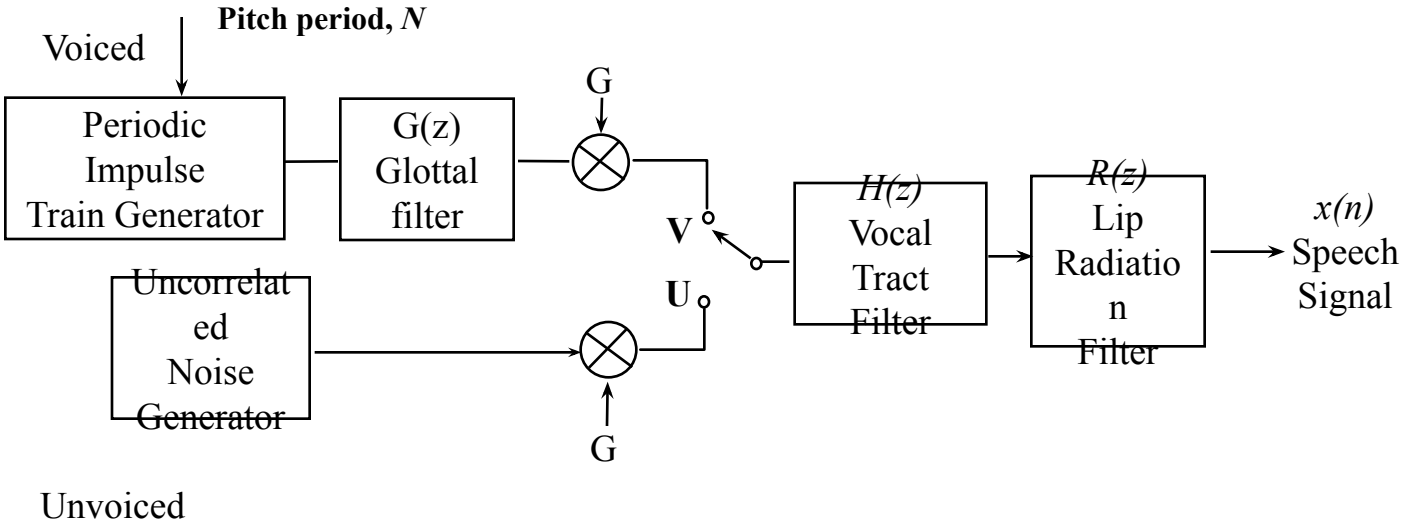
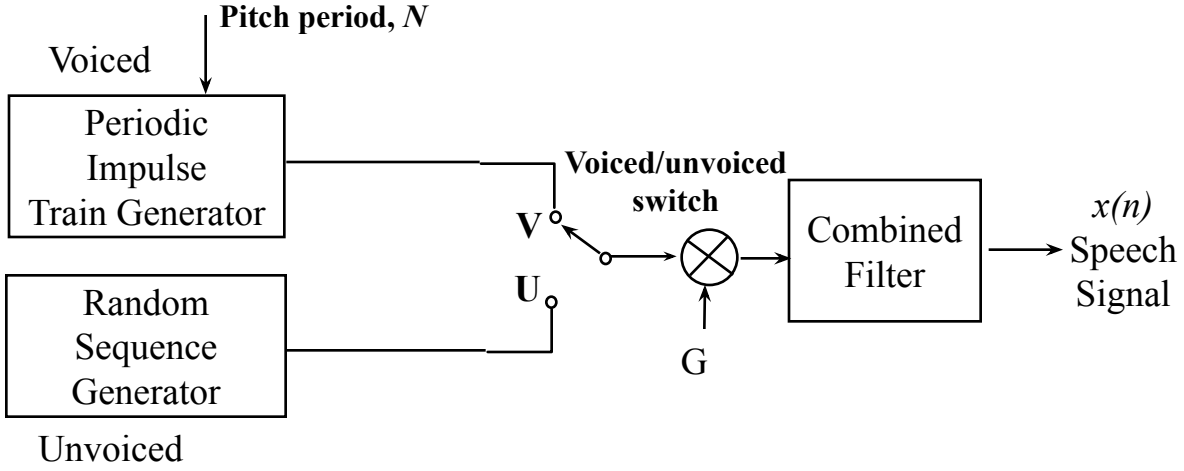
# Fundamental Frequency ($F_0$) Detection

# Speech Source Model

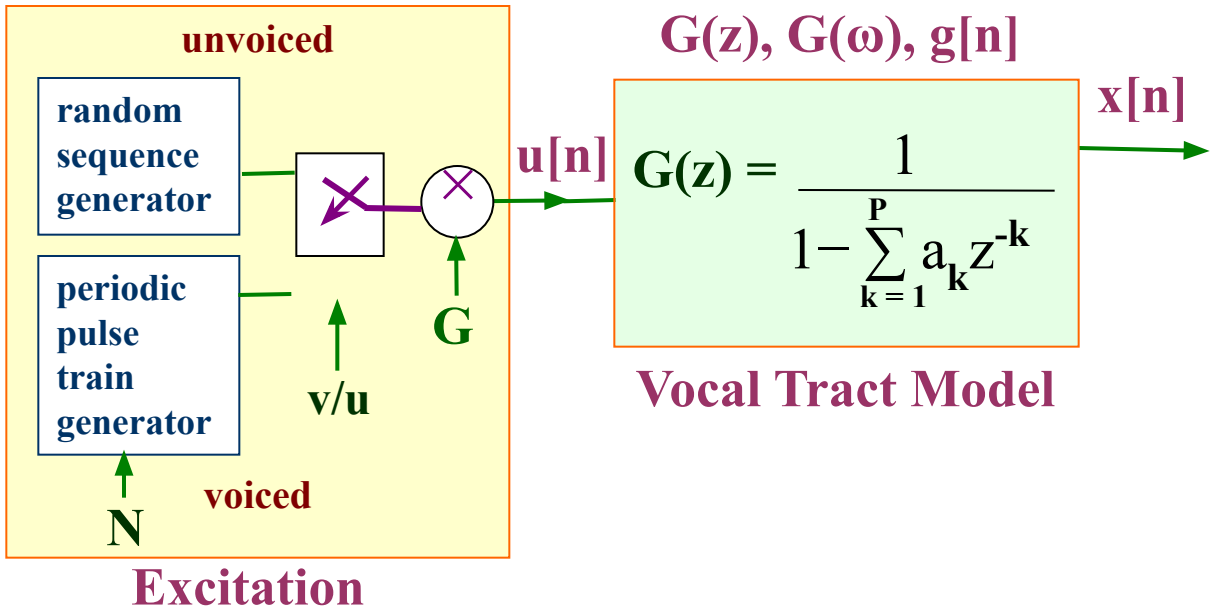
- **Sophisticated model for speech production**




- **Simplified model for speech production**



# Simplified Speech Source Model




– Excitation parameters
  v/u : voiced/ unvoiced
  N : pitch for voiced
  G : signal gain
  → excitation signal u[n]

– Vocal Tract parameters
  $\{a_k\}$ : LPC coefficients
  →formant structure of speech signals
– A good approximation, though not precise enough

# Fundamental Frequency Utilization

- Speech synthesis – Prosody Modeling.
- Speech Recognition
- Speech coding
- Voice conversion
- Spoken language learning

# Fundamental Frequency Characteristics

- F0 takes the values from 50 Hz (males) to 400 Hz (children)
  - If Fs=8000 Hz these frequencies correspond to the lags L=160 to 20 samples. It can be seen, that with low values F0 approaches the frame length 20 ms, which corresponds to 160 samples
- The difference in pitch within a speaker can reach to the 2:1 relation.
- Speech is called quasi periodic signal small changes between the period. These small shifts are called “jitter”
- F0 is influenced by many factors – usually the melody, mood, distress, etc.

# Issues in Fundamental Frequency Detection

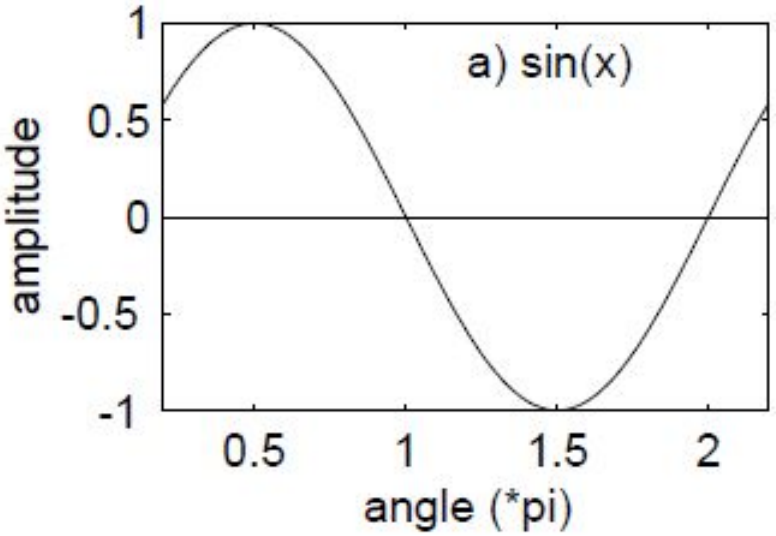
- Purely voiced or unvoiced excitation does not exist either. Usually, excitation is compound (noise at higher frequencies).
- Speech is called quasi periodic signal
- Difficult estimation of pitch with low energy.
- High F0 can be affected by the low formant F1 (females, children).
- During transmission over land line (300–3400 Hz) the basic harmonic of pitch is not presented but its folds (higher harmonics).

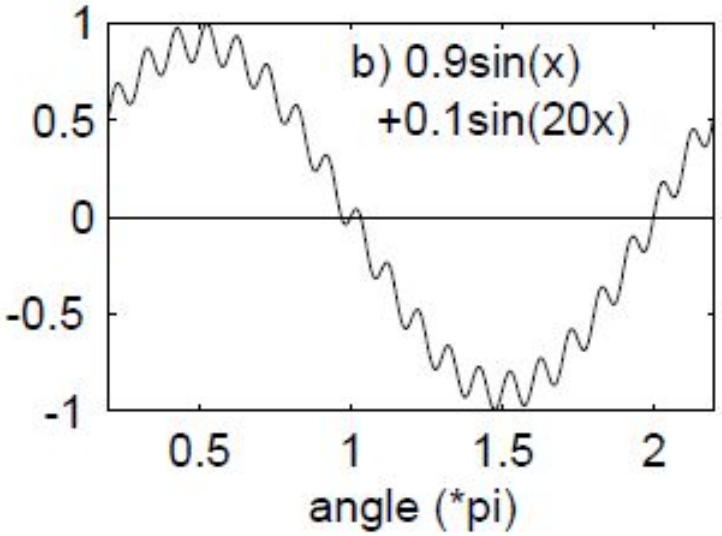
# Time Domain Methods

- Zero-Crossing Detection
- Autocorrelation Function
- Average Magnitude Difference Function

# Zero-Crossing Detection based F0 detection

- Based on a direct application of the definition of periodicity
- Counting the number of time that the signal crosses a reference level
- If the spectral power of the waveform is concentrated around $F_0$, then it will cross the zero line twice per cycle
- Mostly Inexpensive in computation
- Weakness against noise
- Presents weakness when used to analyze signals with energy in high frequencies


a) sin(x)
amplitude
1
0.5
0
-0.5
-1
0.5
1
1.5
2
angle (*pi)


b) 0.9sin(x)
+0.1sin(20x)
1
0.5
0
-0.5
-1
0.5
1
1.5
2
angle (*pi)

# Autocorrelation Technique

- Autocorrelation is a cross-correlation of a signal with itself.

$$\phi(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+\tau)$$

- The maximum of similarity occurs for time shifting of zero.
- An other maximum should occur in theory when the time-shifting of the signal corresponds to the fundamental period.

# Autocorrelation function

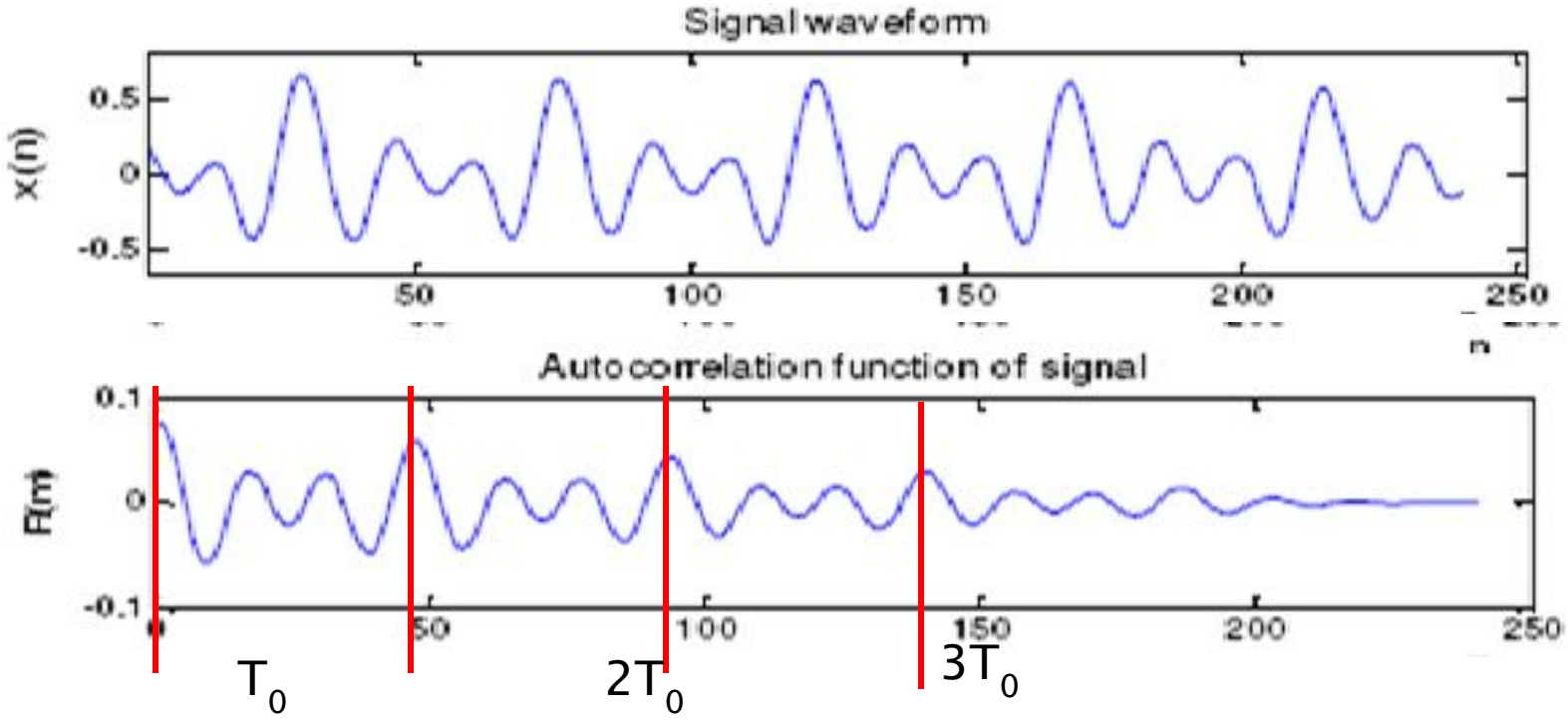
By definition, auto - correlation is

$$R[k] = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x[n] \cdot x[n+k], \quad 0 \le k \le K_0$$

Properties of Autocorrelations is

1. R[k] = R[-k]
2. R[k] is maximum at k = 0

$$R[k] = \frac{1}{N} \sum_{n=0}^{N-1-k} x[n] \cdot x[n+k], \quad 0 \le k \le K_0$$




When a segment of a signal is correlated with itself, the distance (*Lag_time_in_samples*) between the positions of the maximum and the second maximum is defined as the *fundamental period* (pitch) of the signal.

# The result can be enhanced by

- Center clipping
- Using cross-correlation
- Filtering the speech signal

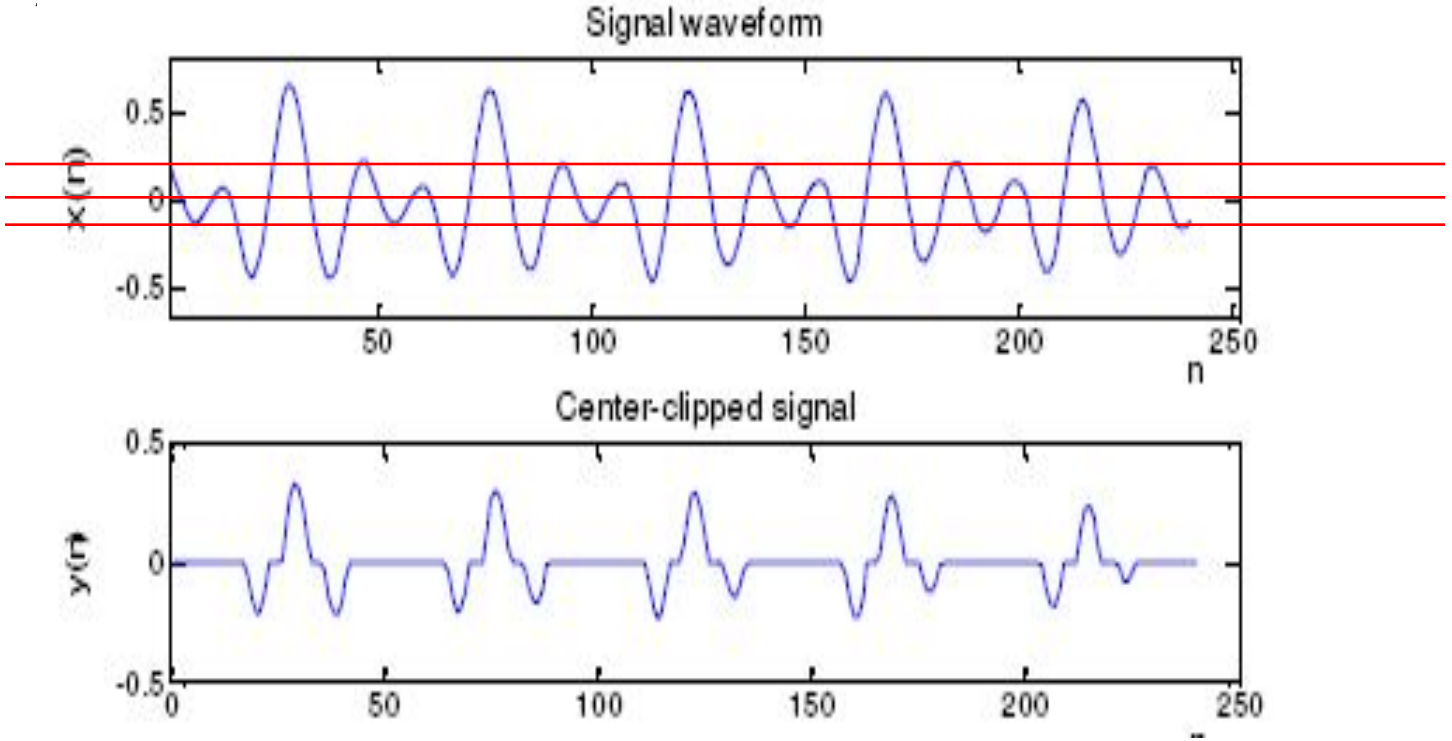
  Pitch range is 40 – 400 Hz (40-200 Hz for male, and 200-400 Hz for female)

# Center clipping

$$y(n) = clc[x(n)] = \begin{cases} (x(n) - C_L), & x(n) \geq C_L \\ 0 & , |x(n)| < C_L \\ (x(n) - C_L), & x(n) \leq C_L \end{cases}$$

Typical $C_L$ =1/4 peak-to-peak of X

$$R'(m) = \sum_{n=0}^{N-1-m} y(n) \cdot y(n+m), \quad 0 \leq m \leq M_0$$



# Clipping Level Value Estimation

As a speech signal s(n) is a non-stationary signal, the slipping level changes and it is necessary to estimate it for every frame, for which pitch is predicted. Simple method is to estimate the clipping level from the absolute maximum value in the frame:

$$c_L = k \max_{n=0\ldots N-1} |x(n)|,$$

*Where the constant k is selected between 0.6 and 0.8.*

Further, subdivision into several micro-frames can be done, for instance x1(n), x2(n), x3(n) of one third of the original frame length. The clipping level is then given by the lowest maximum from the micro-frames:

$$c_L = k \min \{\max |x_1(n)|, \max |x_2(n)|, \max |x_3(n)|\}$$

**Issue:** clipping of noise in pauses, where subsequently can be detected pitch. The method therefore should be preceded by the silence level $s_L$ estimation. In the maximum of the signal is < $s_L$, then the frame is not further processed.

# Normalized cross correlation function (NCCF) method

The normalized cross correlation function (NCCF) is very similar to the autocorrelation function, but is better follows the rapid changes in pitch and the amplitude of speech signal.

The NCCF based PDA overcomes most of the shortcomings of the autocorrelation based algorithms at a slight increase in computational complexity.

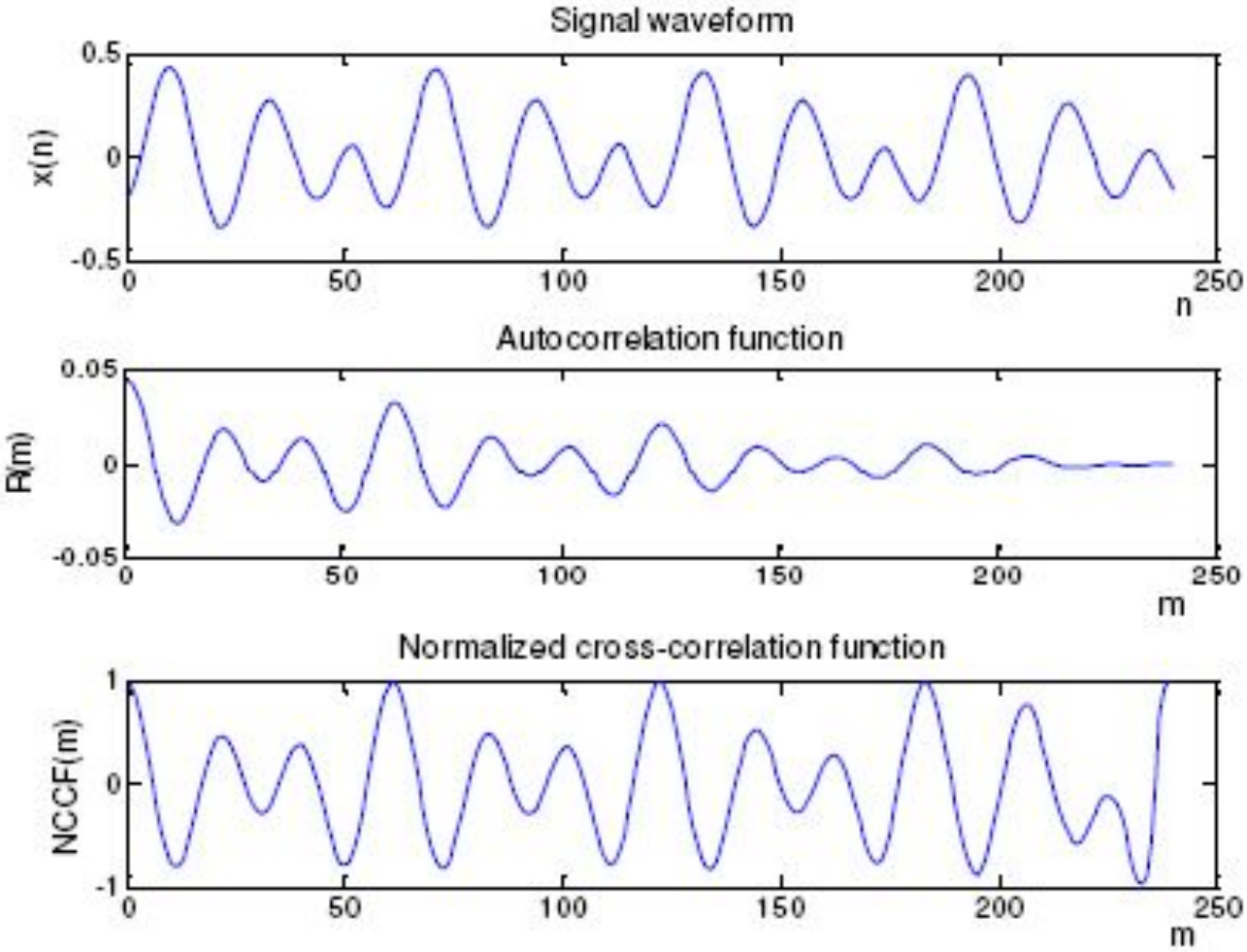
The NCCF function for speech segment x(n), 0 ≤ n ≤ N-1 is defined

$$NCCF(m) = \frac{\sum_{n=0}^{N-m-1} x(n) \cdot x(n+m)}{\sqrt{\sum_{n=0}^{N-m-1} x^2(n) \cdot \sum_{n=0}^{N-m-1} x^2(n+m)}}, \quad 0 \le m < M_0$$

# Utilization of the Linear Prediction Error

$$
\begin{aligned}
e(n) &= s(n) - \hat{s}(n) \\
E(Z) &= S(z)[1 - (1 - A(z))] = S(z)A(z) \\
e(n) &= s(n) + \sum_{i=1}^{P} a_i s(n-i)
\end{aligned}
$$

The signal e(n) contains no information about formants, thus is more suitable for the estimation. Lag estimation from the error signal can be done using the ACF method


Signal waveform
x(n)
0.5
0
-0.5
0
50
100
150
200
250
n
Autocorrelation function
R(m)
0.05
0
-0.05
0
50
100
150
200
250
m
Normalized cross-correlation function
NCCF(m)
1
0
-1
0
50
100
150
200
250
m

# Autocorrelation Technique

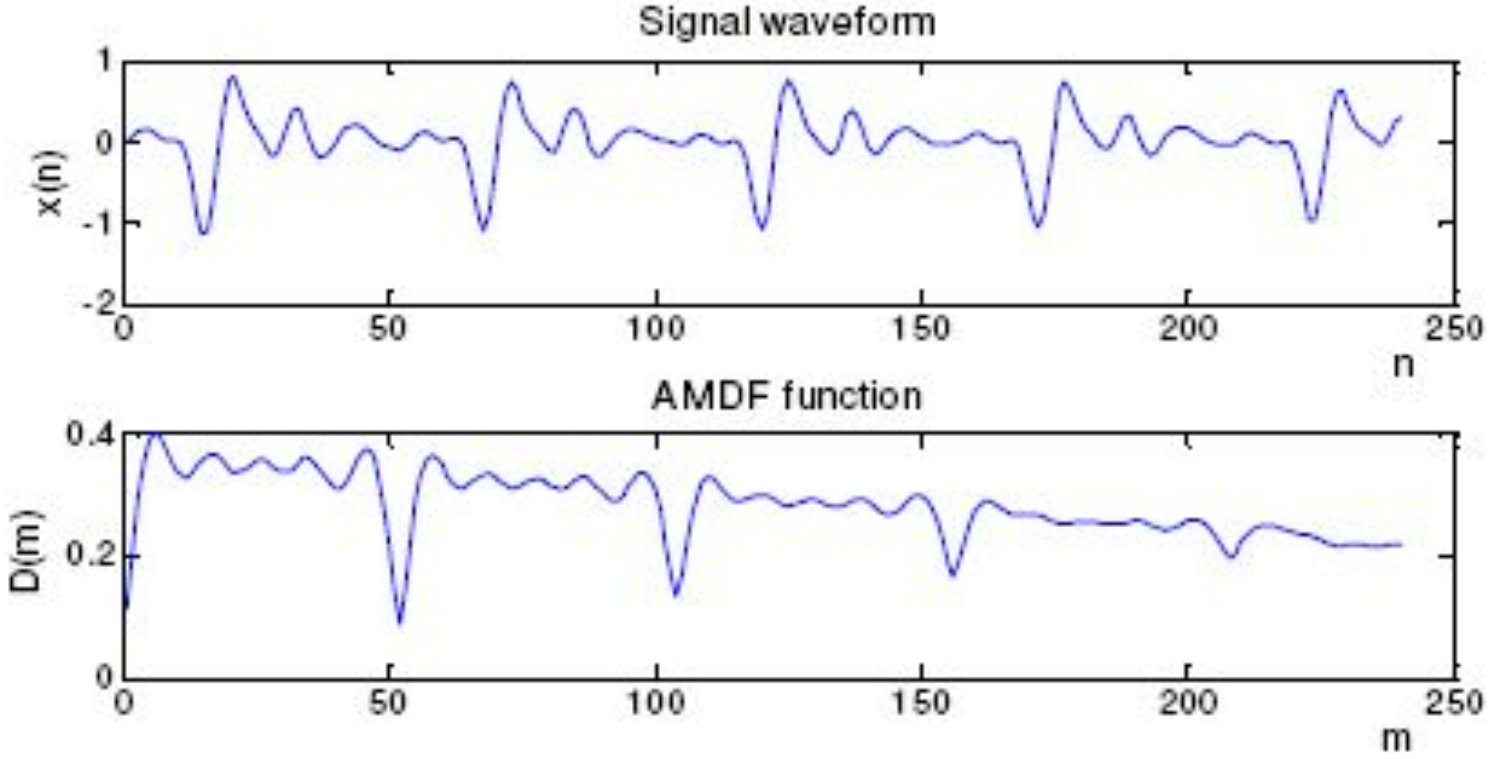
- Not very efficient for high fundamental frequency.
- Convolution is a very expensive process.
- Computation efficiency can be improved using the FFT algorithm instead of convolution. It reduces calculation from N squared to N $\log_2$N.
- Most of the variation of this technique related to the mathematical definition of the autocorrelation used, the way the maximums are localized, and how errors in the maximum identification are attenuated.

# Average Magnitude Difference Function(AMDF)

- It is an alternate to Autocorrelation function.
- It compute the difference between the signal and a time-shifted version of itself.

$$D_x[k] = \frac{1}{N} \sum_{n=0}^{N-1-k} \left| x(k) - x(n+k) \right|, \qquad 0 \le k \le K_0$$

- While autocorrelation have peaks at maximum similarity, there will be valleys in the average magnitude difference function.


Signal waveform
x(n)
1
0
-1
-2
0
50
100
150
200
250
n
AMDF function
D(m)
0.4
0.2
0
0
50
100
150
200
250
m

- Less computational complex so suitable for H/W implementation

- Performance is poor and it does not cater for variation in the energy of the speech

# Other Temporal Algorithm

- Waveform Maximum Detection
- Sum Magnitude Difference Squared Function
- Average Squared Difference Function
- Cumulative Mean Normalized Difference Function
- Circular Average Magnitude Difference Function
- Adaptive Filter

# Frequency Domain Pitch Detection Algorithms (PDAs)

- Frequency domain PDAs operate on speech spectrum.
- Distance between harmonics is the fundamental frequency, or inverse of the pitch period.
- Main drawback of frequency domain PDAs is high computational complexity.

# Frequency Domain Pitch Detection Algorithms (PDAs)

- Two kinds of frequency PDAs are:
  1. Harmonic Peak Detection
  2. Spectrum Similarity

# Harmonic Peak Detection

- Detect all harmonic peaks.
- Get the fundamental frequency by recognizing it as the common divisor or the spacing of adjacent harmonics.
- This can be done using a comb filter.

$$C(\omega,\omega_0)=\begin{cases} W(k\omega_0), & \omega=k\omega_0 \quad k=1,2,\ldots.\dfrac{\Omega_m}{\omega_0} \\ 0 \quad otherwise \end{cases}$$

$$A_c(\omega_0)=\frac{\omega_0}{\Omega_m}\sum_{k=1}^{\Omega_m/\omega_0} S(k\omega_0)W(k\omega_0)$$

$$where \quad \Omega_m is \quad the \quad \max imum \quad frequency$$

# Spectrum Similarity

- Assumes that the spectrum is fully voiced and is composed only of harmonics located at multiples of the pitch frequency.
- Synthetic spectrum is created from each frequency candidate and compared to original spectrum.
- Best one to match the original spectrum is selected as the fundamental frequency.

# Time and Frequency Domain PDAs

- Combine results from time domain algorithms and frequency domain algorithms to provide a better pitch estimate.
- E.g. Autocorrelation for time and frequency domain is given by :

$$R_{ST}(\tau) = \alpha R_T + (1-\alpha) R_S(\tau)$$

# Voiced/Unvoiced Classification

- Periodic Similarity
- Peakiness of Speech
- Zero Crossing
- Spectrum Tilt

# Periodic Similarity

- Most prominent characteristic that separates voiced speech from unvoiced speech is its regularity and fairly well-defined pitch.

$$P_s = \frac{\sum_{i=1}^{N} s[i]s[i-T]}{\sum_{i=1}^{N} s^2[i] \sum_{i=1}^{N} s^2[i-T]}$$

# Peakiness of Speech

- Voiced speech contains regular pulses which do not appear in unvoiced speech.
- Unvoiced speech, however, may contain a few random spikes and in these cases, those frame may be incorrectly labeled as voiced.

LPC residue is used to compute its value

$$P_k = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} r^2(i)}}{\frac{1}{N}\sum_{i=1}^{N} |r(i)|}$$

# Zero Crossing

- Number of times unvoiced speech crosses the zero line is significantly higher than that of voiced speech.
- Gender of speaker can also have an effect on zero crossing.
- Small pitch weighting can be used to weight the decision threshold.

# Spectrum Tilt

- Voiced speech has higher energy in low frequencies and unvoiced speech has higher energy in high frequencies.
- Can avoid individual spikes in low-level signals.

$$S_t = \frac{\sum_{i=1}^{N} s(i)s(i-1)}{\sum_{i=1}^{N} s^2(i)}$$