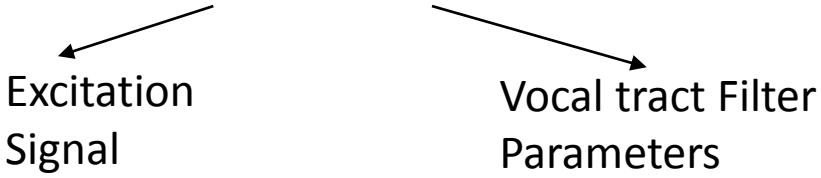
ECE462 – Lecture 23

Speech Compression

Use a speech model of speech synthesis

At transmitter: Analyze speech signal to determine

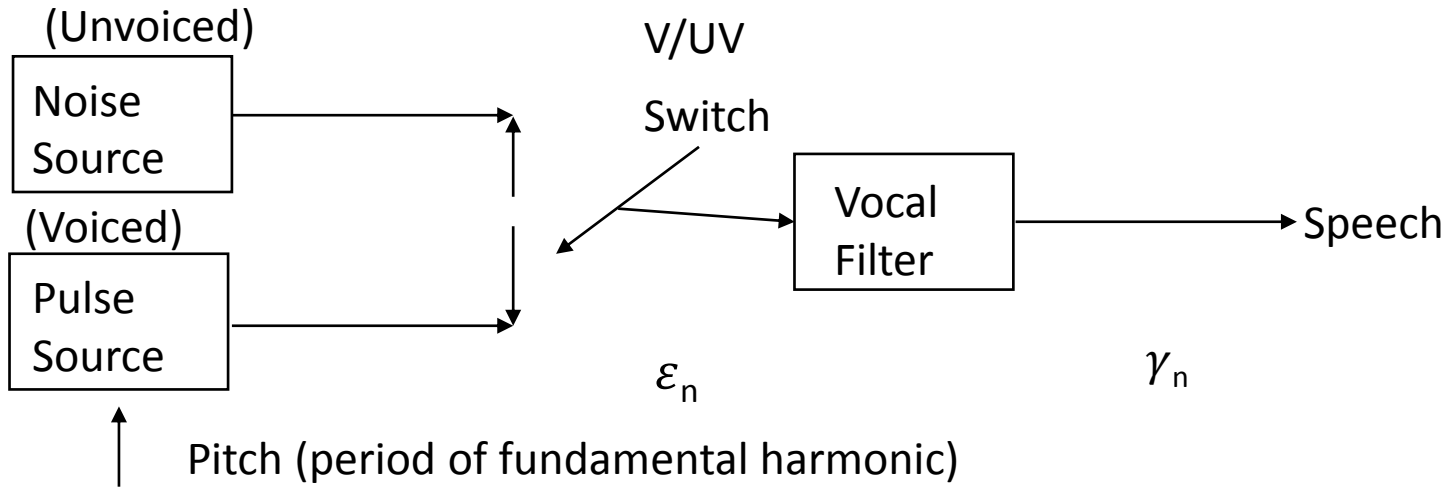


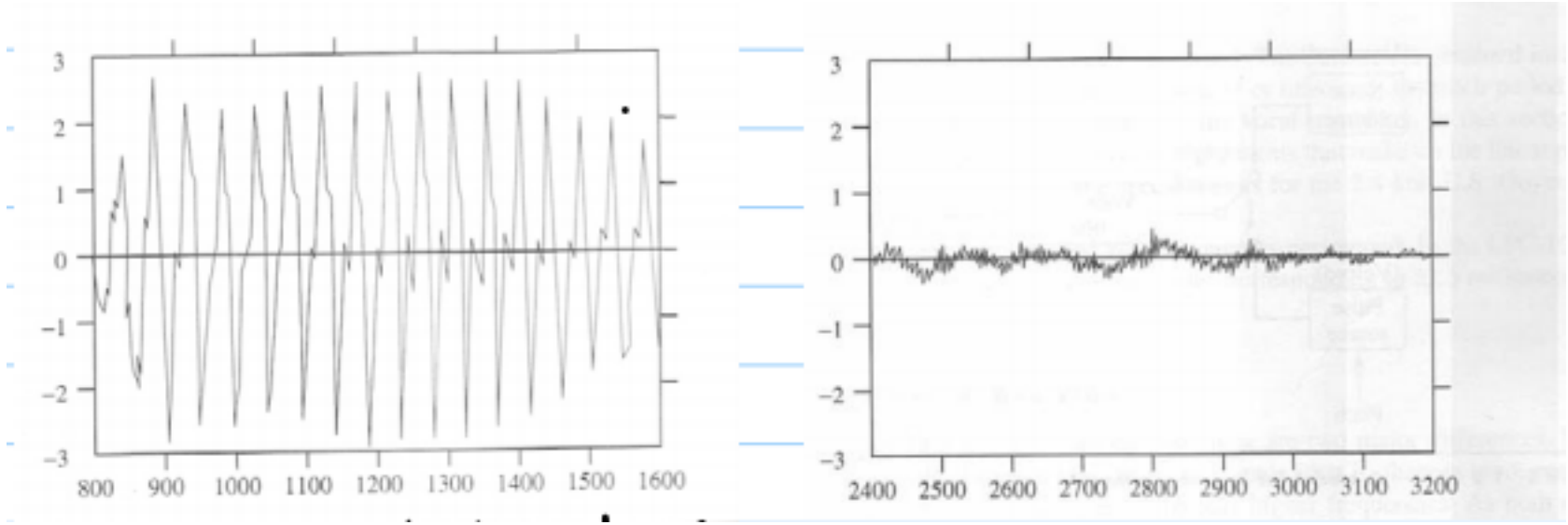
At receiver: Synthesize signal based on speech model

- Channel recorders/Linear predictive coders
- Code excited linear prediction (CELP)

Linear Predictive Coding

A model for speech synthesis





The sound |e| in test (pseudo periodic)

The sound |s| in test (noise like)

$$Y_n = \sum_{i=1}^{m} b_i y_{n-i} + G\varepsilon_n$$

G: gain,$\{b_i\}$ parameters of vocal filter

LPC-10 Standard: 2.4 Kbs

- Input signal is sampled at 8,000 samples/sec
- Signal is broken into segments of 180 samples/segment

  (22.5 msec of speech / segment)

Voiced vs Unvoiced decision

- Voiced Speech contains more energy (higher signal values)
- Unvoiced speech contains higher frequency values

(more zero crossings compared to voiced speech: note that both types of signal are approximately zero mean)

In LPC-10 signal is filtered (LP) with a bandwidth of 1KHZ and then the energy and zero crossings are calculated to decide between voiced - unvoiced

Pitch period estimation

- Computationally intensive process
- Human pitch period is usually between 2.5 to 19.5 msecs

1. Autocorrelation Approach: Given a voiced speech segment

X(n), n= n_0,….,n_0+N

Autocorrelation function: $R_{xx}(k)=\sum_{n_0}^{n_0+N} x(n)x(n-k)$

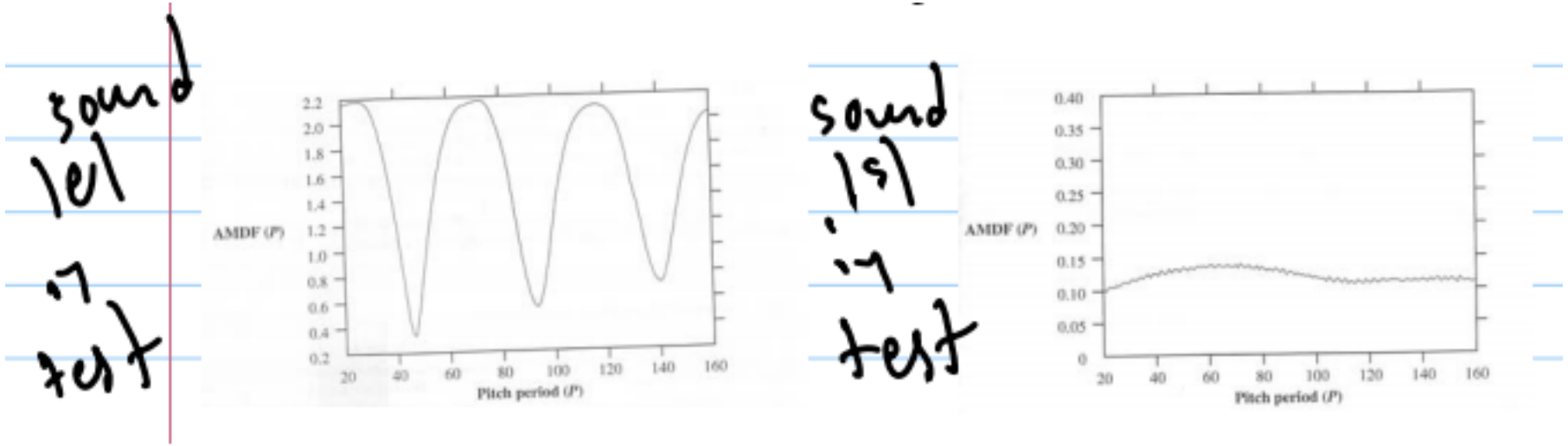
$k \geq 0$

$R_{xx}(k)$ will have a maximum when k is equal to the pitch period

Pitch period estimation

2. Average magnitude difference Approach (AMDF)

$$\mathrm{AMDF(P)} = \frac{1}{N}\sum_{i=n_0+1}^{n_0+N} |x_i - x(i-p)|$$



- Calculating the parameters of the Vocal Filter

According to the model $y_n = \sum_{i=1}^{m} b_i y_{n-i} + G\varepsilon_n$

This implies that y_n is estimated (or predicted) by m past values. In practice of course does not hold exactly, so to estimate the parameters $\{b_i\}$, i=1,…,m and the gain G we minimize the average error

$$e_n^2 = \left(y_n - \sum_{i=1}^{m} b_i y_{n-i} - G\varepsilon_n \right)^2$$

That is we minimize $E\{e_n^2\}$ w.r.t. b_j

Taking the derivative of $E\{e_n^2\}$ w.r.t. b_j

$$\frac{\partial E\{en^2\}}{\partial b_j} = \frac{\partial}{\partial b_j} E\{(y_n - \sum_{i=1}^{m} b_i y_{n-i} - G\varepsilon_n)\}=0$$

$$\Rightarrow -2E[y_n - \sum_{i=1}^{m} b_i y_{n-i} - G\varepsilon_n / y_{n-j}] = 0$$

$\Rightarrow \sum_{i=1}^{m} b_i E\{y_{n-i}\, y_{n-j}\} = E\{y_n\, y_{n-j}\}$ (2)

$j=1,2,\ldots\ldots,m$

because $E\{\epsilon_n\, y_{n-j}\} = 0 \quad j \neq 0$

Note that assuming the speech signal is "stationary" that is its statistics do not change with time, then

$$E\{y_{n-i}\, y_{n-j}\} = R_{yy}(|i-j|)$$

and $$E\{y_n\, y_{n-j}\} = R_{yy}(|j|)$$

where once again the autocorrelation function is estimated as

$$\hat{R}_{yy}(k) = \sum_{n_0+l+k}^{n_0+N} y_n\, y_{n-k}$$

Then from (2) above the following linear system of equations is obtained

$$\Rightarrow \begin{pmatrix} R_{yy}(0) & R_{yy}(1) \cdots & R_{yy}(m-1) \\ R_{yy}(1) & R_{yy}(0) & R_{yy}(m-2) \\ \vdots & \ddots & \vdots \\ R_{yy}(m-1) & R_{yy}(m-2) \ldots & R_{yy}(0) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = \begin{pmatrix} R_{yy}(1) \\ R_{yy}(2) \\ \vdots \\ R_{yy}(m) \end{pmatrix}$$

R $\quad$ $\underline{b} = \underline{r}$

And $b = R^{-1}.r$

Once $\{b_i\}_{i=1,\cdots,m}$ have been obtained from 1 with j=0 we obtain

$$\Rightarrow E\{\ y_n^2\ \} - \sum_{i=1}^{M} b_i E\{y_{n-i}.y_n\} = GE\{\epsilon_n y_n\} = G$$

$$\boxed{G = R(0) - \sum_{i=1} b_i.R(i)}$$

b_i by definition

- Note: The matrix R has a nice structure. It is Toeptitz: (All diagonal elements and those parallel to the diagonal are equal -> Efficient solution of system exists (Levinson – Durbin Algorithm)

- Levinson – Durbin Algorithm
  1. Initialize $E_0 = \widehat{R_{yy}}(0), i = 0$
  2. Increment i by 1
  3. Calculate $k_i = \sum_{j=1}^{i-1} b_j^{(i-1)} R_{yy}(i-j+1) - R_{yy}(i)$
  4. Set $b_i^{(i)} = k_i, \widetilde{(i-1)\ E_{i-1}}$
  5. For $j = 1,2,\cdots,i-1$

  Calculate $b_j^{(i)} = b_j^{(i-1)} + k_i b_{i-j}$

  6. $E_i = \left(1 - k_i^2\right) E_{i-1}$
  7. If i<m go to step 2. Note: the k_i are known 1 as reflection coefficients

- If $|k_i| < 1 \rightarrow$ filter stable
- For effective reconstruction of voiced segment m> 10
- Given $k_i, i = 1, \cdots, m \leftrightarrow \{b_i\}, i = 1, \cdots, m$

Since stability is better controlled by k_i, quantize and transmit $\{k_i\}$.

<u>Parameters to be transmitted</u>

1. v/uv decision (1 bit)
2. Pitch period (quantized to 1 of 60 different values) 6 bits
3. Vocal track filter parameters

(m=10 for voice)

(m=5 for unvoiced)

Plus G (5 bit log compounded quant)

For k_1, k_2: $g_i = \frac{1+k_i}{1-k_i} \rightarrow$ 5 bit uniform quantized

k_3, k_4: 5 bit uniform quantized

k_5, k_8: 4 bit uniform quantized

k_9: 2 bits,

k_{10}: 3 bits

+ 1 bit for synchronization -> 54 bits/frame

- So for LPC-10 coder
- 54 bits/frame of 180 samples 8,000 samples/sec
- So bit rate:(8000/180)x54 = 2400 bps.

- This encoding approach generates low bit rate speech, with “artificial voice quality.”