pitch period estimate. This is expected since a periodic waveform with a period of T
is also periodic with a period of 2T, 3T, . . .

A Real Example of Autocorrelation & LPC



From: Rabiner & Schafer, 2007

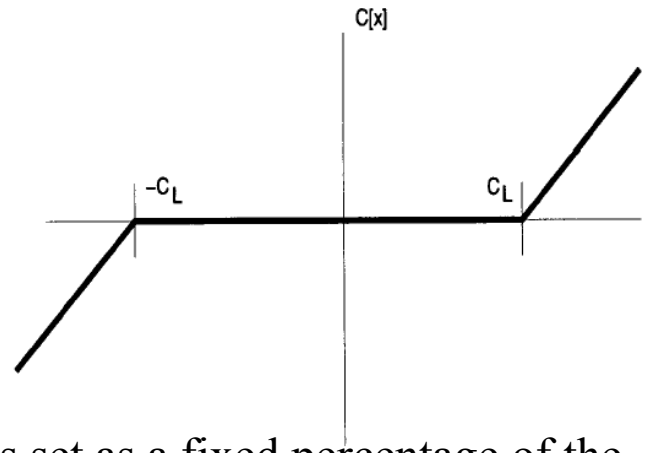
Pitch Estimation II: Autocorrelation of Center-Clipped Speech

- Since speech is not a purely periodic signal and vocal tract resonances produce additional maxima in the autocorrelation, pitch analysis on a direct autocorrelation of the speech signal can result in multiple local maxima.

- The method of center clipping the speech before computing the autocorrelation is one of the methods developed to suppress this local maxima phenomena.

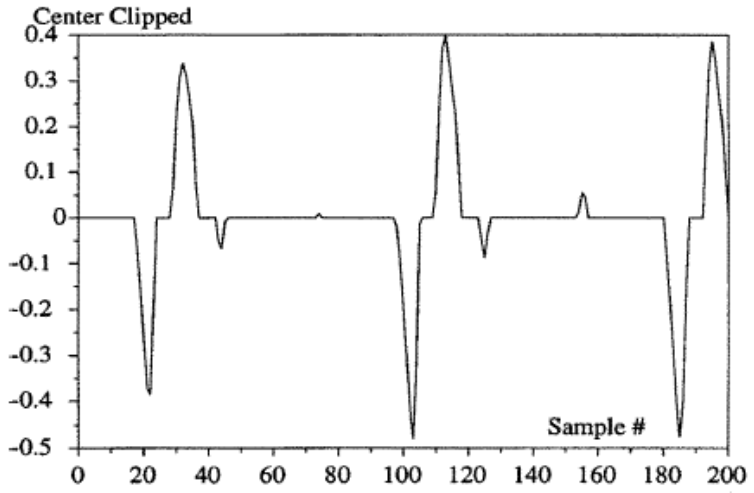
Center-Clipped Autocorrelation Cont'd

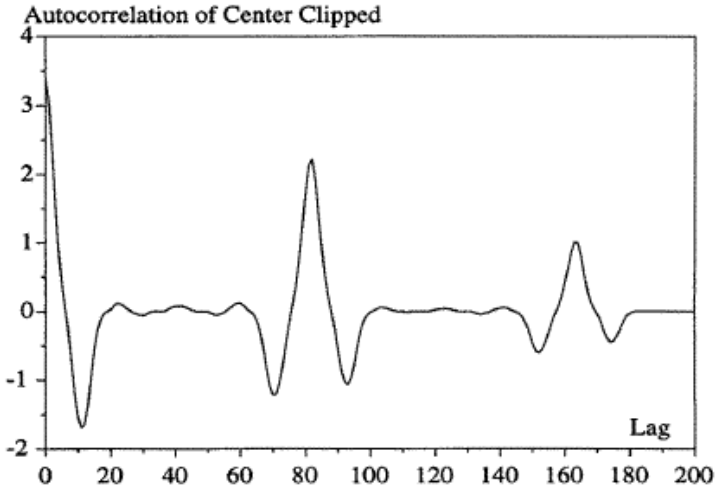
- The center-clipped speech is obtained by the nonlinear transformation: $y(n)=C[s(n)]$
- For samples with amplitude above C_L, the output of the center clipper is equal to the input minus the clipping level.
- For samples with magnitude below the clipping level, the output is zero.



C_L is set as a fixed percentage of the maximum amplitude of the speech signal, typically: 30%

Center-Clipped Autocorrelation Cont'd





center-clipped speech segment and the autocorrelation function of the clipped waveform.

The autocorrelation shows that the peak corresponding to pitch period is prominent, while the other local maxima have been reduced.

The peak of the autocorrelation of the center-clipped speech is much more distinguishable than in the autocorrelation of the original speech.

Note: If the signal is noisy or only mildly periodic (e.g. transition), the clipping operation might remove beneficial signal information.

For segments of rapidly changing energy, setting an appropriate clipping level can be difficult, even if it is adjusted dynamically.

Pitch Estimation III:
The Magnitude Difference Function

- The magnitude difference function (MDF) is defined by:

$$MDF[l,m] = \sum_{n=m-N+1}^{m} \left| S[n] - S[n-l] \right|$$

• For short segments of voiced speech it is reasonable to expect that s[n]-s[n-l] is small for l =0,±T,±2T, . . , with T being the signal's period.

• By computing the MDF for the lag range of interest, we can estimate the period by locating the lag value associated with the minimum magnitude difference.

• Note: **no products are needed for the implementation !**

The MDF Cont'd

Note that the MDF is bounded.
This fact is derived from its definition, where $MDF[l, m] \geq 0$.

From the same equation, each additional accumulation of term causes the result to be greater than or equal to the previous sum since each term is positive.
Thus, it is not necessary to calculate the sum entirely: if the accumulated result at any instance during the iteration loop is greater than the minimum found so far, calculation stops and resumes with the next lag.
The idea is implemented with the following pseudocode.

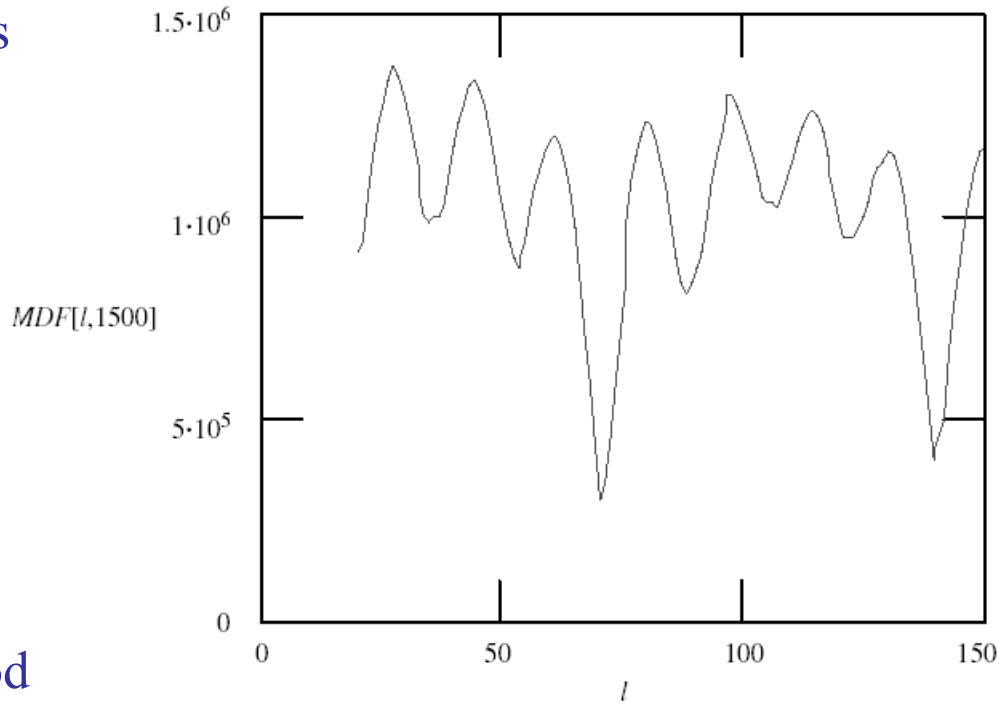
```
PITCH_MD1(m, N)
1.  min ← ∞
2.  for l ← 20 to 150
3.      mdf ← 0
4.      for n ← m-N+1  to  m
5.          mdf ←  mdf + abs(s[n]-s[n-l])
6.          if mdf ≥ min  break
7.      if mdf < min
8.          min ← mdf
9.          lag ← l
10. return lag
```

The MDF: Example

The same situation as in the previous Example is considered, where magnitude difference is computed for l = 20 to 150 .

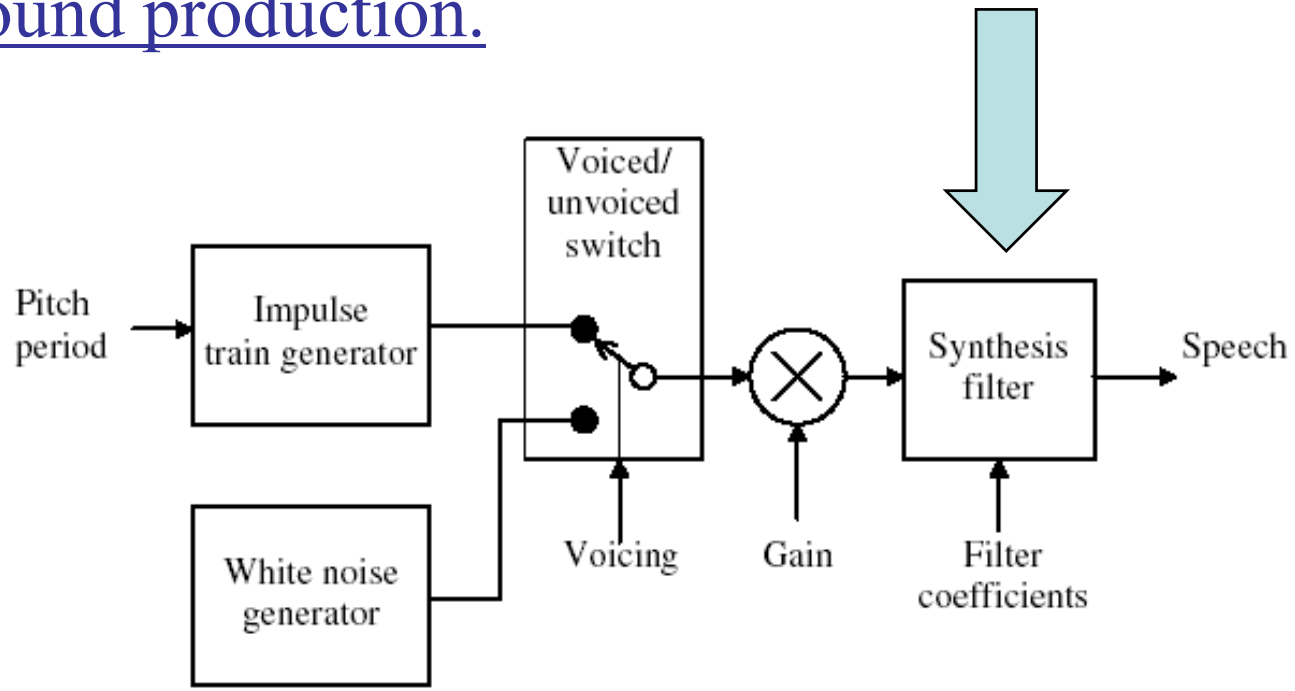
Lowest MDF occurs at l =70 with the next lowest MDF point located at l = 139.

Compared with the results of the previous example, the present method yields a slightly lower estimate.



Linear Prediction
Vocal Tract Modeling

- Linear Prediction (LP) is a widely used and successful method that represents the frequency shaping attributes of the vocal tract in the source-filter model of human sound production.



What is Linear Prediction ?

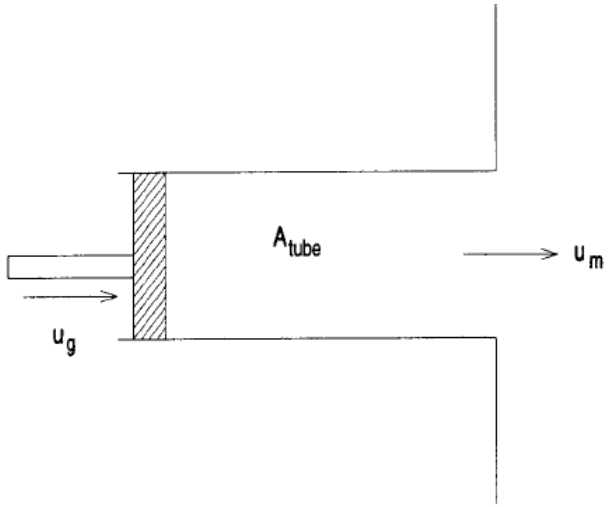
- Linear prediction, also frequently referred to as Linear Predictive Coding (LPC), predicts a time-domain speech sample based on a linearly weighted combination of previous samples.
- LP analysis can be viewed simply as a method to remove the redundancy in the short-term correlation of adjacent samples.

- However, additional insight can be gained by presenting the LP formulation in the context of lossless tube modeling of the vocal tract.

Sound Propagation in the Vocal Tract

- Sound waves are pressure variations that propagate through air by the vibrations of the air particles.
- Modeling the vocal tract as a uniform lossless tube with constant cross-sectional area is a simple but useful way to understand speech production.

Ug and *Um* represent the volume velocity flow at the glottis and mouth, respectively;
Atube is the constant cross-sectional area of the tube.



The Tube Model

- A system of partial differential equations describes the changes in pressure and volume velocity over time and position along the tube.
- wave equations characterize this system as:
  - Assuming ideal conditions:
    - no losses due to viscosity or thermal conduction
    - no variations in air pressure at the open end of the tube

x:	Location inside the tube
t:	Time
p(x,t):	Sound pressure at location *x* and time *t*
u(x,t):	Volume velocity flow at location *x* and time *t*
ρ :	Density of air inside the tube
c:	Velocity of sound
A9x,t):	Cross-sectional area of the tube at location *x* and time *t*

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial (u/A)}{\partial t}$$

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho c^2}\frac{\partial (pA)}{\partial t} + \frac{\partial A}{\partial t}$$

Single Tube Solution

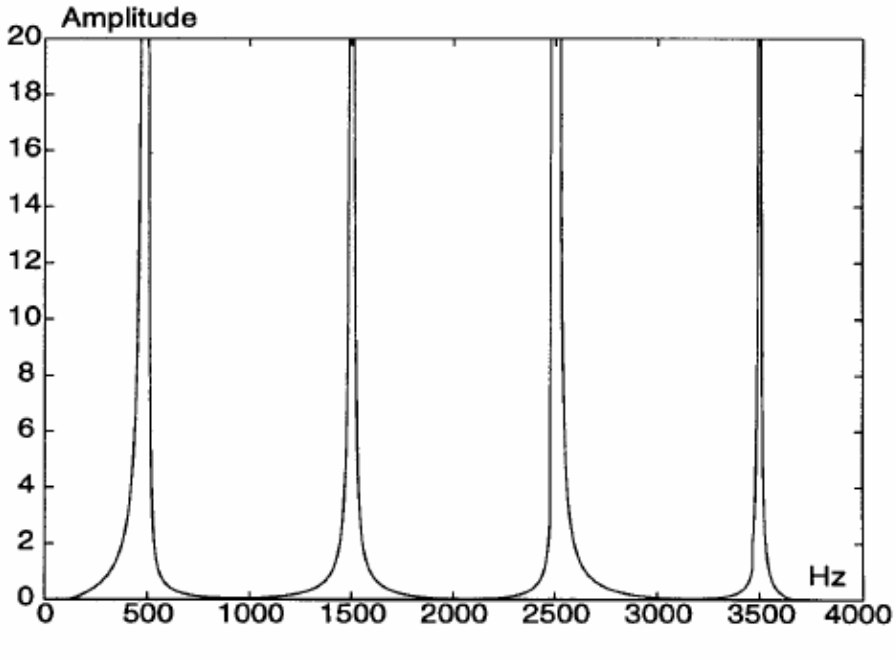
- The frequency response of the lossless tube system is not dependent on the source, just as the impulse response of an electrical system is not dependent on its input.
- The resonant frequencies of the vocal tract are called *formant frequencies*.
  - If the tube is 17.5 cm long, and 35,000cm/sec is used as *c* (the speed of sound), then the equally spaced formant frequencies of this system are:

$$\frac{35,000cm/\sec}{4(17.5cm)} \pm n \times \frac{35,000cm/\sec}{2(17.5cm)} = 500Hz \pm n \times 1000Hz$$

Single Tube Formants

The system has an infinite number of poles on the $j\omega$ axis corresponding to the tube resonant frequencies of

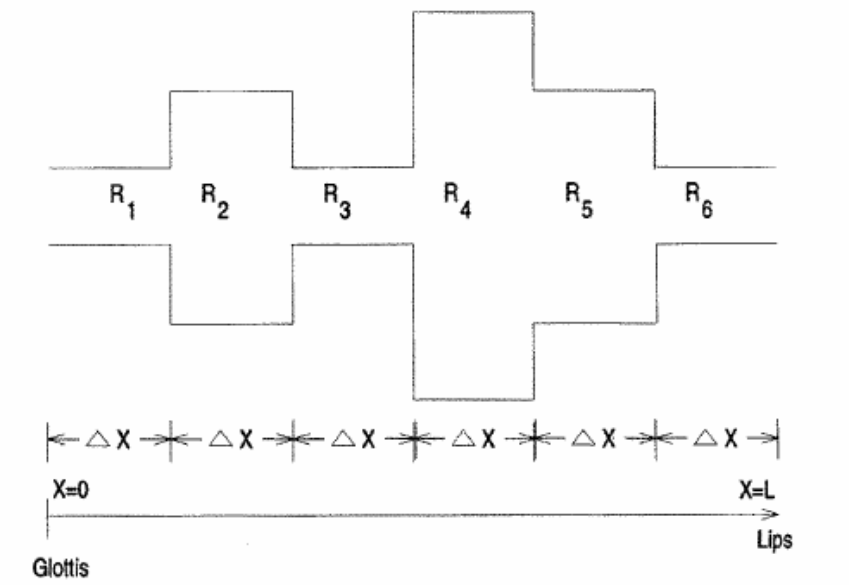
$$\frac{c}{4l} \pm \frac{nc}{2l} \qquad n = 0,1,\ldots,\infty$$



In an actual vocal tract, which is not uniform in area and is not lossless, formant frequencies are generally not as evenly spaced. A human vocal system also changes over time as the person articulates sounds. Therefore, the formant frequencies also change over time.

Multiple Tube Model

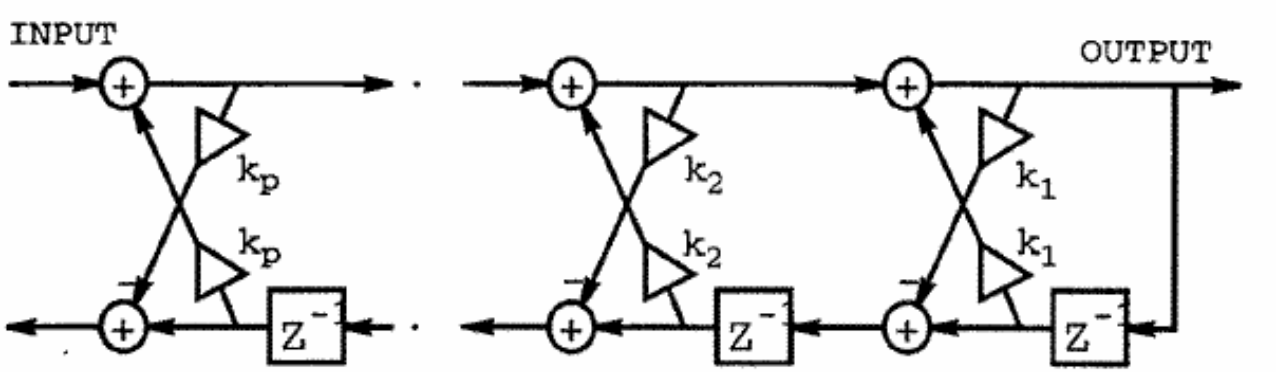
- In a physical vocal tract, the cross-sectional area varies based on position along the tract and over time.
- These variations create different speech sounds with the same excitation.
- To better model the varying cross-sectional area of the vocal tract, the single lossless tube can be extended to many lossless tubes concatenated to one another:



The vocal tract is excited at $x = 0$, which is either at the glottis or at some constriction in the vocal tract. The excitation propagates through the series of tubes with some of the energy being reflected at each junction and some energy being propagated.

Reflection Coefficients

- *The reflection coefficients* signify how much energy is reflected and how much is passed.
- These reflections cause spectral shaping of the excitation which acts as a digital filter with the order of the system equal to the number of tube boundaries:

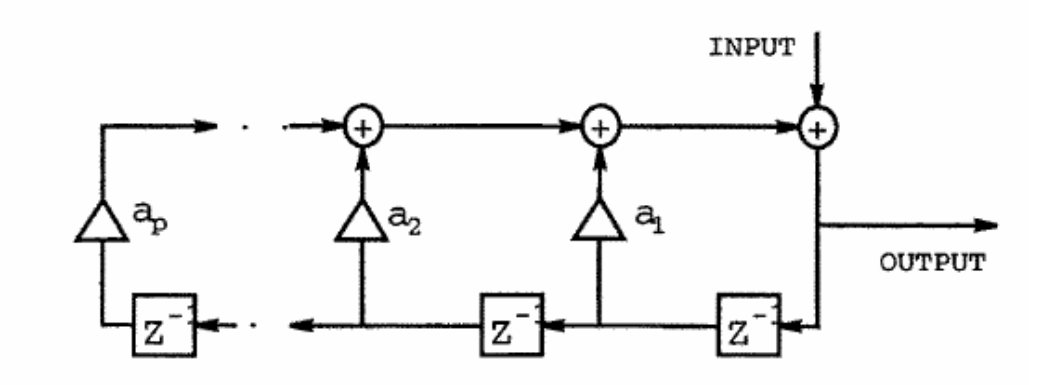


Reflection Coefficients Realization

- The digital filter can be realized with a lattice structure, where the reflection coefficients are used as weights in the structure.
- The k_i is the reflection coefficient of the i^{th} stage of the filter.
- The input is the excitation, and the output is the filtered excitation, that is, the output speech.

The Direct Form Realization

The lattice structure can be rearranged into the direct form of the standard all-pole filter model:



- Each tap, or *predictor coefficient*, of the digital filter delays the signal by a single time unit and propagates a portion of the sample value.
- There is a direct conversion between the reflection coefficients, k_i and predictor coefficients, a_i and they represent the same information in the LP analysis

Linear Prediction Analysis

- From either the direct-form filter realization or the mathematical derivation of lossless tube model, linear prediction analysis is based on the all-pole filter:

$$H(z) = \frac{1}{A(z)} \qquad A(z) = 1 - \sum_{k=1}^{p} a_k Z^{-k}$$

- where $\{ak, 1 \le k \le p\}$ are the predictor coefficients, and p is the order of the filter.

Time Domain Representation

- By transforming to the time domain, it can be seen that the system predicts a speech sample based on a sum of weighted past samples:

$$S'(n) = \sum_{k=1}^{k} a_k S(n-k)$$

- where $s'(n)$ is the predicted value based on the previous values of the speech signal $s(n)$.

Estimation of LP Parameters

- To utilize the LP model for speech analysis, it is necessary to estimate the LP parameters for a segment of speech.
- The idea is to find the a_k's that provides the closest approximation to the speech samples, so that:

$s'(n)$ is closest to $s(n)$ for all the values of n

- For this discussion, the spectral shape of $s(n)$ is assumed to be stationary across the frame
  - frame = a short segment of speech.

The Prediction Error

- The error between a predicted value and the actual value is: $e(n) = s(n) - s'(n)$

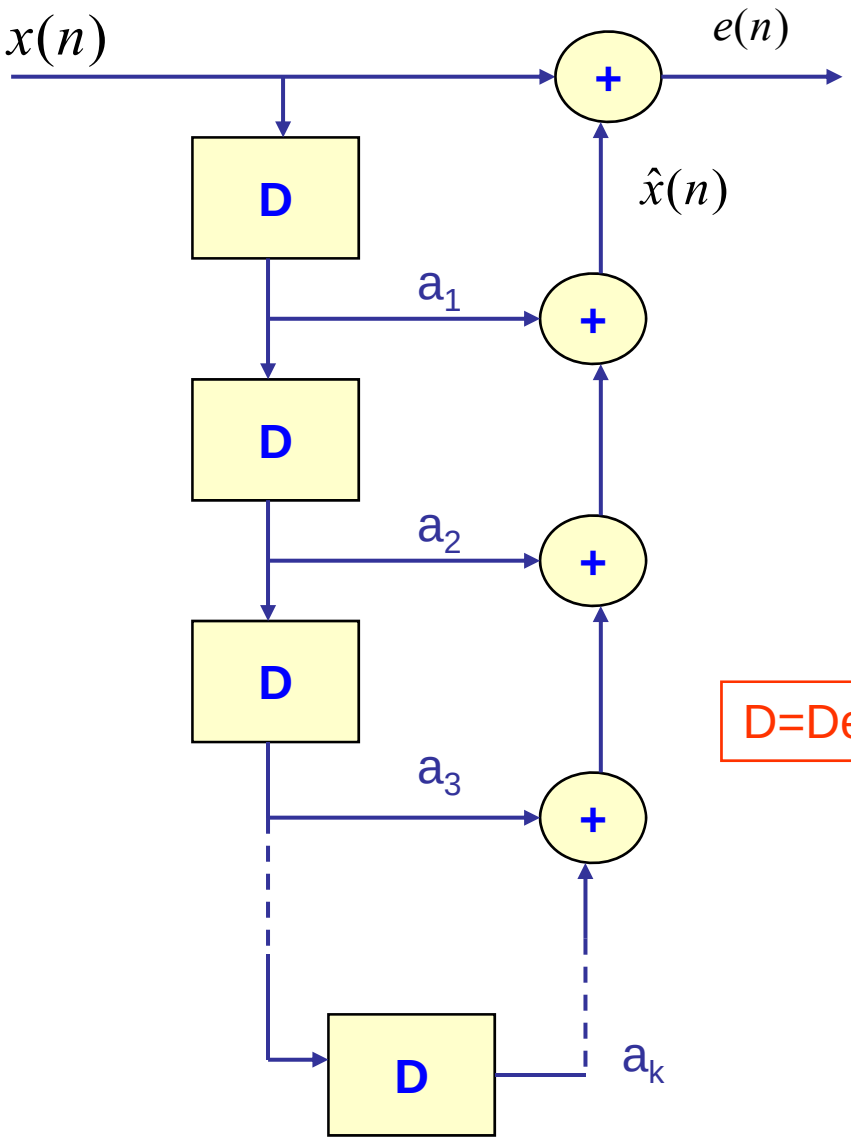
$$e(n) = s(n) - \sum_{n=1}^{k} a_k s(n-k)$$

- The values of a_k can be computed by minimizing the total squared error E over the segment:

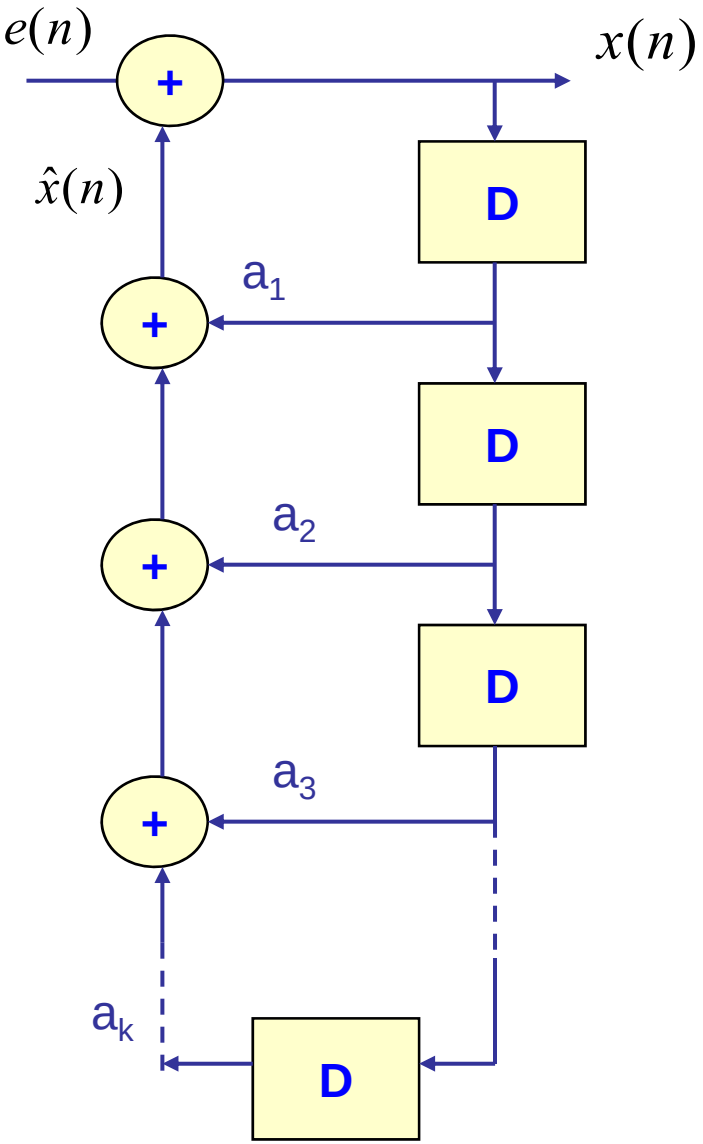
$$E = \sum_{n} e^2(n)$$

By setting the partial derivatives of E with respect to the a_k's **to zero**, a set of equations results that minimizes the error.

Implementation



D=Delay Unit



So….what's left to do ?

Autocorrelation Method of Parameter Estimation

- the speech segment is assumed to be zero outside the predetermined boundaries.
- The range of summation is $0 \le n \le N + p - 1$.
- The equations for the a_k's are compactly expressed in matrix form as:

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(p-1) \\ r(1) & r(2) & \cdots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{bmatrix}$$

- where $r(l)$ is the autocorrelation of lag l computed as:

$$r(l) = \sum_{m=0}^{N-1-l} s(m)s(m+l)$$

N: The length of the speech segment *s*(*n*).

Levinson-Durbin Solution

- Because of the Toeplitz structure (symmetric, diagonals contain same element) of the matrix, the efficient Levinson-Durbin recursion can be used to solve the system.
  - The equations are:

$$[1]\ E^{(0)} = r(0)$$

$$[2]\ k_i = \frac{r(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} r(i-j)}{E^{(i-1)}}$$

$$[3]\ a_i^{(i)} = k_i$$

$$[4]\ a_j(i) = a_j^{(i-1)} - k_i a_{i-j}^{(i)}$$

$$[5]\ E^{(i)} = (1 - k_i^2) E^{(i-1)}$$

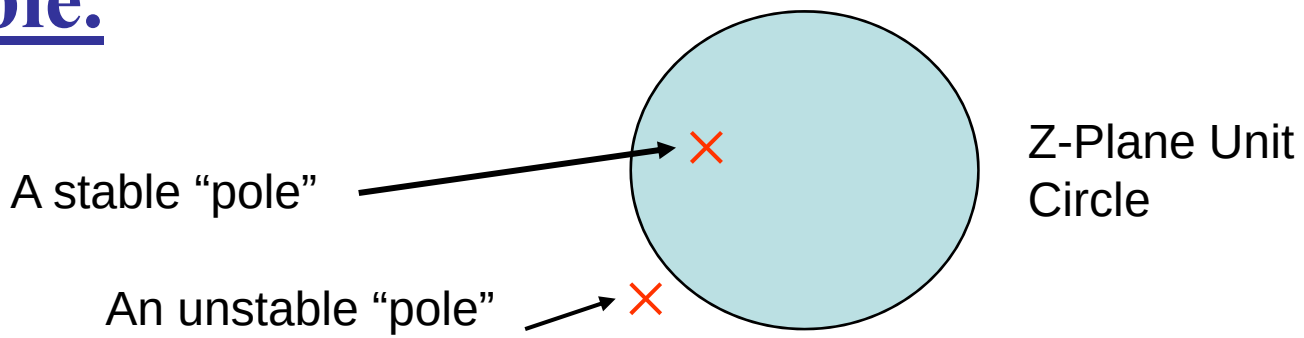
- Where $1 \le j \le i-1$.
- In all equations, i is the current order in the recursion, and the equations are solved in turn for all orders of $i = 1, 2, \cdots, p$.

Levinson-Durbin Solution Cont'd

- The i^{th} order coefficient of Eq. [3] for values

$1 \leq i \leq p$ is the i^{th} reflection coefficient as discussed before.

- $k_i < 1$ *for* $1 \leq i \leq p$ is met, the roots of the predictor polynomial will all lie within the *unit circle in the z-plane*, and the all-pole filter will be **stable.**



The Covariance Method

- In the covariance method, the range of the summation of : $E = \sum_n e^2(n)$ is limited to the range of the indices in the speech segment.
- This formulation results in the solution of the error minimization as:

$$\begin{bmatrix} c(1,1) & c(1,2) & \cdots & c(1,p) \\ c(2,1) & c(2,2) & \cdots & c(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ c(p,1) & c(p,2) & \cdots & c(p,p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} c(1,0) \\ c(2,0) \\ \vdots \\ c(p,0) \end{bmatrix}$$

- where the covariance c is:

$$C(i,k) = \sum_{m=0}^{N-1} s(m-i)s(m-k)$$

- and includes values of $s(n)$ outside the original segment range of $0 \le n \le N\text{-}1$

The Covariance Method Cont'd

- Although the form for the covariance method is not Toeplitz, and does not allow the Levinson-Durbin recursion solution, efficient methods such as the Cholesky decomposition can be used to solve the system of equations.

More information can be found in the following reference:
J. Makhoul. Linear Prediction: A tutorial review. *Proc. IEEE,* 1975, pp. 561-580.