Basic Parameter Extraction

- There are a number of very basic speech parameters which can be easily calculated for use, in simple applications:
  - Short Time Energy
  - Short Time Zero Cross Count (ZCC)
  - Pitch Period
- All of the above parameters are typically estimated for frames of speech between 10 and 20 ms long

Short Time Energy

- The short-time energy of speech may be computed by dividing the speech signal into frames of N samples and computing the total squared values of the signal samples in each frame.
- Splitting the signal into frames can be achieved by multiplying the signal by a suitable window function w(n) {n=0, 1, 2, 3, …, N-1}, which is zero for n outside the range (0, N-1)

Rectangular Window

- A simple rectangular window of duration of 12.5 ms is suitable for this purpose. For a window starting at sample m, the short-time energy E_m is defined as

$$E_m = \sum_n [s(n)\, w(m-n)]^2$$

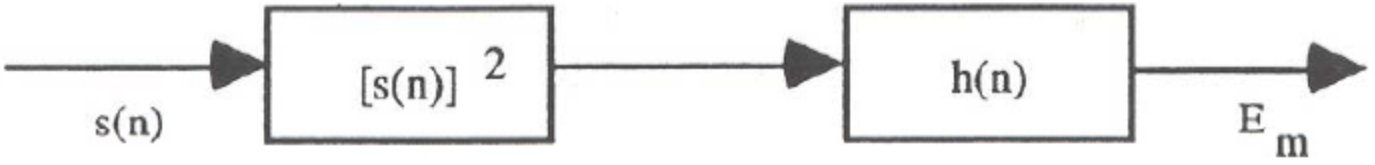
$$w(n) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & otherwise \end{cases}$$

$$\rightarrow \quad E_m = \sum_n [s(n)]^2\, h(m-n)$$

$$h(n) = [w(n)]^2$$

Linear filter representation

- The above equation (see previous slide) can thus be interpreted as



The signal $s(n)^2$ is filtered by a linear filter with impulse response h(n) .

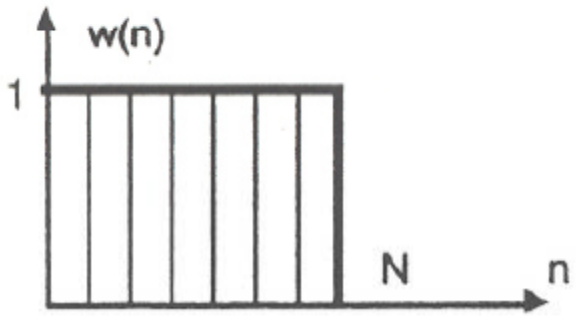
The choice of the impulse response ,h(n) or equivalently the window, determines the nature of the short-time energy representation.

To see how the choice of window affects the short-time energy, let us observe that if h(n) was very long and of constant amplitude E_m would change very little with time

Such a window would be equivalent of a very narrowband lowpass filter. Clearly what is desired is some lowpass filtering, so that the short-time energy reflects the amplitude variations of the speech signal.

We wish to have a short duration window to be responsive to rapid amplitude changes. But a window that is too short will not provide sufficient averaging to produce a smooth energy function.

Note: Rectangular window



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

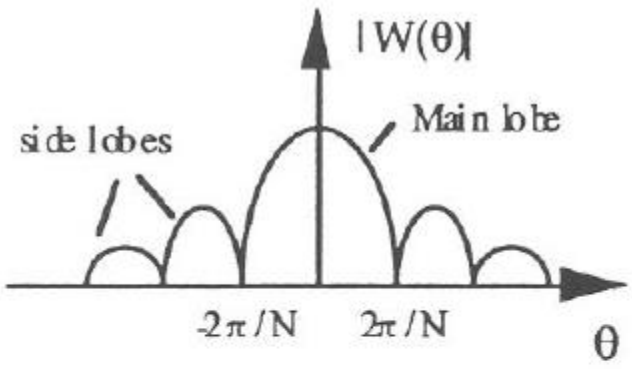
$$W(z)=\sum_{n=0}^{N-1}w(n)z^{-n}=1+z^{-1}+z^{-2}+z^{-3}+....+z^{-(N-1)}=\frac{1-z^{-N}}{1-z^{-1}}$$

$$W(\theta)=W(z)\big|_{z=e^{j\theta}}=\frac{1-e^{-jN\theta}}{1-e^{-j\theta}};$$

$$W(\theta) = e^{-j\frac{N-1}{2}\theta} \quad \frac{Sin\frac{N\theta}{2}}{Sin\frac{\theta}{2}}$$

↑ Phase term ↑ Magnitude

If N is too small, E_m will fluctuate very rapidly depending on exact details of the waveform.

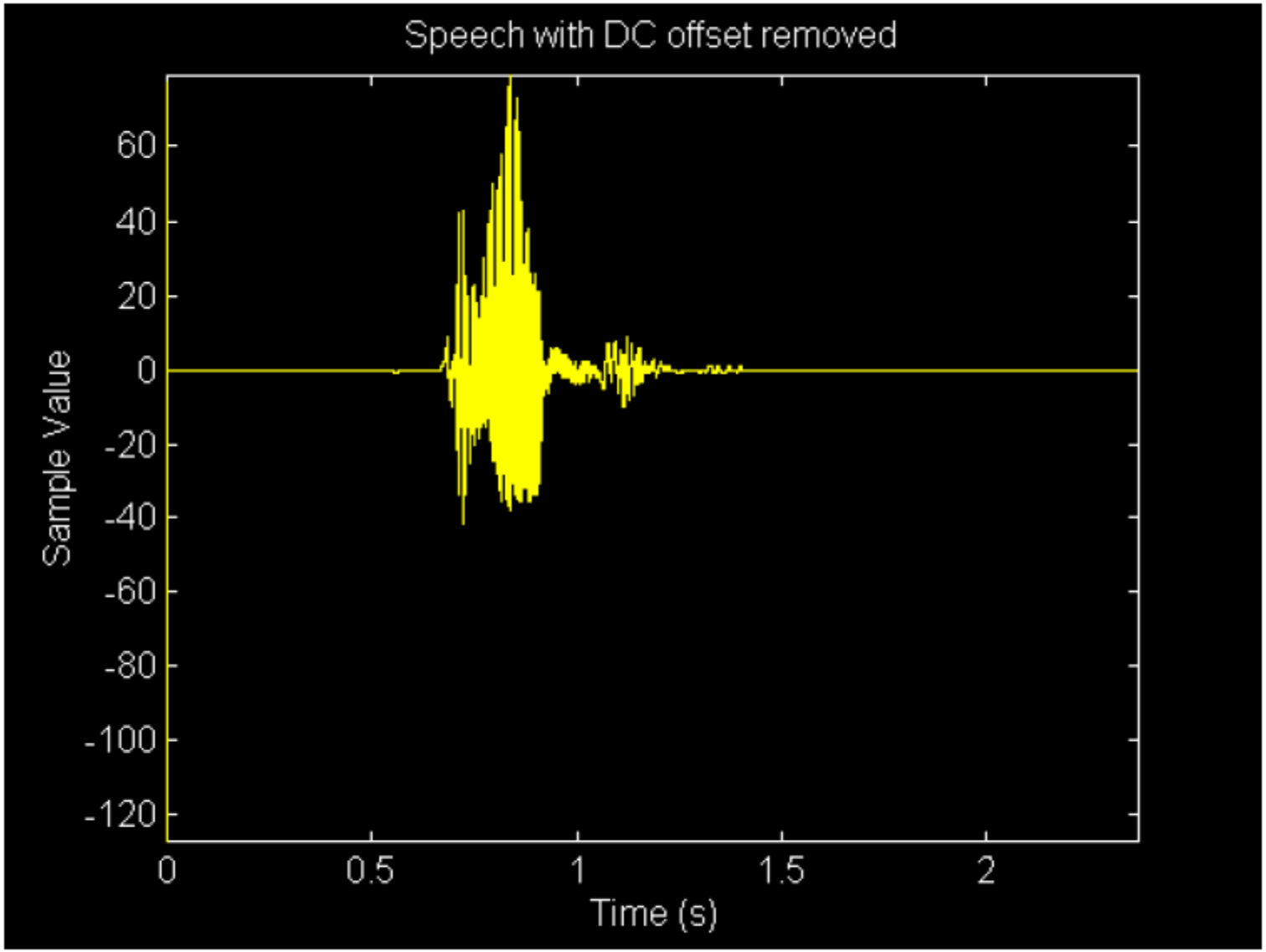


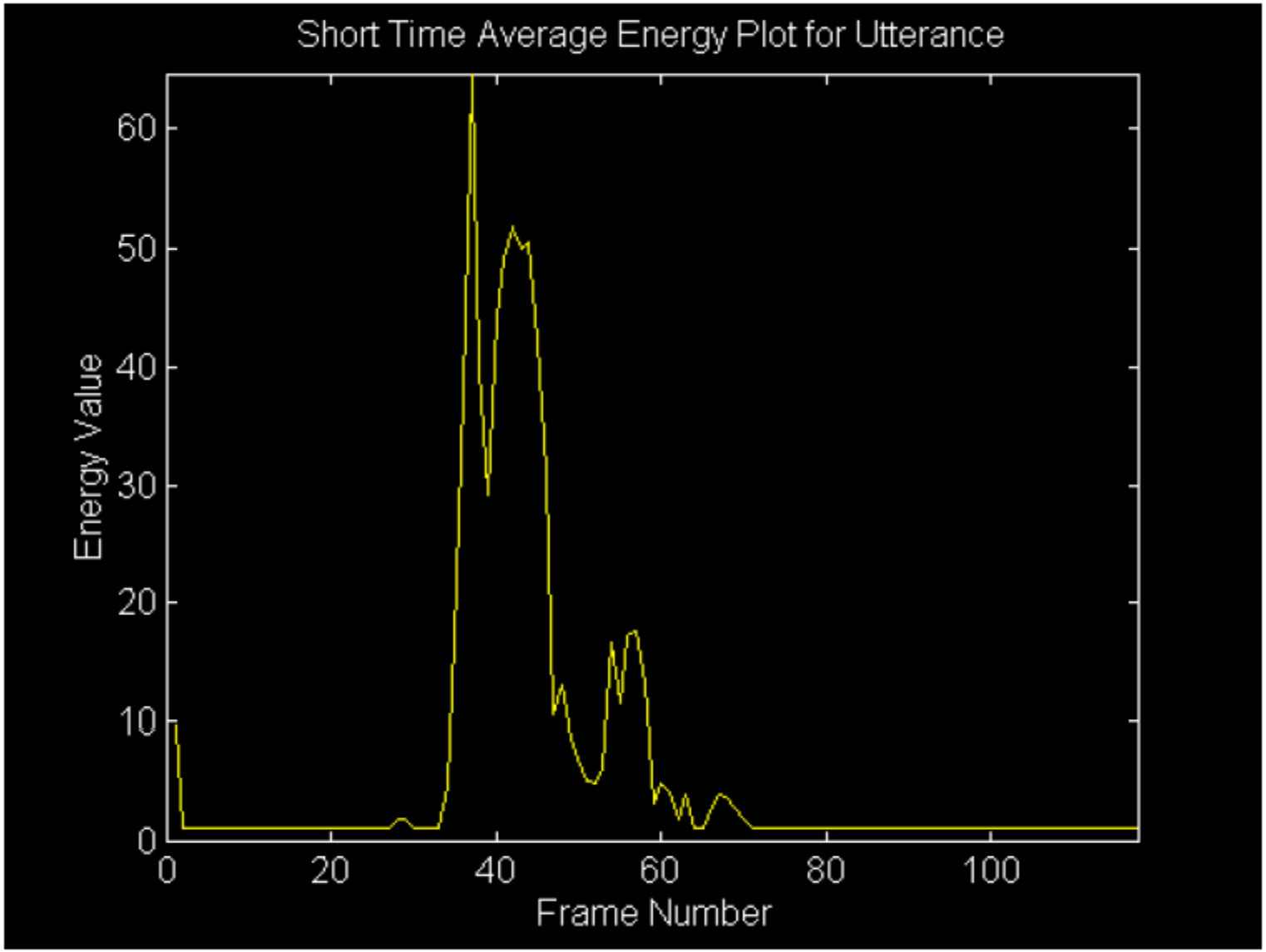
In N is too lager, E_m will change very slowly and thus will not adequately reflect the changing properties of the speech signal

Choice of Window Size

- Unfortunately this implies that no single value of N is entirely satisfactory.

- A suitable practical choice for N is on the order of 100-200 samples for a 10 kHz sampling rate (10-20 ms duration)





Note that a recursive lowpass filter H(z) can also be used to calculate the short-time energy:

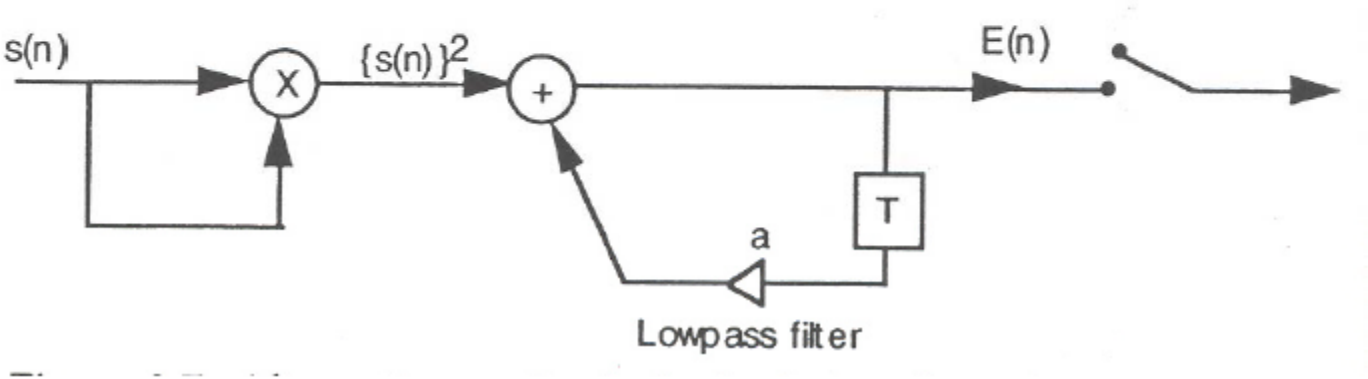
$$H(z) = \frac{1}{1 - az^{-1}} \qquad 0 < a < 1$$

It can be easily verified that the frequency response H(θ) has the desired lowpass property. Such a filter can be implemented by a simple difference equation:

$$E(n) = a\ E(n-1) + [s(n)]^2$$

E(n) is the energy at the time instant n

The structure for calculating the short-time energy recursively



The quantity E(n) must be computed at each sample of input speech signal, even though a much lower sampling rate suffice.

The value 'a' can be calculated using $a = e^{(-f_c 2\pi / f_s)}$

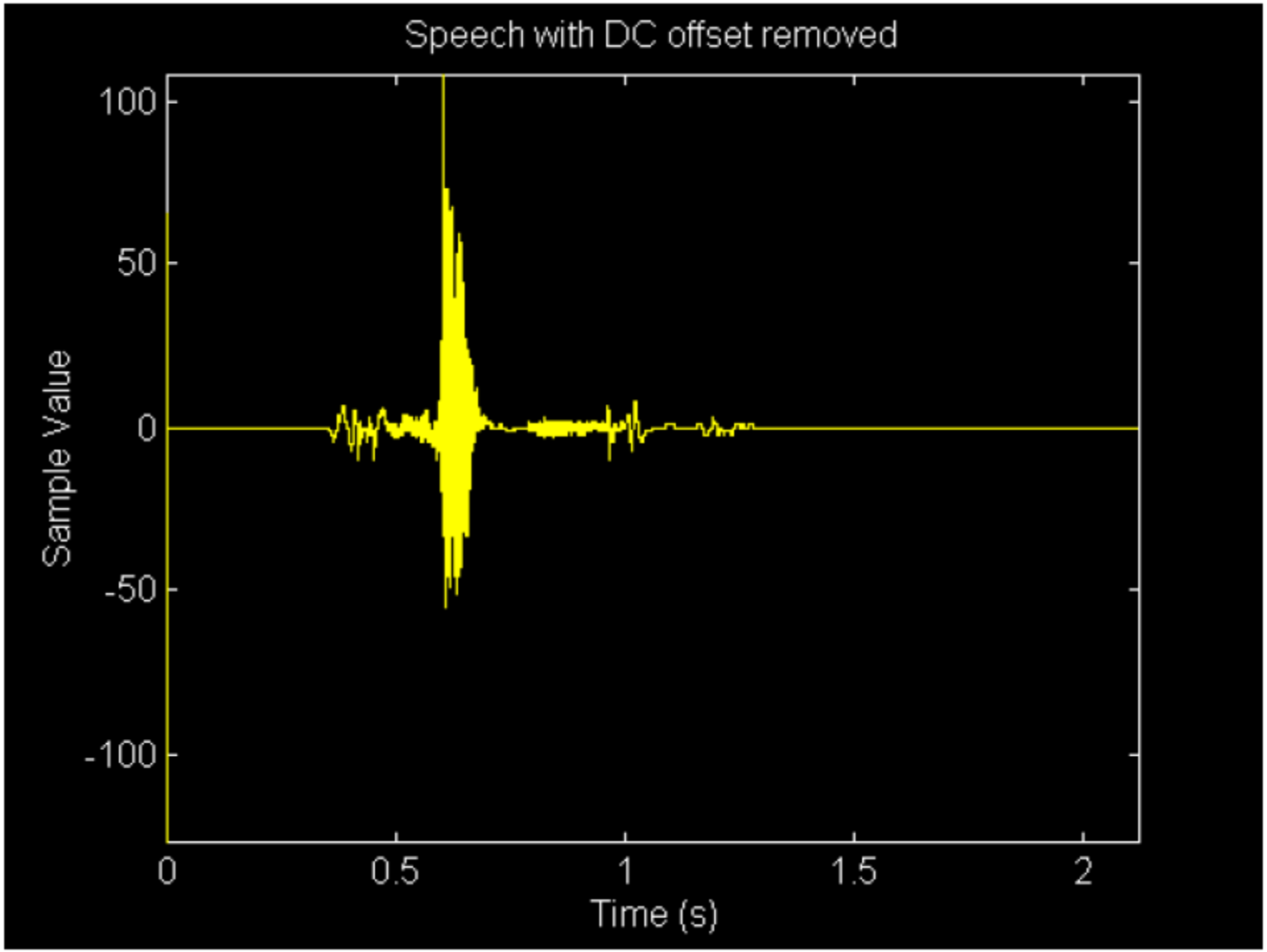
Fc is the cut-off frequency and fs is the sampling frequency (e.g fc=30 Hz, fs = 8000Hz)

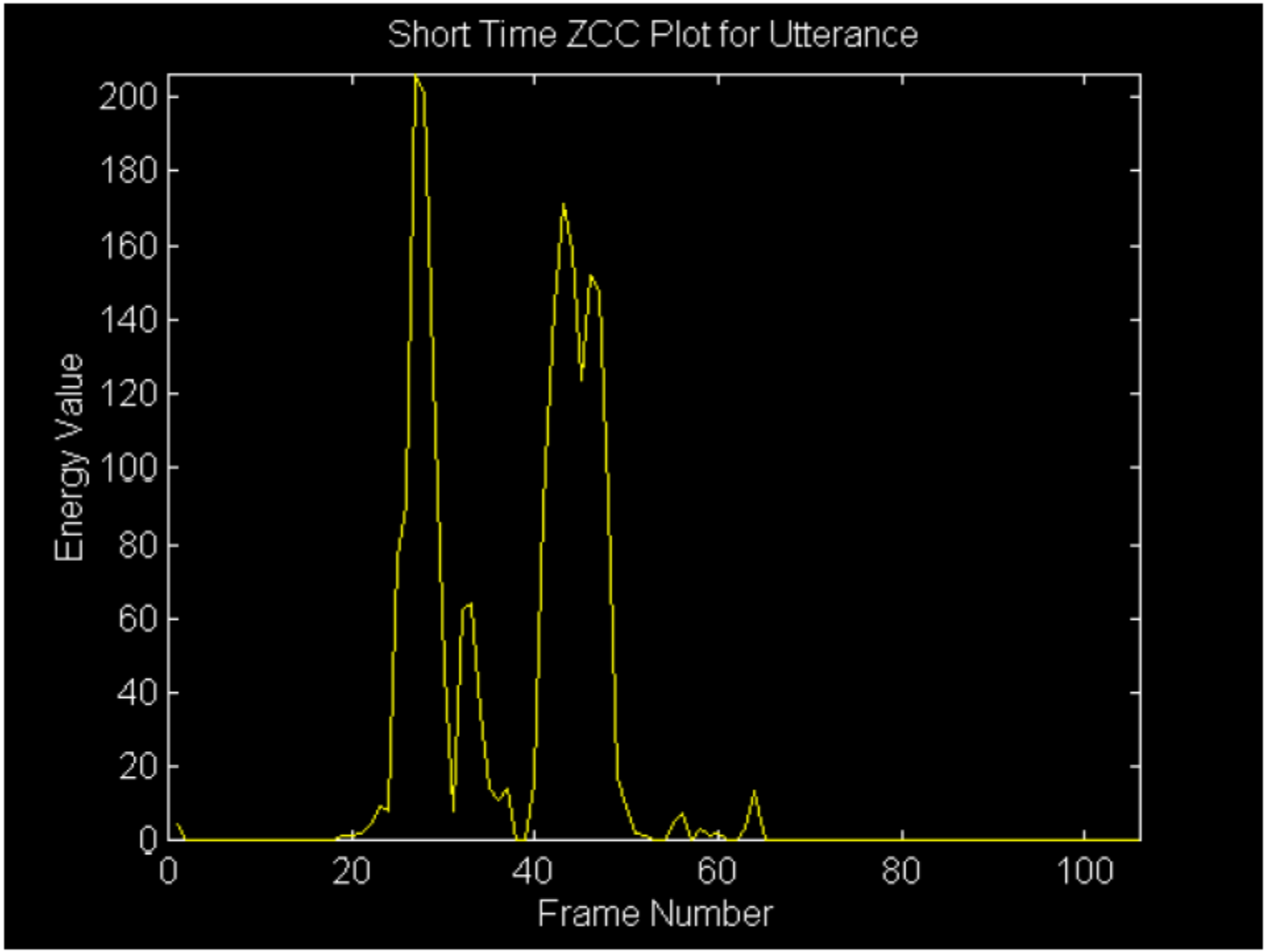
Short Time Zero Crossing Count

- The Short Time ZCC is calculated for a block of N samples of speech as

$$ZCC_i = \sum_{k=1}^{N-1} 0.5 \,|\, sign(s[k]) - sign(s[k-1]) \,|$$

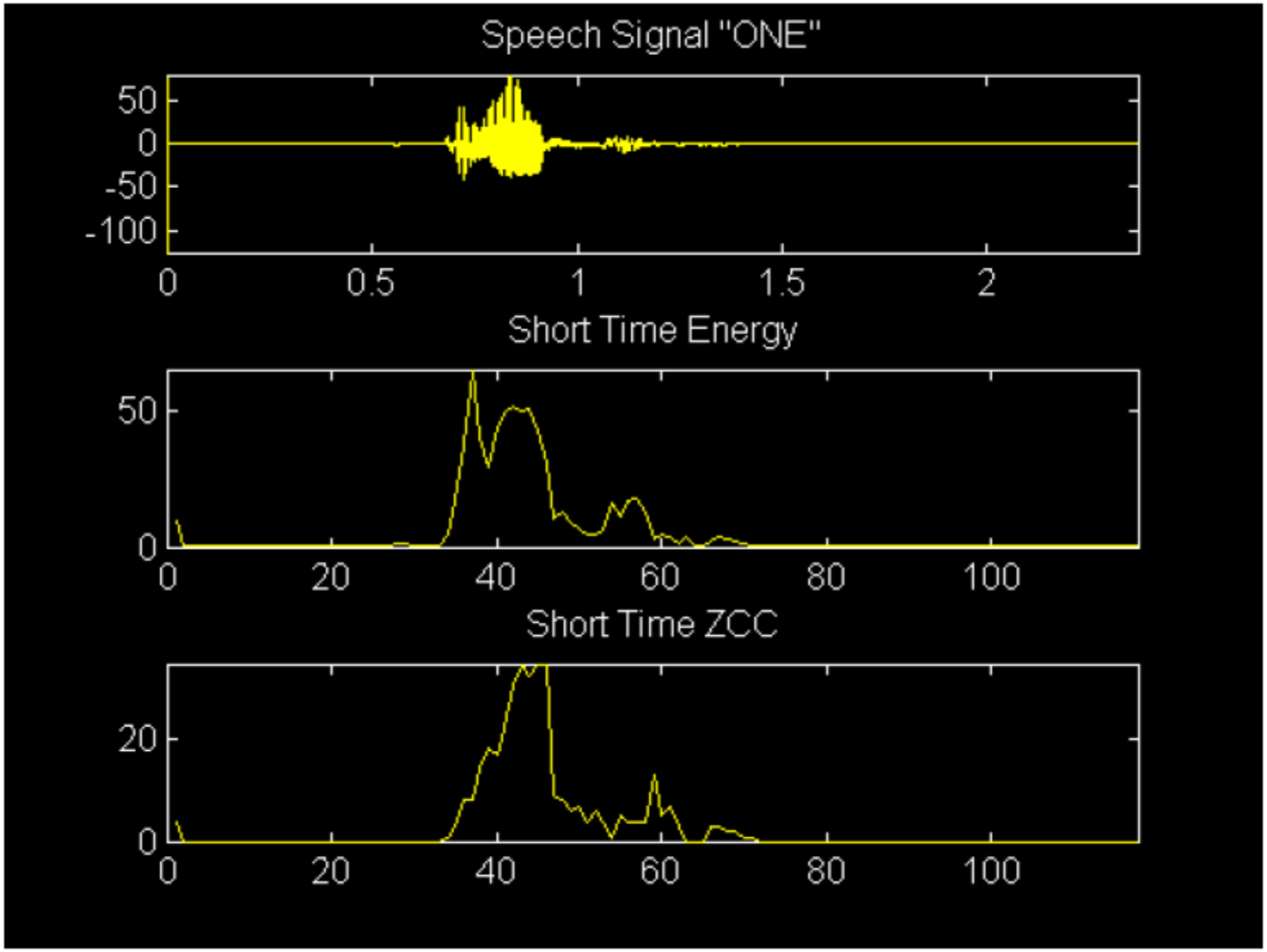
- The ZCC essentially counts how many times the signal crosses the time axis during the frame
  - It "reflects" the frequency content of the frame of speech
    - High ZCC implies high frequency
- It is essential that any constant DC offset is removed from the signal prior to ZCC calculation

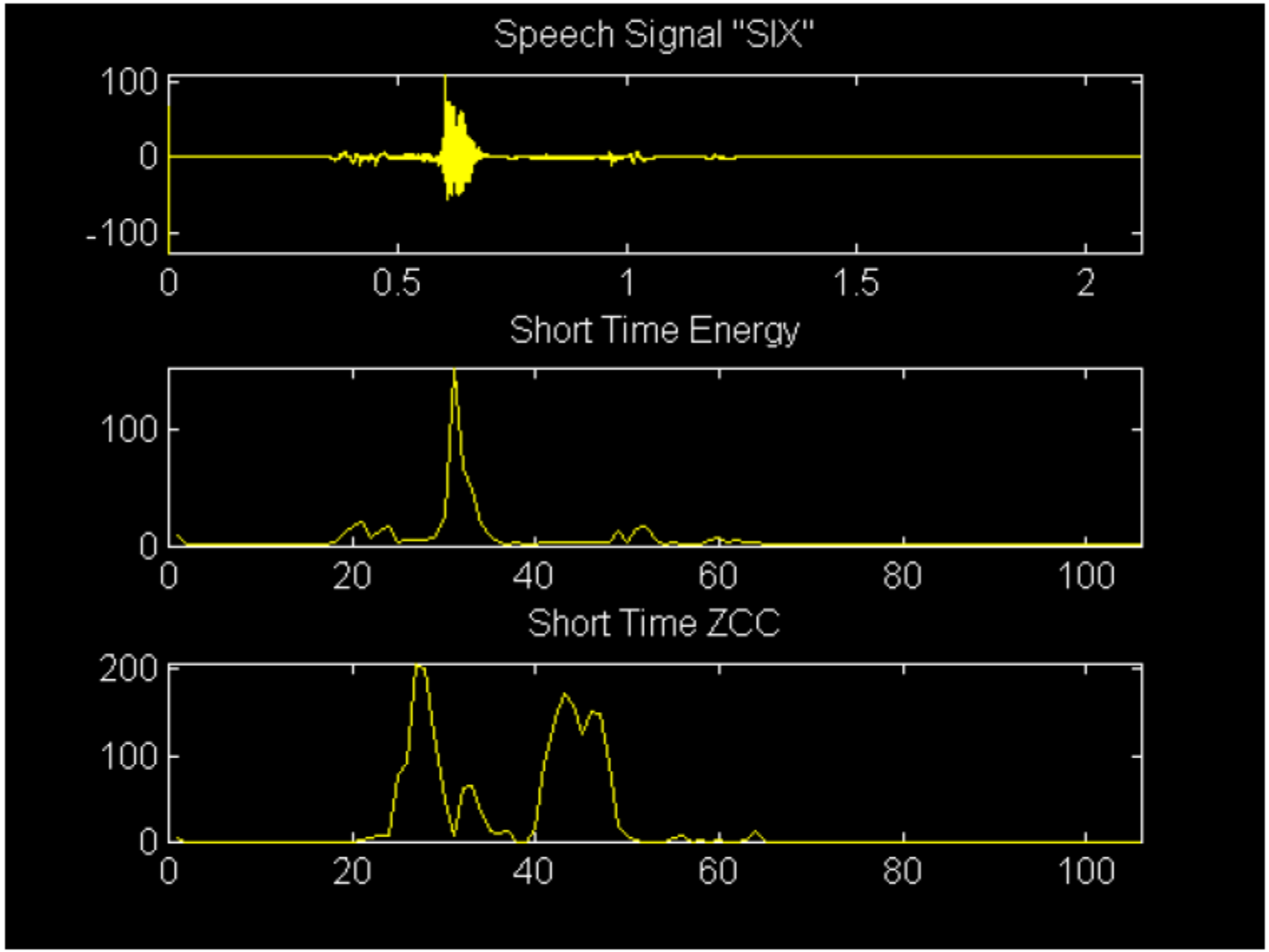


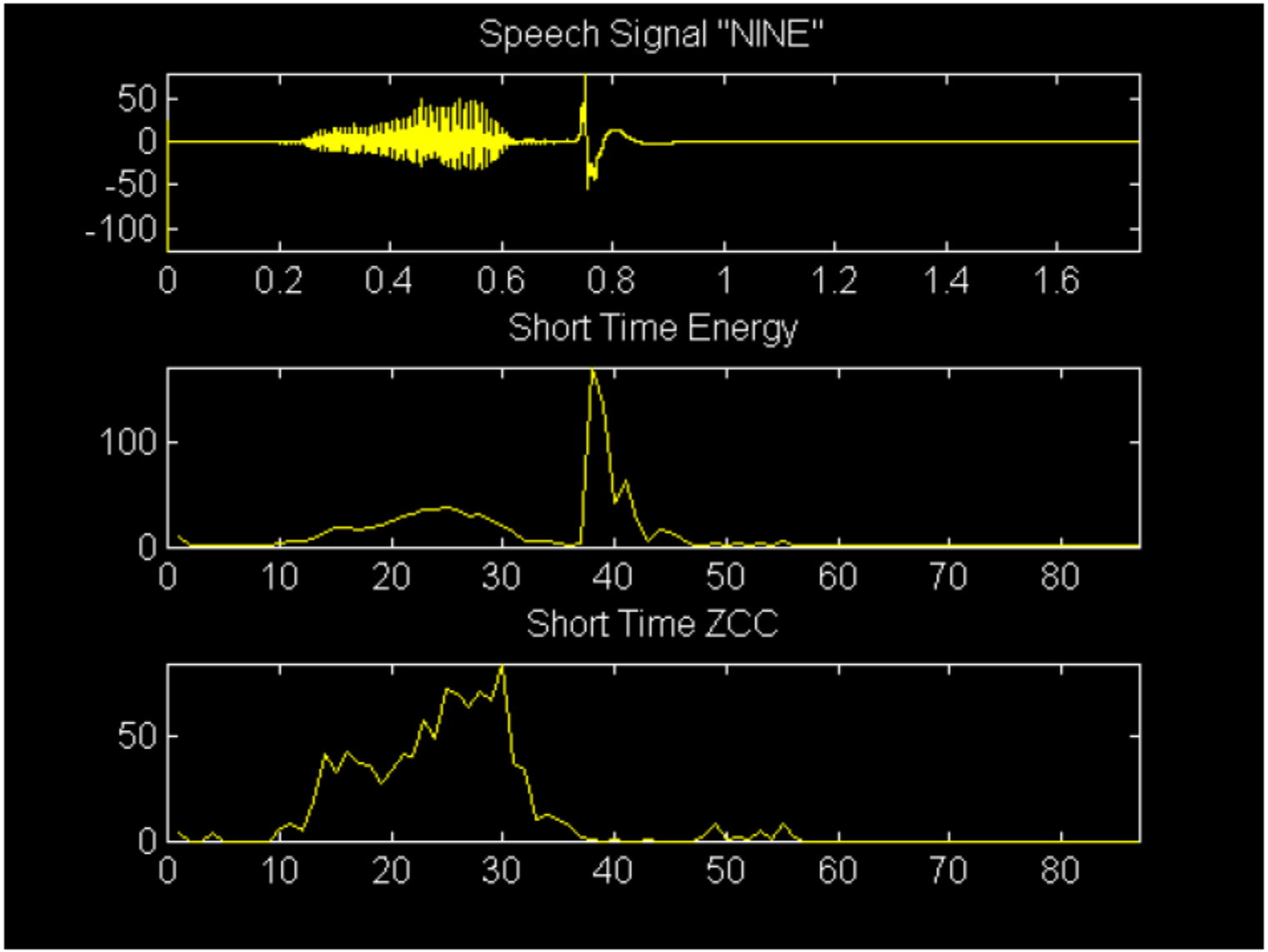


Uses of Energy and ZCC

- Short Time Energy and ZCC can form the basis for :
  - Automated speech "end point" detection
    - Needs to be able to operate with background noise
    - Needs to be able to ignore "short" background noises and intra-word silences (temporal aspects)
  - Voiced\Unvoiced speech detection
    - High Energy + Low ZCC – Voiced Speech
    - Low Energy + High ZCC – Unvoiced Speech
  - Parameters on which simple speech recognition\speaker verification\identification systems could be based

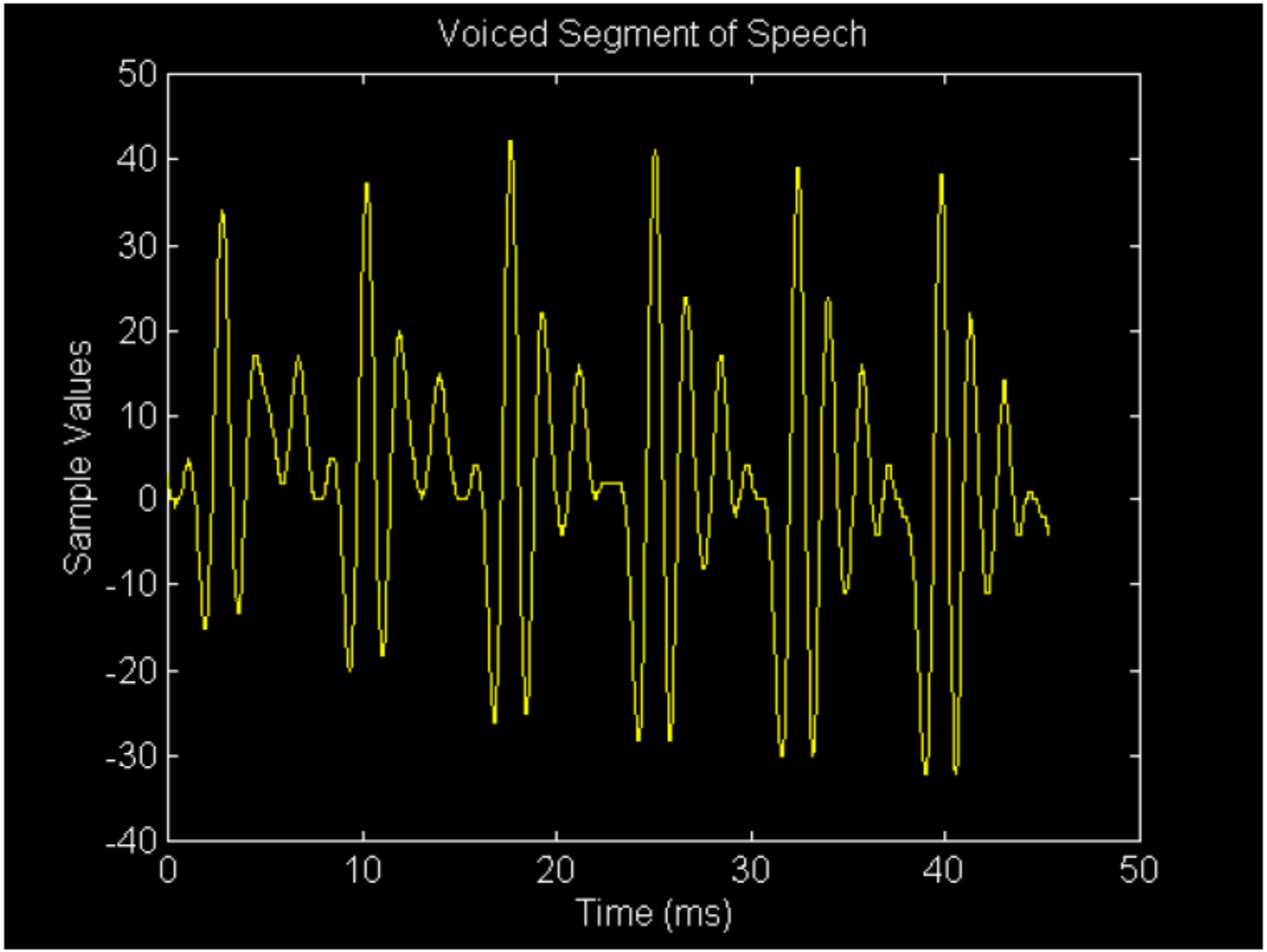






Pitch Period Estimation

- Pitch period is equal to the inverse of the fundamental frequency of vibration of the vocal chords
- It only makes sense to speak about the pitch period of a VOICED frame of speech
- Number of techniques used to determine pitch period
  - Time Domain
  - Frequency Domain



Time Domain Methods

- Since pitch frequency is typically less then 600-700 Hz, the speech signals are first low passed filtered to remove components above this frequency range
- The two most commonly used techniques are:
  - Short Time Autocorrelation Function
  - Average Magnitude Difference Function (AMDF)
- During voiced speech, the speech signal is "quasi-periodic"
- Either technique attempts to determine the period (in samples between "repetitions" of the voiced speech signal

Autocorrelation Function

- Correlation is a very commonly used technique in DSP to determine the "time difference" between two signals, where one is a "nearly perfect" delayed version of the other
- Autocorrelation is the application of the same technique to determine the unknown "period" of a quasi-periodic signal such as speech
- The autocorrelation function for a delay value of k samples is:

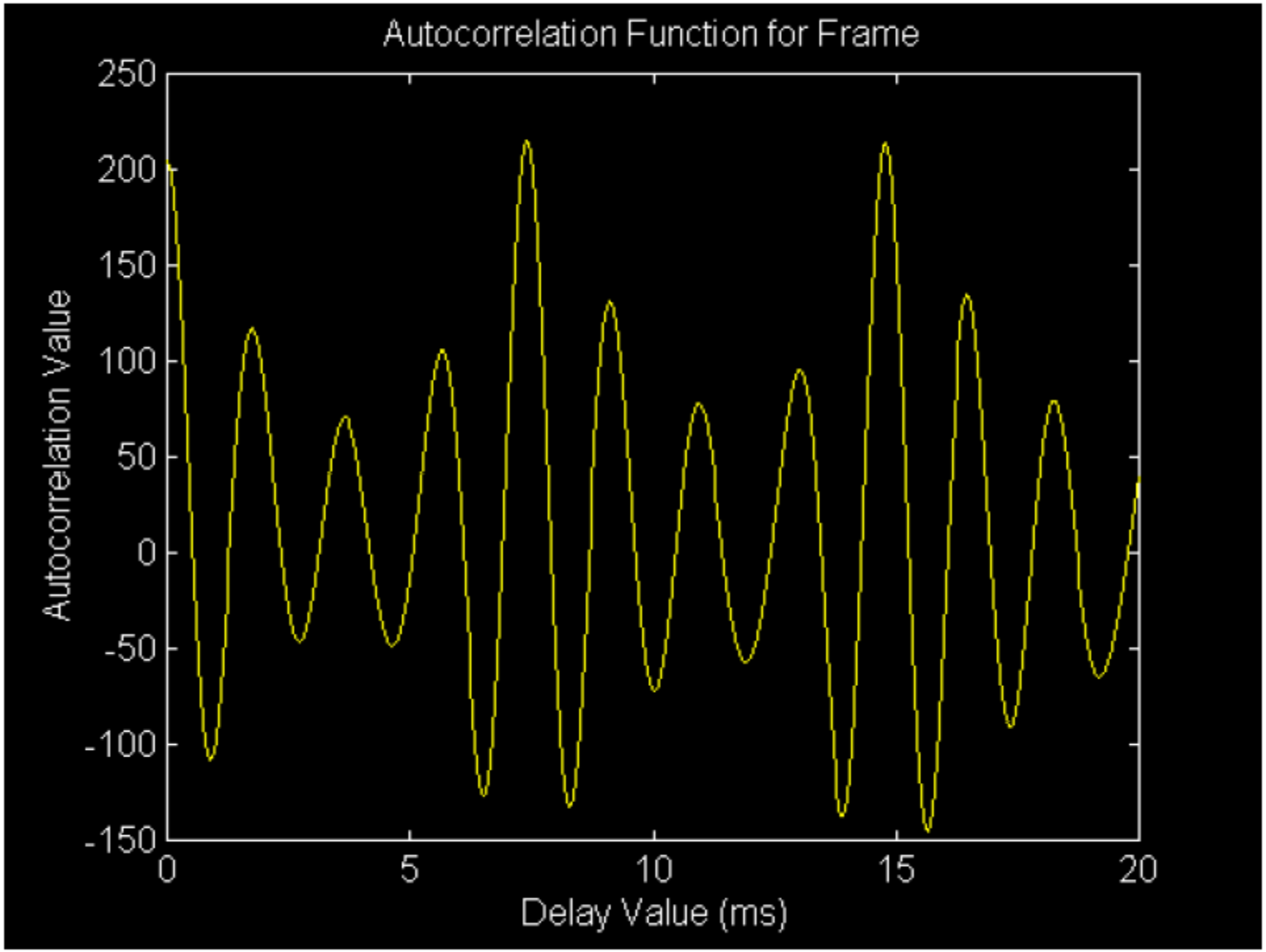
$$\phi(k) = \frac{1}{N} \sum_{n=0}^{N-1} s[n]s[n+k]$$

Autocorrelation Function

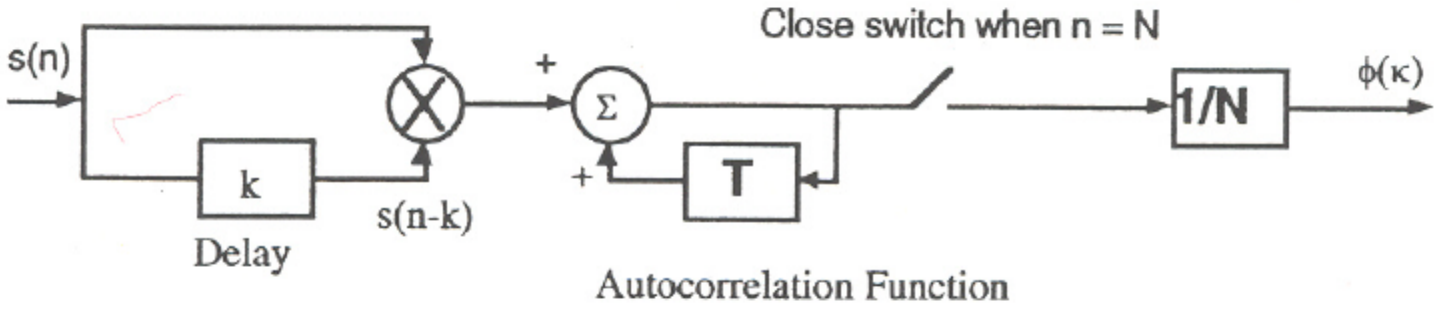
- Clearly, $\phi(k=0)$ would be equal to the average energy of the signal s[n] over the N sample frame
- If s[n] was perfectly periodic with a period of P samples then $s[n+P]=s[n]$
- Therefore, $\phi(k=P)=\phi(k=0)=$Average Energy
- While this is NOT exactly true for speech signals, the autocorrelation function with k equal to the pitch would result in a large value
- For the various k values between 0 and P, the various terms $(s[n]s[n+k])$ in the autocorrelation function would tend to be a mixture of positive and negative values
- These would tend to cancel each other out in the autocorrelation sum to yield very low values for $\phi(k)$

Autocorrelation Function

- This, for a given frame of N samples of VOICED speech, a plot of $\phi(k)$ versus k would exhibit distinct peaks at k values of 0, P, 2P …., where P is the pitch period
- The graph of $\phi(k)$ would be of quite small values between these peaks
- This pitch period for that frame is simply got by measuring the distance, in samples, between the peaks of the graphs of the autocorrelation function



A block diagram of the implementation of the autocorrelation function is shown below:



$$\phi(k) = \frac{1}{N}\sum_{n=0}^{N-1} s[n]s[n+k]$$

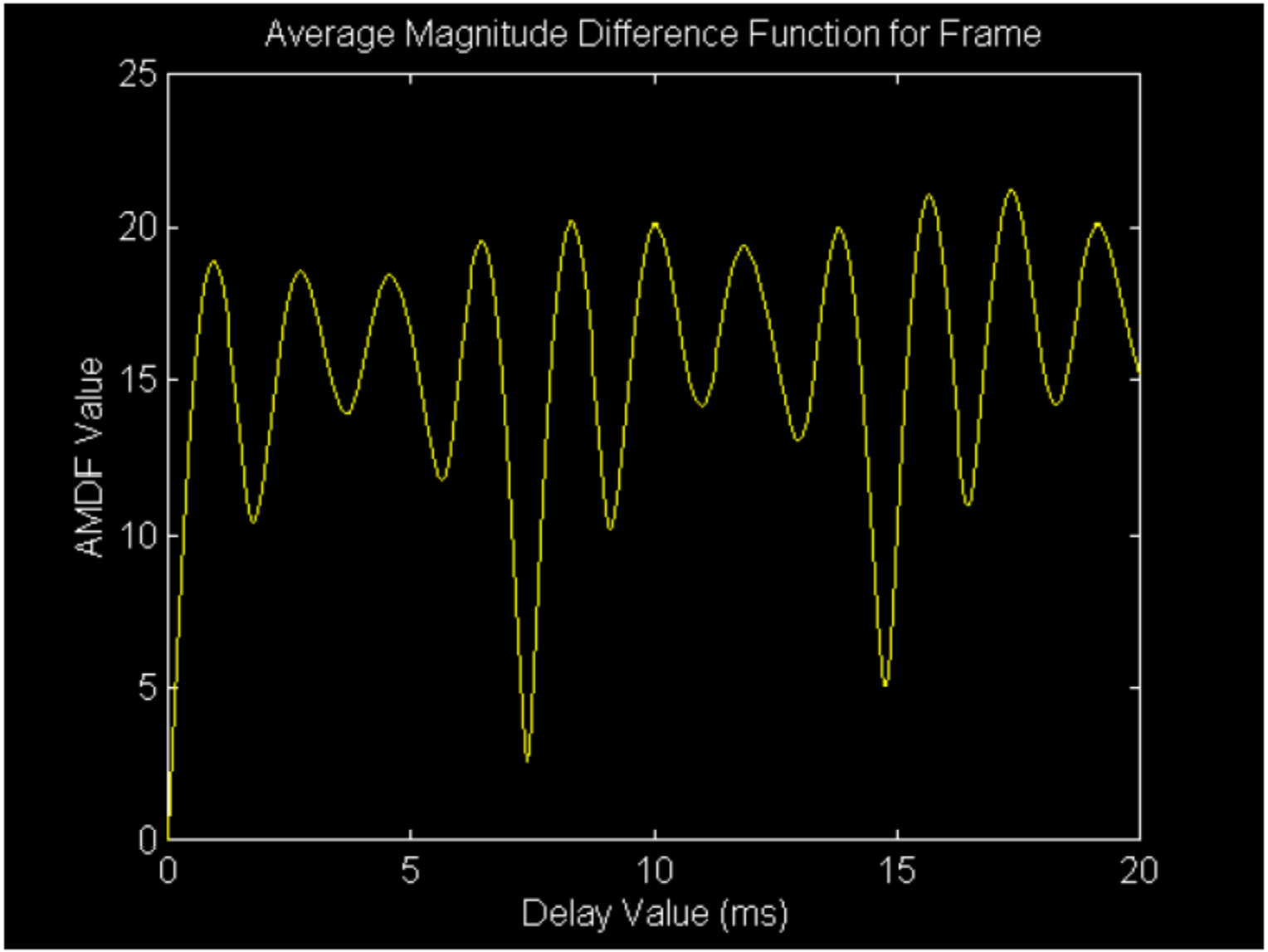
Average Magnitude Difference Function

- The AMDF is similar but opposite to the Autocorrelation Function
- For a delay of k samples, the AMDF is defined as

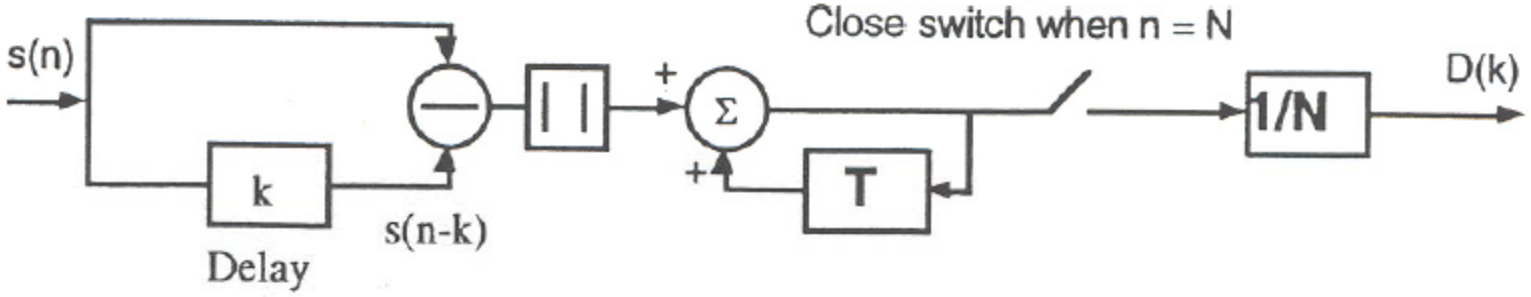
$$D(k) = \frac{1}{N} \sum_{n=0}^{N-1} | s[n] - s[n+k] |$$

Average Magnitude Difference Function

- For a given frame of VOICED speech, a plot of AMDF (D(k)) versus different values of delays (k), will exhibit deep "nulls" at k=0, P , 2P……
- If is used as an alternative to autocorrelation as on some processor architectures, it may be less computationally intensive to implement
- Care should be taken with both techniques to support the "overlap" into adjacent frames introduced by the the autocorrelation and AMDF



A block diagram implementation of the AMDF function:



$$D(k) = \frac{1}{N}\sum_{n=0}^{N-1} | s[n] - s[n+k] |$$

Pre-emphasis Filter

- Recall transfer function of vocal tract:

$$\frac{S(z)}{E(z)} = A_v \frac{1}{(1-z^{-1})^2} \frac{1}{1+\sum_{k=1}^{p} a_k z^{-k}} (1-z^{-1})$$

- There is an -6dB/octave trend as frequency increases .

- It is desirable to compensate for this by pre-processing the speech. This has an effect of cancelling out the effect of glottis and is known as pre-emphasis.

Pre-emphasis

- The high-pass filtering function can be achieved by use of the following difference equation:

$$y(n) = s(n) - as(n-1)$$

- Normally a is chosen between 0.9 and 1

Exercise: Pre-emphasis filter

1- Use MATLAB to plot the frequency response of a pre-emphasis filter with the following transfer function:

$$H(z) = 1 - 0.95z^{-1}$$

2 – Plot the spectra of a frame of speech before and after pre-emphasis filter has been applied?

Short-time Fourier Transfer - review

- Spectrogram maybe attained through use of STFT.
- FT is carried out on a short sequence of signal
- The signal maybe windowed e.g. Hamming window
- Overlapping should be also carried out.
- Following formula for calculating STFT with window w of length N:

$$STFT(k,b) = \sum_{m=0}^{N-1} w(m-b)s(m)e^{\frac{-j2\pi km}{N}}$$

STFT Exercise

1. Generate a signal composed of 4 tones of different frequencies
   - Two tones should be present constantly and other two tones occurring at different times.
   - signal should be about 1 sec in length in total and tones should have different levels.
2. Write a Matlab script to perform the STFT
   - include Hamming window
   - 50% overlapping of frames

3. Plot spectrogram of a signal

4. Investigate effect of
   - changing frame size
   - changing number of points in FFT
5. Record your own speech signal and generate spectrograms.

Exercises …

- Find and plot short-time energy of your recorded speech.
- Find and plot Short-Time Zero-Crossing Counts (ZCC) of your speech signal
- Use Auto-Correlation and Average Magnitude functions to Find and plot Fundamental frequency (Pitch period) of yourself.