Speech and Audio Coding Theory

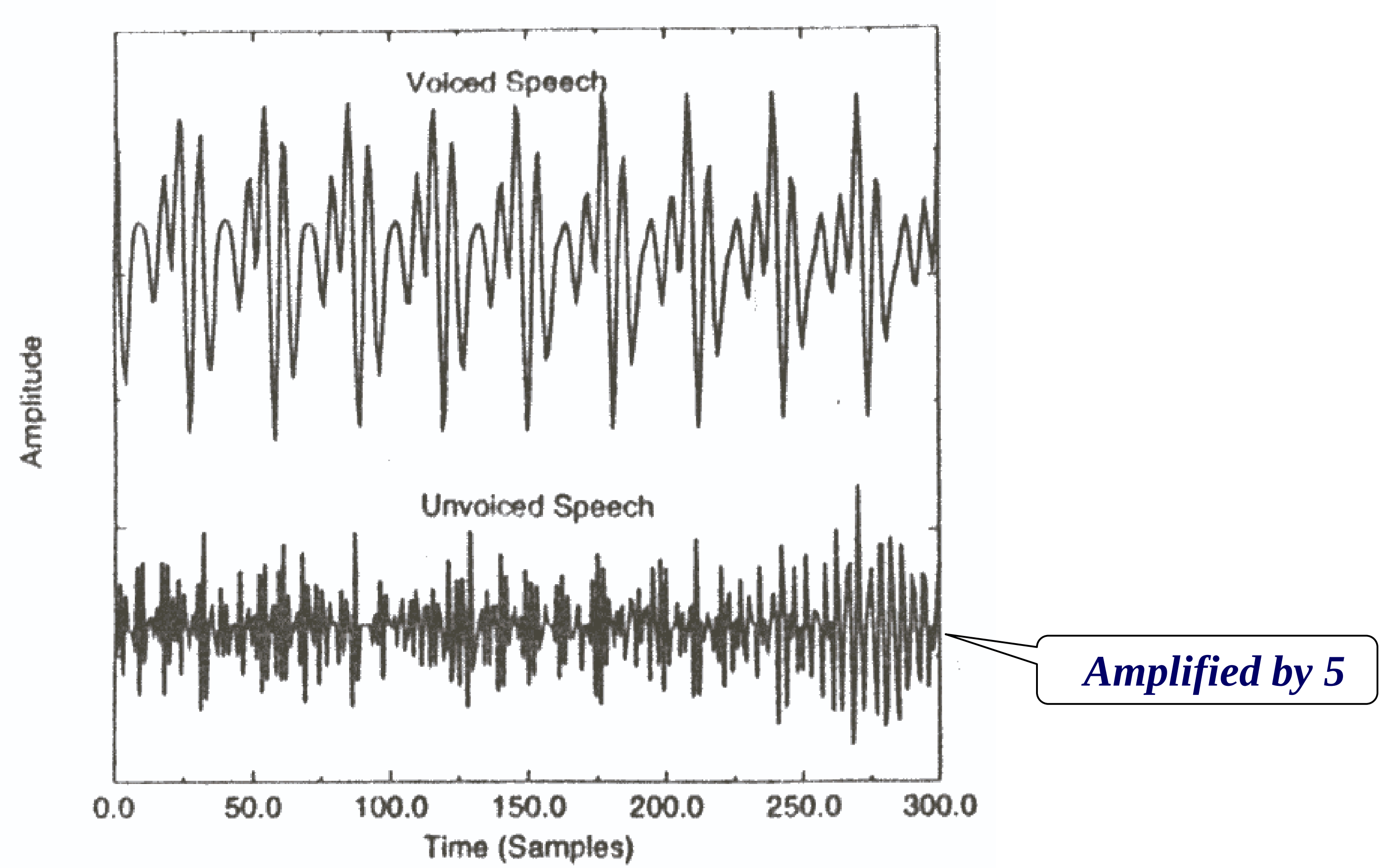
Contents of lecture

- General speech characteristics
- Frequency domain analysis: Short-time spectral analysis
- Time domain analysis: Linear predictive modeling
  - Source-filter model of speech production
  - Solutions to LPC analysis
    - Auto-correlation method (AM)

General speech characteristics

- Analysis of the speech signal not on the phoneme level (linguistic unit level), but on the general speech characteristics (physical waveform level)
  - Voiced signal: high energy, quasi-periodicity (due to pitch)
  - Unvoiced signal: relatively low energy, like random noise with no periodicity
  - Mixture of voiced and unvoiced signals: transition region (voiced-to-unvoiced or unvoiced-to-voiced region) or inherently mixed characteristics
  - Example of voiced and unvoiced speech signals (Next slide)

General speech characteristics



Short-time spectral analysis

- Frequency domain analysis of speech signal
  - Short-time Fourier transform
    - Time-dependent Fourier transform

$$S_k(e^{j\omega}) = \sum_{n=-\infty}^{\infty} w(k-n)s(n)e^{-j\omega n}$$

    - $w(k-n)$: Real window sequence to isolate the portion of the input signal

- Ideal window frequency response
  - Very narrow main lobe: to increase frequency resolution
  - No side lobe: for no frequency leakage
  - In practice, no ideal window

Window functions

- Rectangular window

$$w(n) = \begin{cases} 1 & ; \quad 0 \le n \le N-1 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

- Bartlett window

$$w(n) = \begin{cases} \frac{2n}{N-1} & ; \quad 0 \le n \le \frac{N-1}{2} \\ 2 - \frac{2n}{N-1} & ; \quad \frac{N-1}{2} \le n \le N-1 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

- Hamming window

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi \frac{n}{N-1}) & ; \quad 0 \le n \le N-1 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

Window functions

- Hanning window

$$w(n)=\begin{cases}0.5-0.5\cos(2\pi\frac{n}{N-1}) & ;\quad 0\le n\le N-1\\ 0 & ;\quad \text{otherwise}\end{cases}$$

- Blackman window

$$w(n)=\begin{cases}0.42-0.5\cos(2\pi\frac{n}{N-1})+0.08\cos(2\pi\frac{n}{N-1}) & ;\quad 0\le n\le N-1\\ 0 & ;\quad \text{otherwise}\end{cases}$$

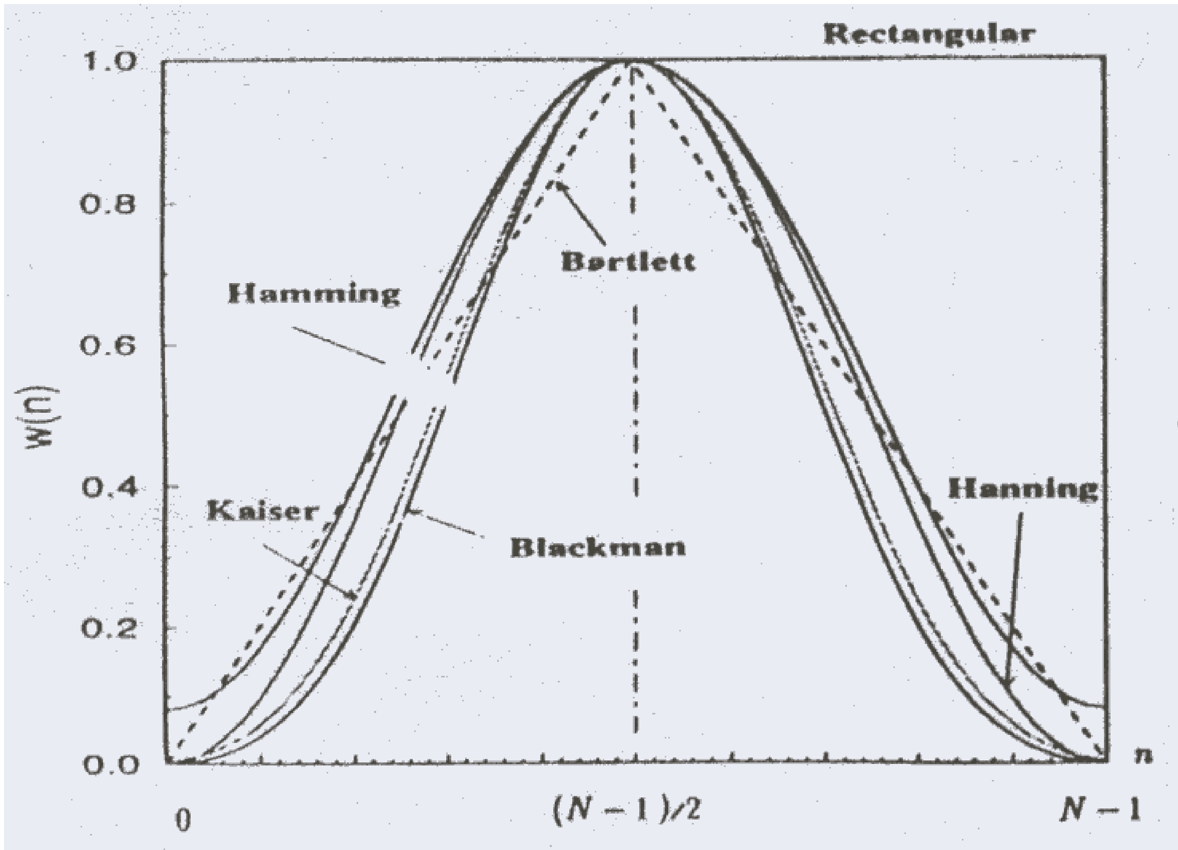
- Kaiser window

$$w(n)=\begin{cases}\dfrac{I_0(\beta\sqrt{1-(\frac{2n}{N-1}-1)^2})}{I_0(\beta)} & ;\quad 0\le n\le N-1\\ 0 & ;\quad \text{otherwise}\end{cases}$$

  - I_0 is zero-order Bessel function given by $I_0(\beta)=\sum_{k=0}^{\infty}\frac{\beta^{2k}/2}{(k!)^2}$

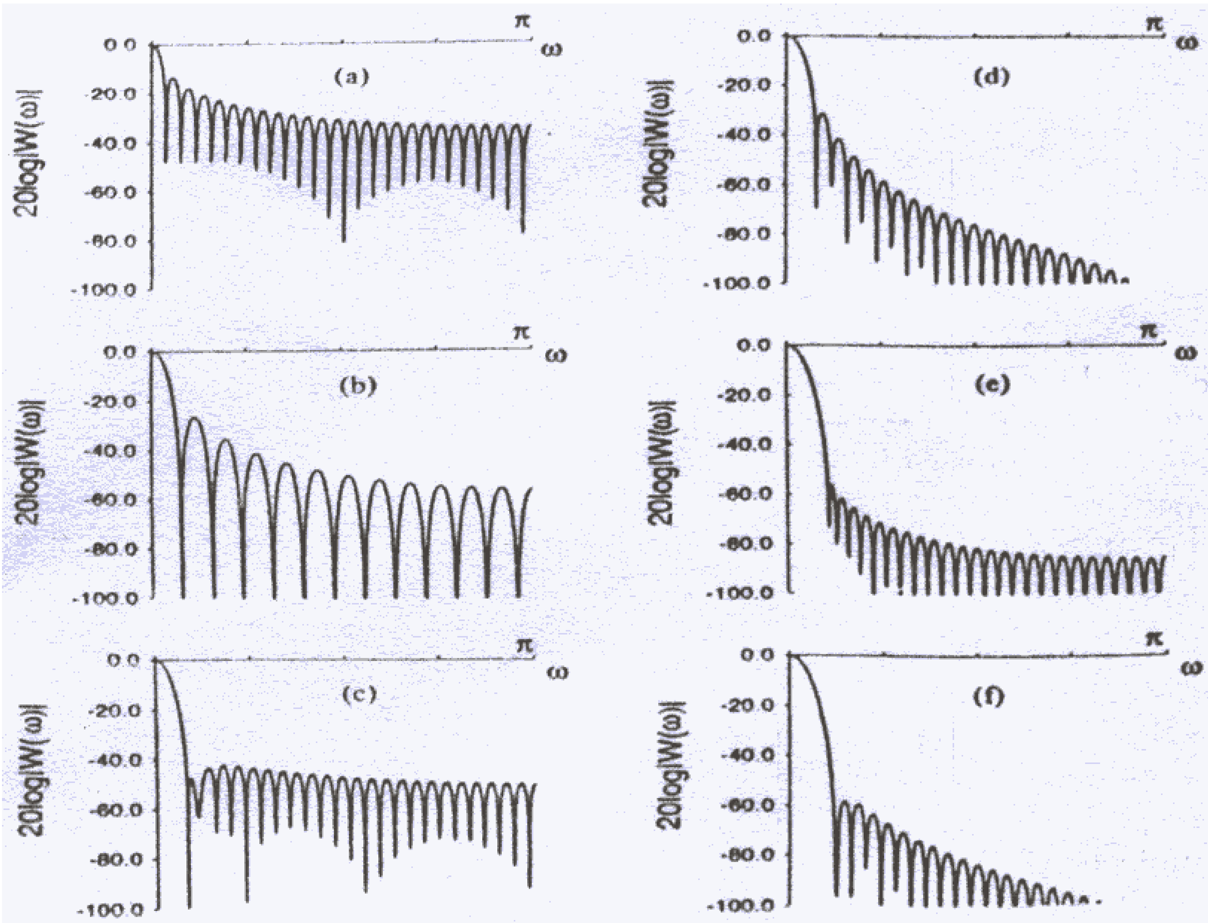
Window functions

- Time domain shapes for the windows



Window functions

- Frequency domain shapes for the windows



(a) Rectangular

(b) Bartlett

(c) Hamming

(d) Hanning

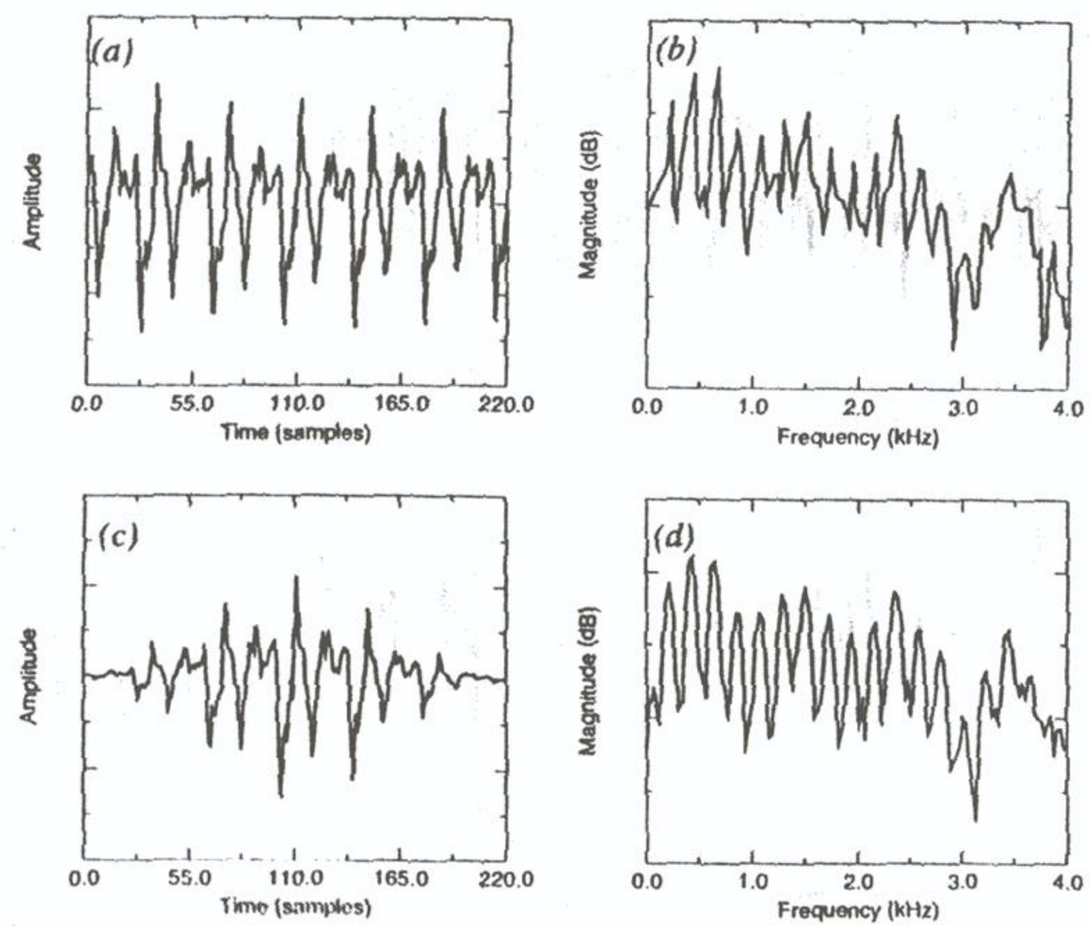
(e) Kaiser (β=7.8)

(f) Blackman

- Rectangular window: highest resolution, largest leakage
- Blackman window: lowest resolution, smallest leakage

Window functions

- Voiced signal for rectangular and Hamming (220 samples)



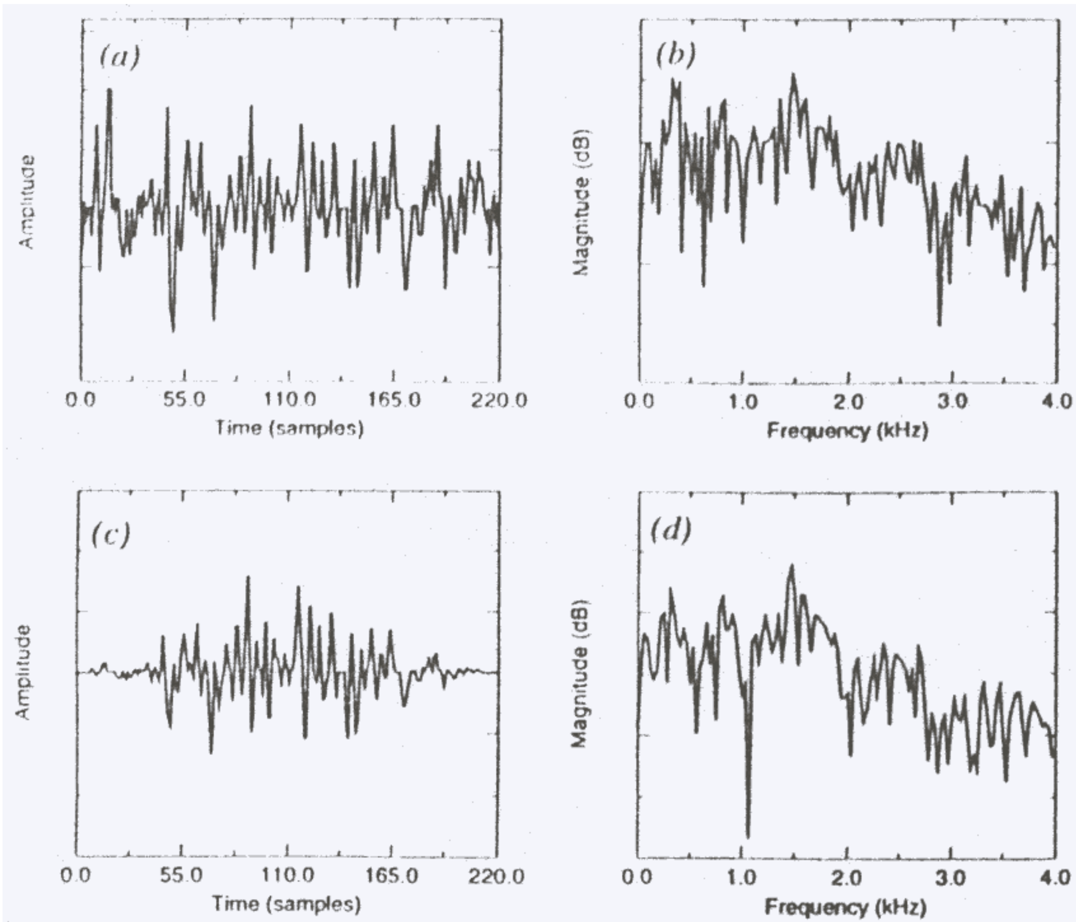
(a), (b): using rectangular

(c), (d): using Hamming

- Similarity: pitch harmonics, formant structure, gross spectral shape
- Rectangular window: shaper, but noise-like due to high leakage
- So, rectangular window is not used generally for STFT.

Window functions

- Unvoiced signal for rectangular and Hamming (220 samples)



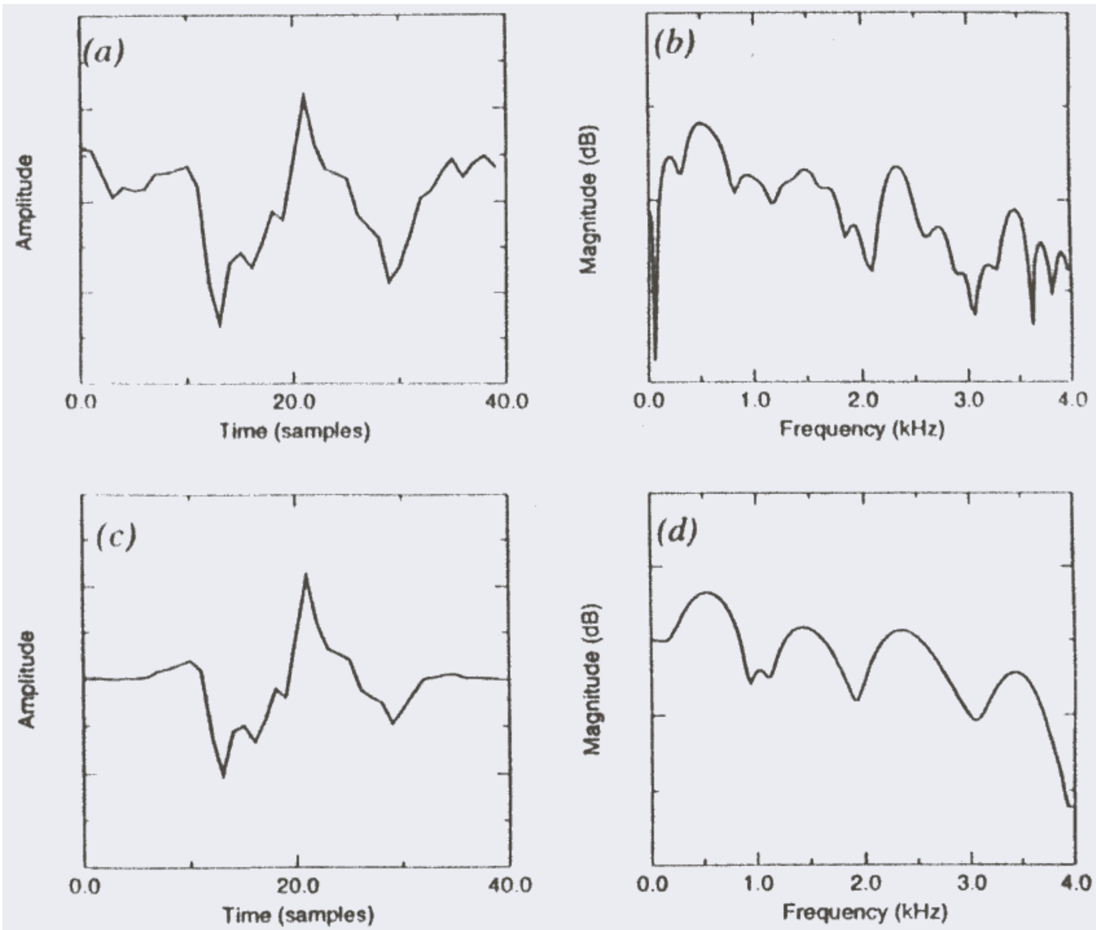
(a), (b): using rectangular

(c), (d): using Hamming

- Hamming is still smoother than rectangular.

Window functions

- Voiced signal for rectangular and Hamming (40 samples)



(a), (b): using rectangular

(c), (d): using Hamming

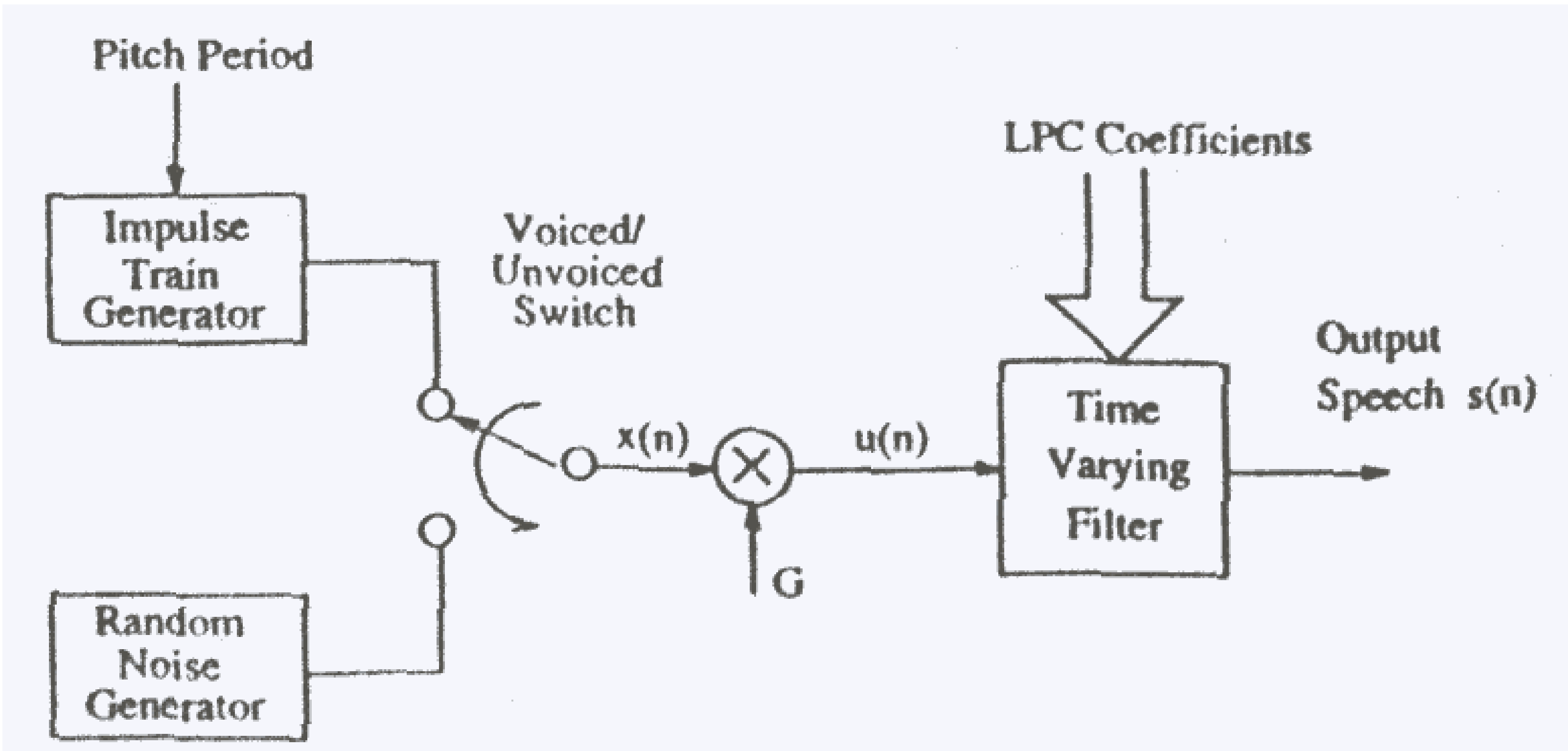
- Different spectrum according to window position
- Good temporal resolution with a short window
- Good frequency resolution with a longer window
- Trade-off between short and long windows → Therefore, it is reasonable to set a window size to 120-240 samples (i.e. 15-30msec duration).

Linear predictive modeling of speech signals

- Linear predictive coding (LPC) analysis
  - Very accurate representation of speech with a small set of parameters
  - Short-term correlations between speech samples
    - To capture the formant information
  - Long-term correlations between speech samples (Pitch prediction)
    - To capture the fundamental frequency (pitch period) information
- Difference between "term" and "time" in the literature
  - "term": sample interval to obtain the correlation
  - "time": analysis frame size to obtain the correlation

Source-filter model of speech production

- Block diagram of the simplified source-filter production model



Source-filter model of speech production

- Speech production modeling by time-varying digital filter
  - Glottal flow + vocal tract + lip radiation
  - $H(z) = \dfrac{S(z)}{X(z)} = \dfrac{G\left(1 - \sum_{j=1}^{M} b_j z^{-j}\right)}{1 - \sum_{i=1}^{N} a_i z^{-i}}$ ：Pole−zero modeling
- Approximation to all-pole model if *N* is large enough
  - $H(z) = \dfrac{G}{1 - \sum_{j=1}^{p} a_j z^{-j}} = \dfrac{G}{A(z)} = \dfrac{S(z)}{X(z)}$
  - Then, the difference equation becomes $s(n) = Gx(n) + \sum_{j=1}^{p} a_j s(n-j)$

Source-filter model of speech production

- Error or residual signal
  - If speech production model is really same as the above all-pole model, then we can decompose the given $s(n)$ to the excitation signal $x(n)$ and the filter coefficients a_j.
  - However, since the all-pole model is not exact, we may approximate the above difference equation to

$$e(n) = s(n) - \sum_{j=1}^{p} \alpha_j s(n-j)$$

    - $e(n)$: error (or residual) signal
    - α_j: the estimates of a_j

Source-filter model of speech production

- Determine α_j by minimizing the MSE
  - $MSE = E\{e^2(n)\} = E\left\{\left[s(n) - \sum_{j=1}^{p}\alpha_j s(n-j)\right]^2\right\}$
    - $E\{\ \}$ is ensemble average, not time average.
  - Using $\dfrac{\partial E}{\partial \alpha_i} = 0, \quad 1 \le i \le p,$

$$E\left\{\left[s(n) - \sum_{j=1}^{p}\alpha_j s(n-j)\right]s(n-i)\right\} = 0, \quad \text{for } i = 1, \ldots, \mathrm{p}$$

  - $E\{s(n)s(n-i)\} = E\left\{\sum_{j=1}^{p}\alpha_j s(n-j)s(n-i)\right\} = \sum_{j=1}^{p}\alpha_j E\{s(n-j)s(n-i)\}$

Source-filter model of speech production

- Determine α_j (cont.)
  - $\sum_{j=1}^{p} \alpha_j \phi_n(i, j) = \phi_n(i, 0), \quad \text{for } i = 1, \ldots, p$
    - $\phi_n(i, j) = \mathrm{E}\{s(n-i)s(n-j)\}$
    - Therefore, given $\phi_n(i, j)$ and $\phi_n(i, 0)$, we can obtain α_j.
  - Assumption
    - Signal is stationary.
    - Not true over a long duration, but realistic for short segments since speech signal can be considered as quasi-stationary signal.
    - So, the ensemble average function can be approximated as the time average function.

Solutions to LPC analysis

- Expectation operation is replaced by time average operation.
  - $$\begin{aligned}\phi_n(i,j) &= \mathrm{E}\{s(n-i)s(n-j)\} \\ &= \sum_m s_n(m-i)s_n(m-j), \quad \text{for } i=1,\dots,p, \quad j=0,\dots,p\end{aligned}$$

- Auto-correlation method (AM)
  - Assumption: $s_n(m) = 0$ outside $0 \le m \le N-1$.
  - That is, there is a constraint on the signal itself, but not on the analysis frame.
  - Therefore, we should consider the prediction error in $0 \le m \le N-1+p$.

Auto-correlation method (AM)

- Solution of AM
  - Since $0 \le m-i \le N-1$ and $0 \le m-j \le N-1$, the range of the summation becomes $0 \le m \le N+p-1$ as in the above.
  - So, $\phi_n(i,j) = \sum_{m=0}^{N+p-1} s_n(m-i)s_n(m-j), \quad 1 \le i \le p, \; 0 \le j \le p$
  - To rearrange the eq., let $m-i=m'$.
    - Then, $m=m'+i$ and $m-j= m'+i-j$.
    - When $m=0$, $m'=-i$, and when $m=N+p-1$, $m'= N+p-1-i$.
    - Therefore, $\phi_n(i,j) = \sum_{m'=-i}^{N+p-1-i} s_n(m')s_n(m'+i-j).$
    - And, since $0 \le m' \le N-1$ and $0 \le m'+i-j \le N-1$ (or, $-(i-j) \le m' \le N-1-(i-j)$), we can obtain $0 \le m' \le N-1-(i-j)$.
    - Using this, $\phi_n(i,j) = \sum_{m'=0}^{N-1-(i-j)} s_n(m')s_n(m'+i-j).$

Auto-correlation method (AM)

- Solution of AM (cont.)
  - Consequently, $\phi_n(i,j) = \sum_{m=0}^{N-1-(i-j)} s_n(m)s_n(m+i-j), \quad 1 \le i \le p, \ 0 \le j \le p$
  - Now, we define $R_n(j) = \sum_{m=0}^{N-1-j} s_n(m)s_n(m+j)$
  - Then, the short-time autocorrelation function, ϕ_n

    $\phi_n(i,j) = R_n(i-j) = R_n(|i-j|), \quad \text{for } i = 1,\ldots,p \quad j = 0,\ldots,p$
    - This result can be easily derived if examining $R_n(1)$ and $R_n(-1)$.
  - Therefore, $\sum_{j=1}^{p} \alpha_j \phi_n(i,j) = \phi_n(i,0)$ is represented by

    $\sum_{j=1}^{p} \alpha_j R_n(|i-j|) = R_n(i), \quad 1 \le i \le p$

Auto-correlation method (AM)

- Solution of AM (cont.)
  - In normal matrix form,

$$\begin{bmatrix} R_n(0) & R_n(1) & . & R_n(p-1) \\ R_n(1) & R_n(0) & . & R_n(p-2) \\ \vdots & \vdots & \vdots & \vdots \\ R_n(p-1) & R_n(p-2) & . & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ \vdots \\ R_n(p) \end{bmatrix}$$

  - The above matrix eq. can be solved by normal matrix inversion formula, but this method requires a lot of computations and generally accumulates numerical errors due to finite precision computation.
  - However, if we utilize the property that the matrix is symmetric and has Toeplitz characteristics, we can efficiently solve the matrix eq. → Durbin's algorithm

Auto-correlation method (AM)

- Durbin's algorithm
  - Initialization: $E_n^{(0)} = R_n(0)$
  - For $1 \le i \le p$,

$$k_i = \left[R_n(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R_n(i-j) \right] / E_n^{(i-1)}$$

$$\alpha_i^{(i)} = k_i$$

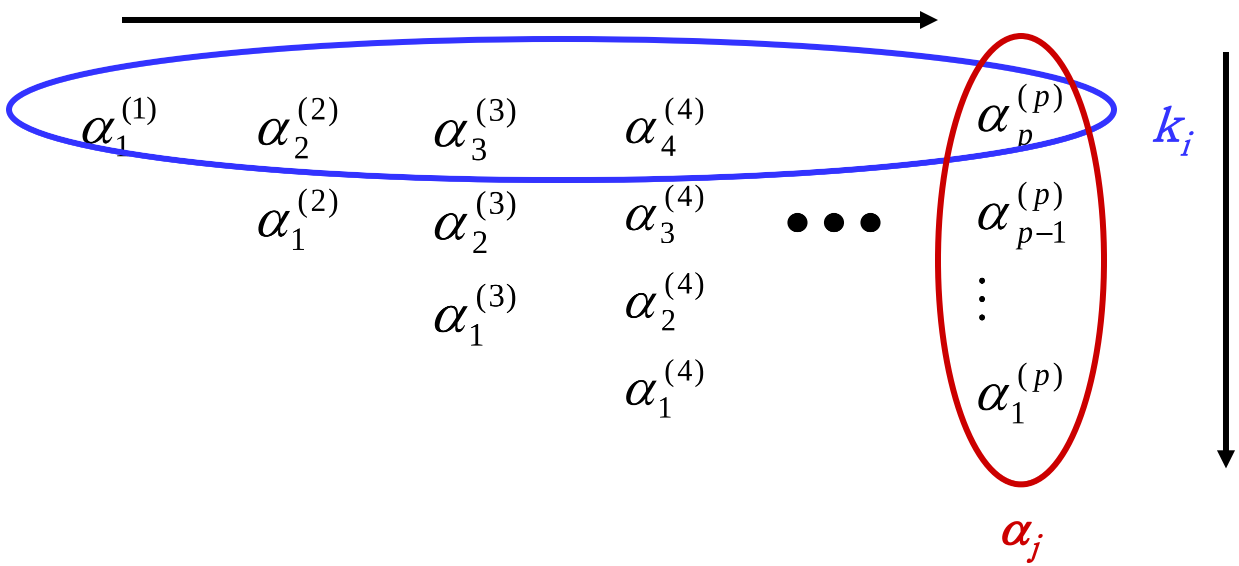
$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \quad 1 \le j \le i-1$$

$$E_n^{(i)} = (1 - k_i^2) E_n^{(i-1)}$$

  - Finally, $\alpha_j = \alpha_j^{(p)} \quad 1 \le j \le p$

Auto-correlation method (AM)

- The order of coefficient computation in the Durbin's recursion



Summary of lecture

- General speech characteristics
- Frequency domain analysis of speech signal
  - Short-time spectral analysis
  - Effects of different window functions
- Time domain analysis of speech signal
  - Linear predictive modeling of speech signals
  - Source-filter model of speech production
  - One of solutions to LPC analysis
    - Auto-correlation method (AM)
      - Durbin's solution