21AIE315 AI IN SPEECH PROCESSING

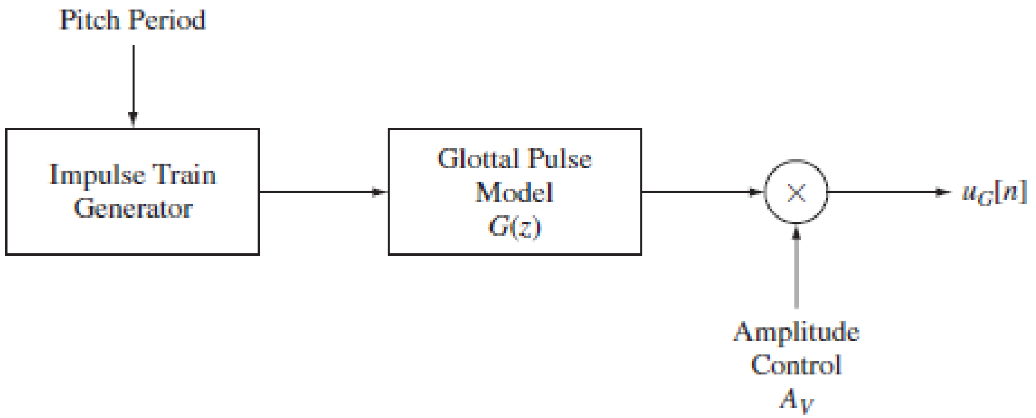
Time domain processing

Dr. Jyothish Lal G, Assistant Professor (Sr. Gr)

Department of AI , Amrita School of AI, Coimbatore

AMRITA
VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

Generation of the excitation signal for voiced speech.



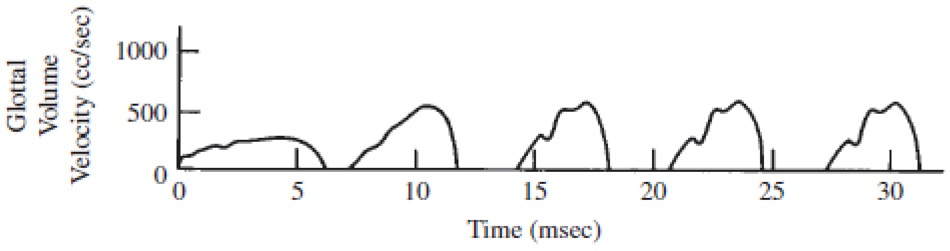
The impulse train generator produces a sequence of unit impulses that are spaced by the desired fundamental (pitch) period.

This signal, in turn, excites a linear system whose impulse response $g[n]$ has the desired glottal wave shape.

A gain control, A_v, controls the intensity of the voiced excitation.

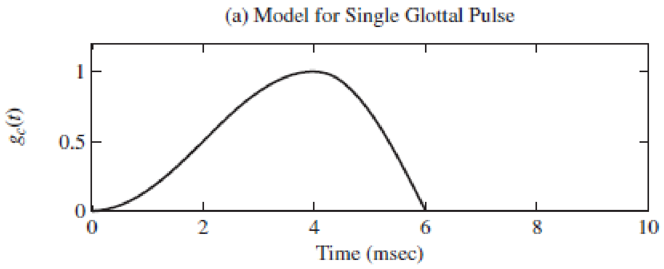
$$g[n] = g_c(nT) = \begin{cases} \frac{1}{2}[1 - \cos(\pi n/N_1)] & 0 \leq n \leq N_1 \\ \cos(\pi(n - N_1)/(2N_2)) & N_1 \leq n \leq N_1 + N_2 \\ 0 & \text{otherwise,} \end{cases}$$

where $N_1 = T_1/T$ and $N_2 = T_2/T$



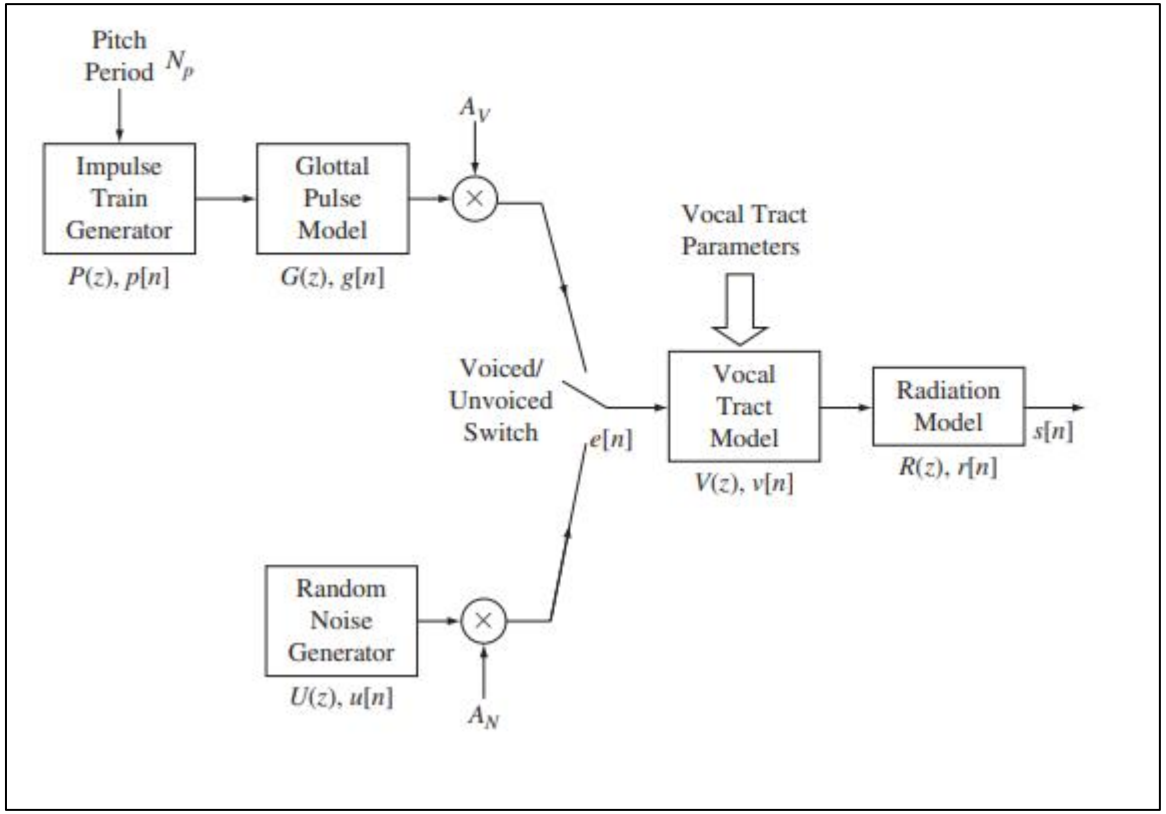
An example : Rosenberg's Glottal Pulse Approximation

$$g_c(t) = \begin{cases} 0.5[1 - \cos(2\pi t/(2T_1))] & 0 \leq t \leq T_1 \\ \cos(2\pi(t - T_1)/(4T_2)) & T_1 < t \leq T_1 + T_2, \end{cases}$$

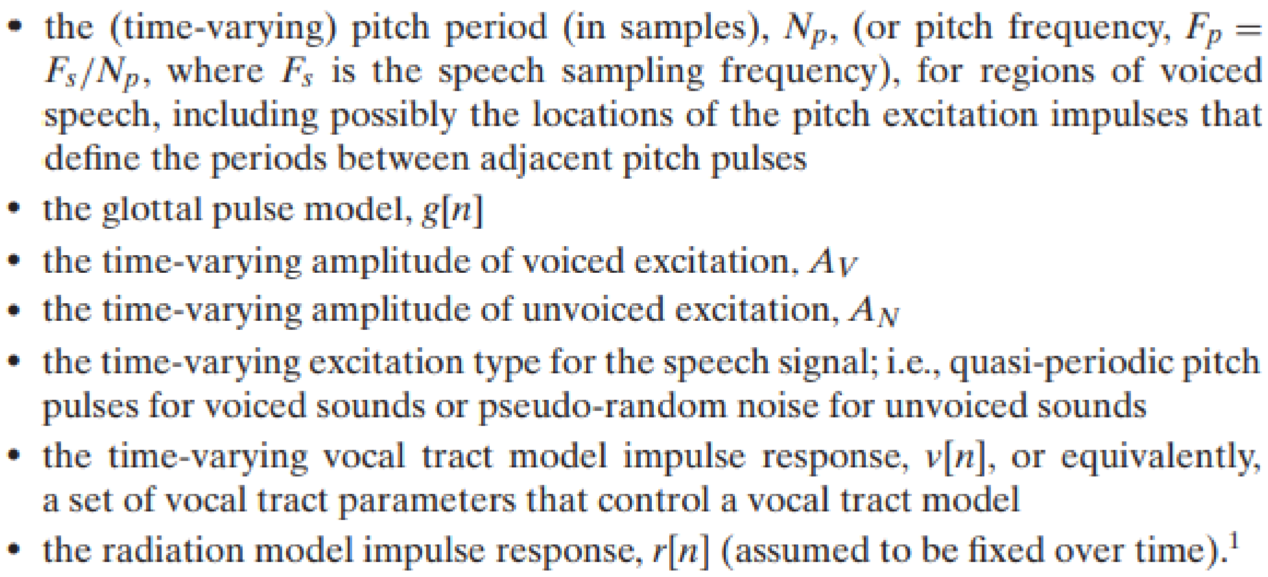


Here, $T_1 = 4$ msec and $T_2 = 2$ msec

Abstract Model for speech production and synthesis

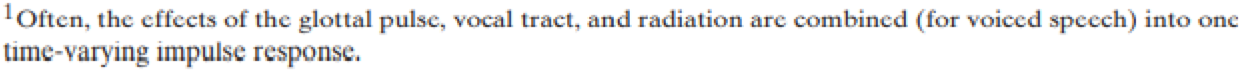


The information carried by the speech signal includes, but is not limited to, the following:



- the (time-varying) pitch period (in samples), N_p, (or pitch frequency, $F_p = F_s/N_p$, where F_s is the speech sampling frequency), for regions of voiced speech, including possibly the locations of the pitch excitation impulses that define the periods between adjacent pitch pulses
- the glottal pulse model, $g[n]$
- the time-varying amplitude of voiced excitation, A_V
- the time-varying amplitude of unvoiced excitation, A_N
- the time-varying excitation type for the speech signal; i.e., quasi-periodic pitch pulses for voiced sounds or pseudo-random noise for unvoiced sounds
- the time-varying vocal tract model impulse response, $v[n]$, or equivalently, a set of vocal tract parameters that control a vocal tract model
- the radiation model impulse response, $r[n]$ (assumed to be fixed over time).[1]

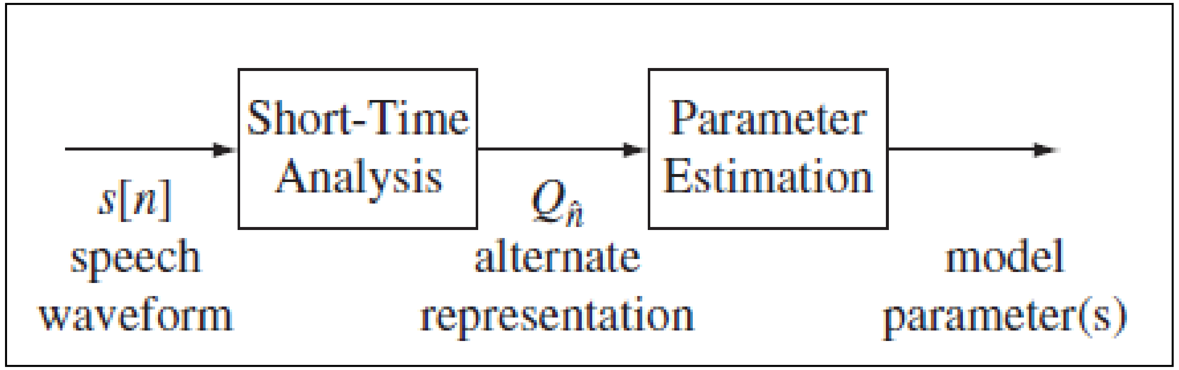
The goal of speech analysis is to estimate (as a function of time) parameters of a speech representation such as in Figure, and use them as a basis for an application such as a speech coder, a speech synthesizer, a speech recognizer, etc.



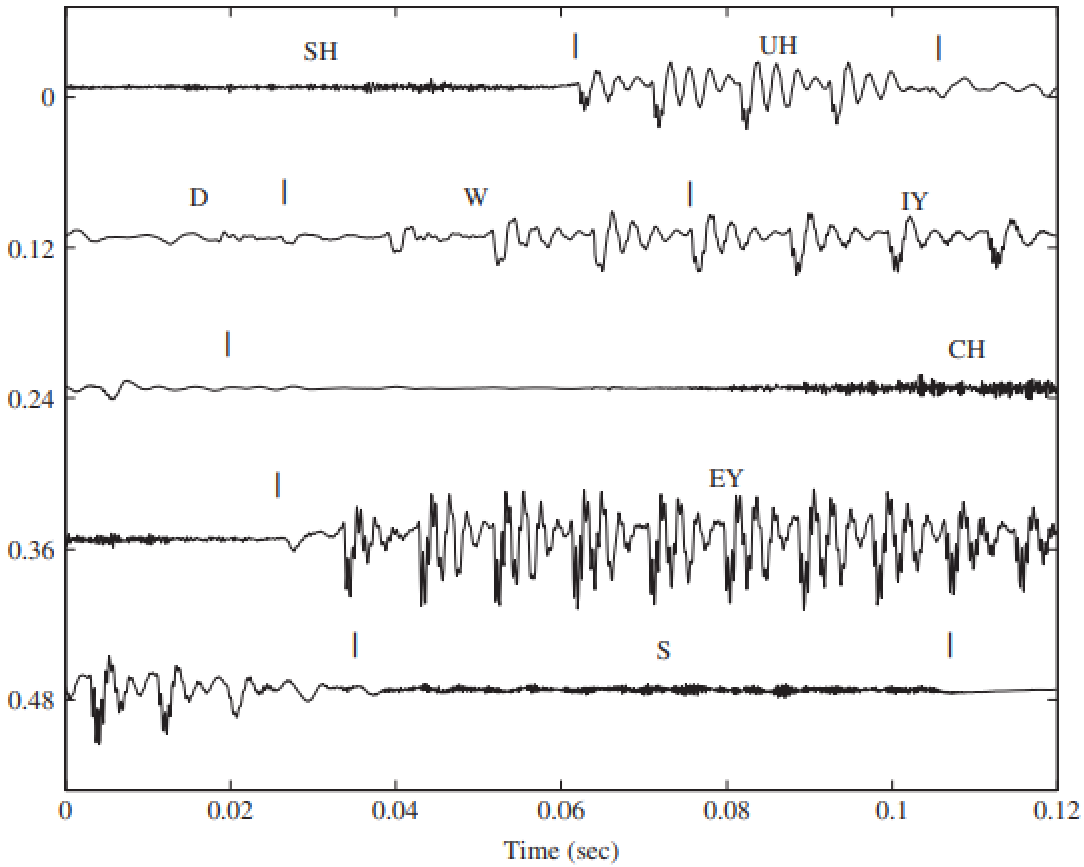
[1] Often, the effects of the glottal pulse, vocal tract, and radiation are combined (for voiced speech) into one time-varying impulse response.

Parametric representation

Digital signal processing for conversion of the speech waveform to an alternate representation that is more suitable for speech analysis; i.e., model parameter estimation.



Time domain : the alternate representation methods involve direct operations on the waveform of the speech signal.



See how the properties of the speech signal change slowly with time !

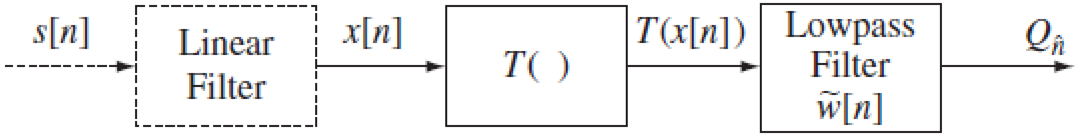
- the excitation mode changes
- there is a significant variation in the peak amplitude of the signal, and there is a steady variation of fundamental (pitch) frequency within voiced regions

Waveform of an utterance of /SH UH D - W IY - CH EY S/ (“should we chase”). The sampling rate is 10 kHz, and the samples are connected by straight lines. Each line in the plot corresponds to 1200 speech samples or 0.12 seconds of signal

Short-Time Processing of speech signal

- Localization: Speech is non-stationary in nature
- Finding parameters from these 20-30 ms blocks of speech segments is known as short-time processing of speech
- Short-Time processing of these blocks of data gives insights or time localized information of the signal

General Framework for Short-Time Analysis



$$\boxed{Q_{\hat{n}} = \sum_{m=-\infty}^{\infty} T(x[m])\tilde{w}[\hat{n}-m]}$$

Time Domain Parameters

- Short-time energy
- Short-time magnitude
- Short-time zero crossing
- Short-time auto correlation
- Short-time average magnitude difference function
- Linear prediction analysis (will be covered separately)

Short-time Energy

- Energy of a discrete-time signal $E = \sum_{m=-\infty}^{\infty} (x[m])^2$

- Computing total energy of blocks of speech signal

$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n}-m])^2 = \sum_{m=-\infty}^{\infty} (x[m])^2 \tilde{w}[\hat{n}-m]$$

where $w[\hat{n}-m]$ is a window that is applied directly to the speech samples before squaring, and $\tilde{w}[\hat{n}-m]$ is a corresponding window that can be applied equivalently

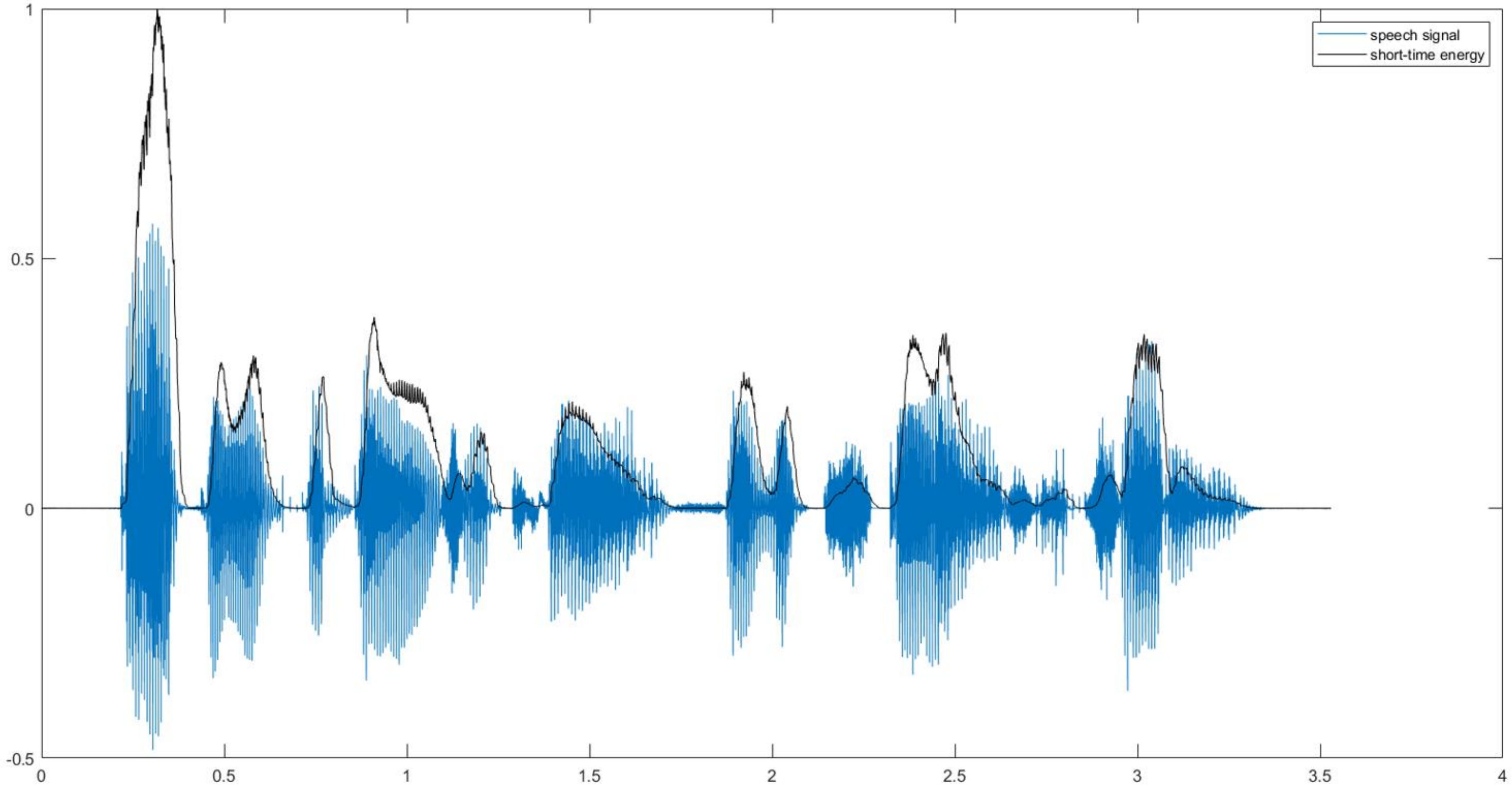
- For a rectangular window, $E_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} (x[m])^2$

Significance of short-time energy (STE):

- Sequence of ST energy values give time varying nature of the speech signal
- Voiced sound blocks: Having high energy
- Unvoiced sound blocks: Having low ST energy
- Silence: No or Negligible energy
- ST energy is used for voiced/unvoiced classification of speech
- STE feature is a versatile and widely used feature in speech processing, with applications in areas such as voice activity detection, speaker recognition, emotion recognition, speech segmentation, and speech enhancement

Short-Time Energy: Illustration

- The frame size and overlap(Full, Half or No overlap) must be specified for the computation



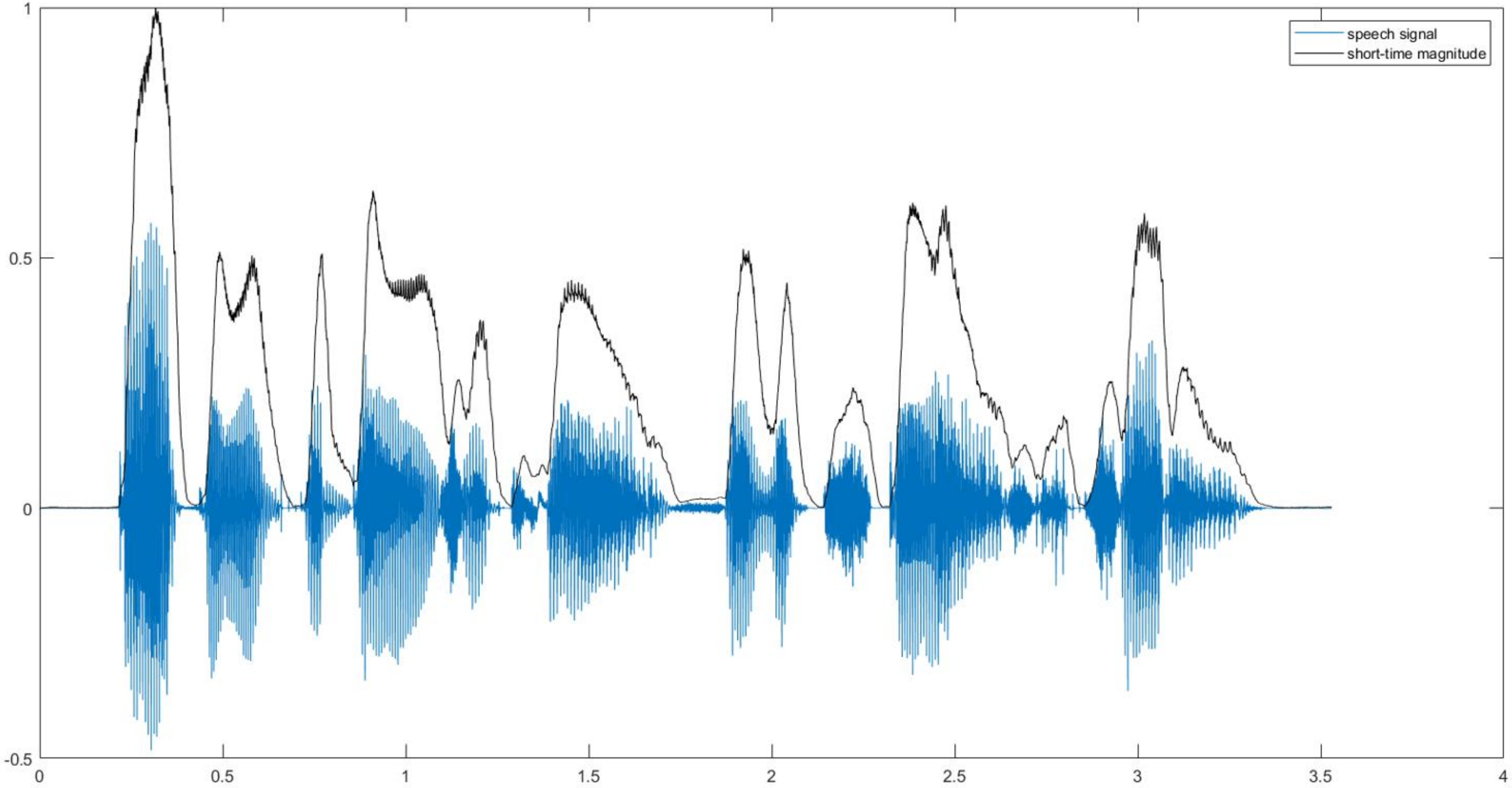
Variants of ST Energy

- Short-time magnitude and Square root S.T energy
- For example, Short-time magnitude function can be derived as follows

$$M_{\hat{n}} = \sum_{m=-\infty}^{\infty} |x[m]w[\hat{n}-m]| = \sum_{m=-\infty}^{\infty} |x[m]|\tilde{w}[\hat{n}-m].$$

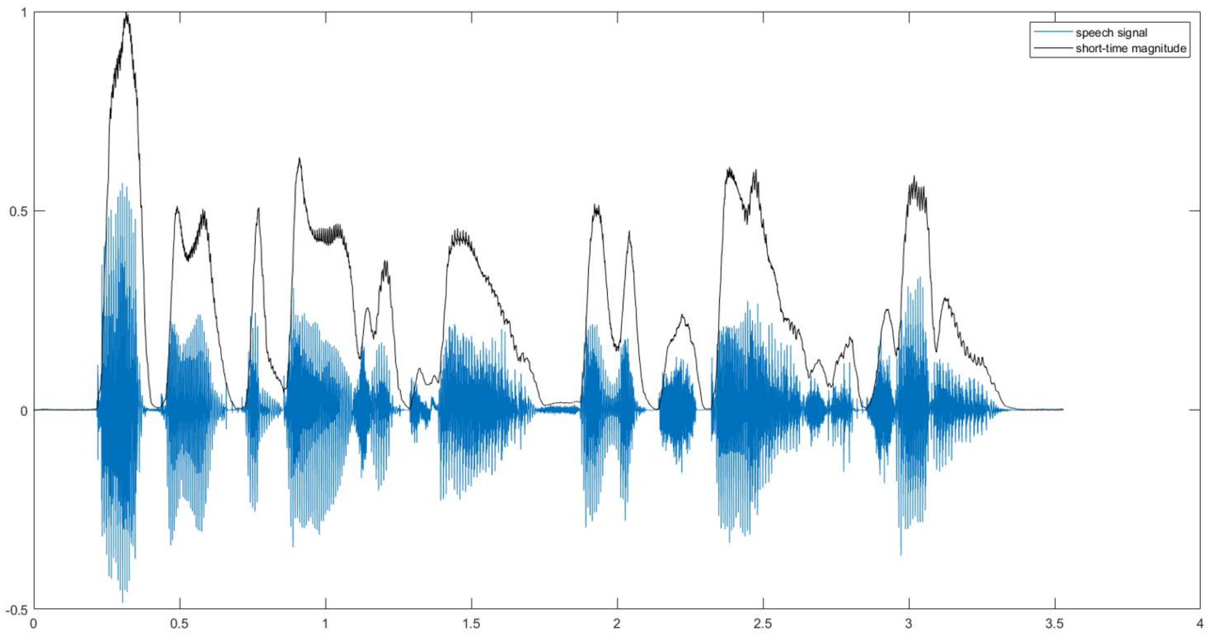
$$M_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} |x(m)|$$

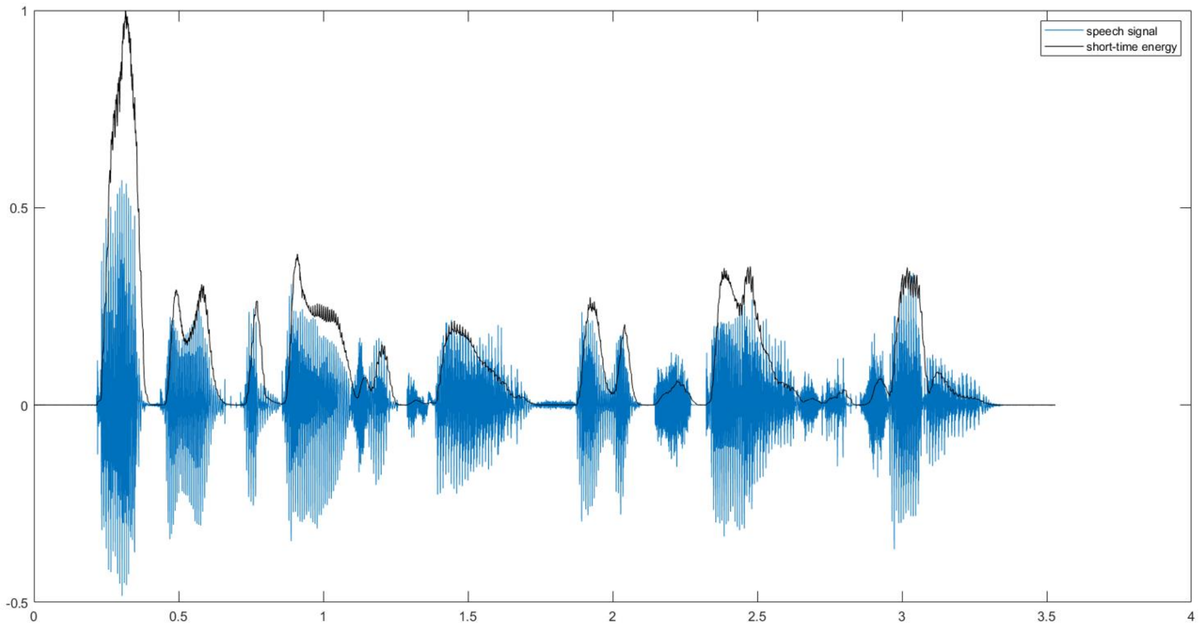
Short-time magnitude: Illustration



Comparison

The dynmaic range in STM is appr. the square root of that of STE





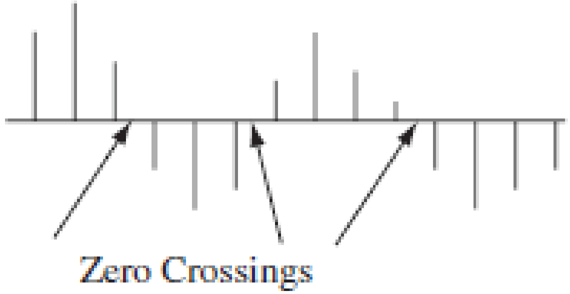
Short-time zero crossing rate

Assume sinusoidal signal of frequency *F*0, sampled at a rate *Fs*,

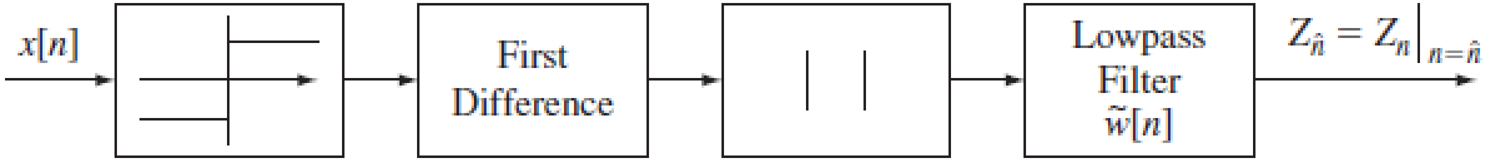
$$z^{(1)} = \frac{2\,\frac{crossings}{cycle} * F_o\,\frac{cycles}{sec}}{F_s\,\frac{samples}{sec}}$$

$z^{(1)} = 2F_o/F_s$ crossings/ sample

$z^{(M)} = 2MF_o/F_s$ crossings/M samples



Short-time zero crossing rate: Computation


$x[n]$
First Difference
| |
Lowpass Filter $\tilde{w}[n]$
$Z_{\hat{n}} = Z_n\big|_{n=\hat{n}}$

$$Z_{\hat{n}}^{(1)} = \frac{1}{2L_{eff}} \sum_{m=-\infty}^{\infty} \left| sgn\big(x(m)\big) - sgn\big(x(m-1)\big) \right| \tilde{w}(\hat{n}-m)$$

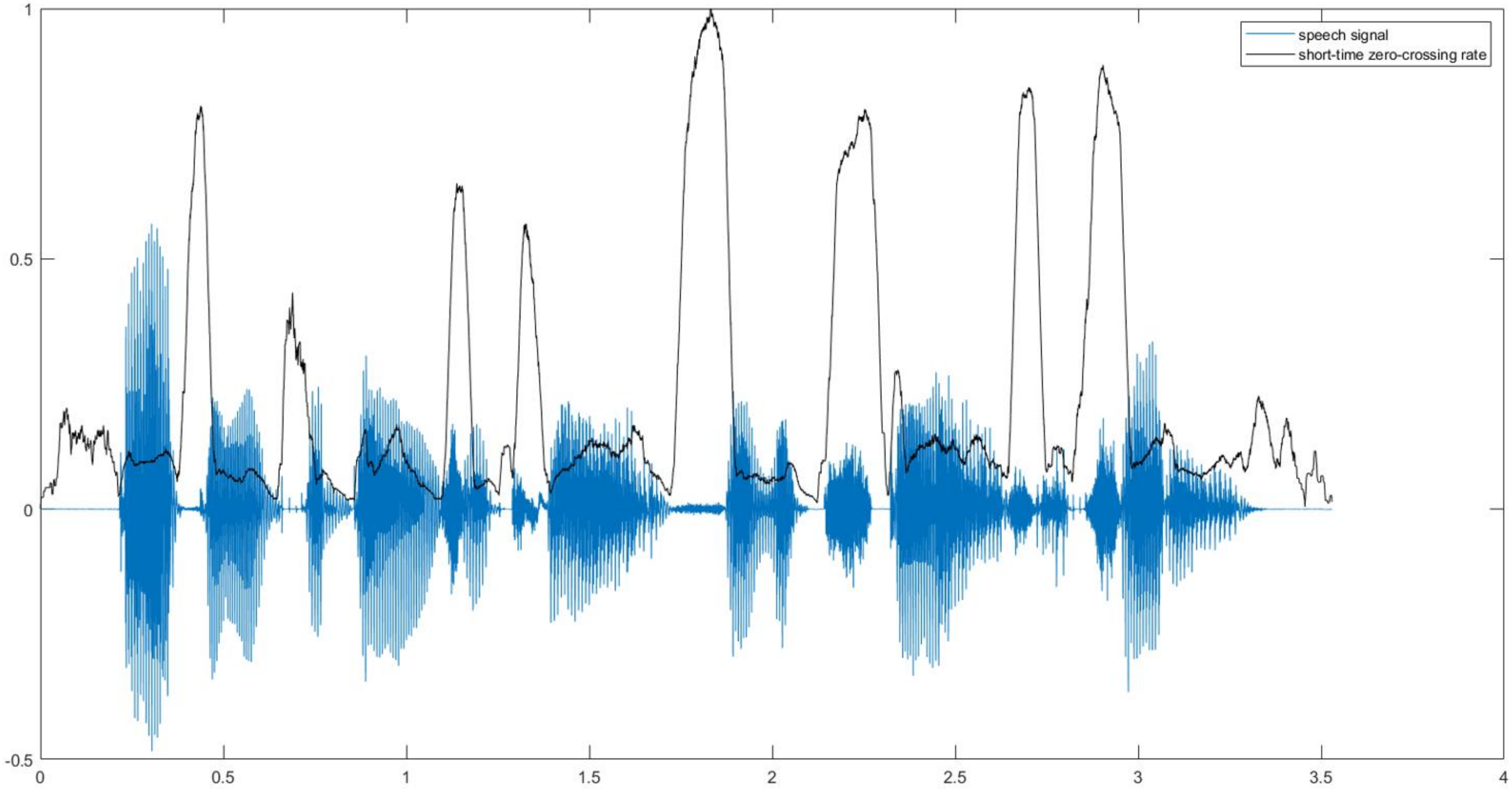
$$Z_{\hat{n}}^{(1)} = \frac{1}{2L} \sum_{m=\hat{n}-L+1}^{\hat{n}} \left| sgn\big(x(m)\big) - sgn\big(x(m-1)\big) \right|$$

where $sgn\big(x(m)\big) = \begin{cases} 1, & x(m) \geq 0 \\ -1, & x(m) < 0 \end{cases}$

Significance of ZCR

- No. of zero crossings→ No. of Sign changes
- ZCR is the ratio of total sign changes over a block to total no. of samples
- ZCR is high for unvoiced speech
- ZCR is low for voiced speech
- voiced/unvoiced classification can be performed using ZCR
- Used in speech segmentation tasks
- ZCR also provides gross information about the frequency contents
- More ZCR indicates high frequency content
- By analyzing the ZCR values in a speech signal, it is possible to determine the pitch contour and other prosodic features of the speech→ speech synthesis

Short-Time ZCR: Illustration



Short-Time Auto-Correlation Function

- Autocorrelation function helps in finding the self similarity (a measure of periodicity)
- Zero lag provides the total energy of the signal or block of signal
- The interval between central peak value and the second largest peak will be the period
- Hence useful in finding the pitch period

Short-Time Auto-Correlation: Computation

- Autocorrelation function of a discrete-time signal $\phi[k] = \sum_{m=-\infty}^{\infty} x[m]x[m+k]$

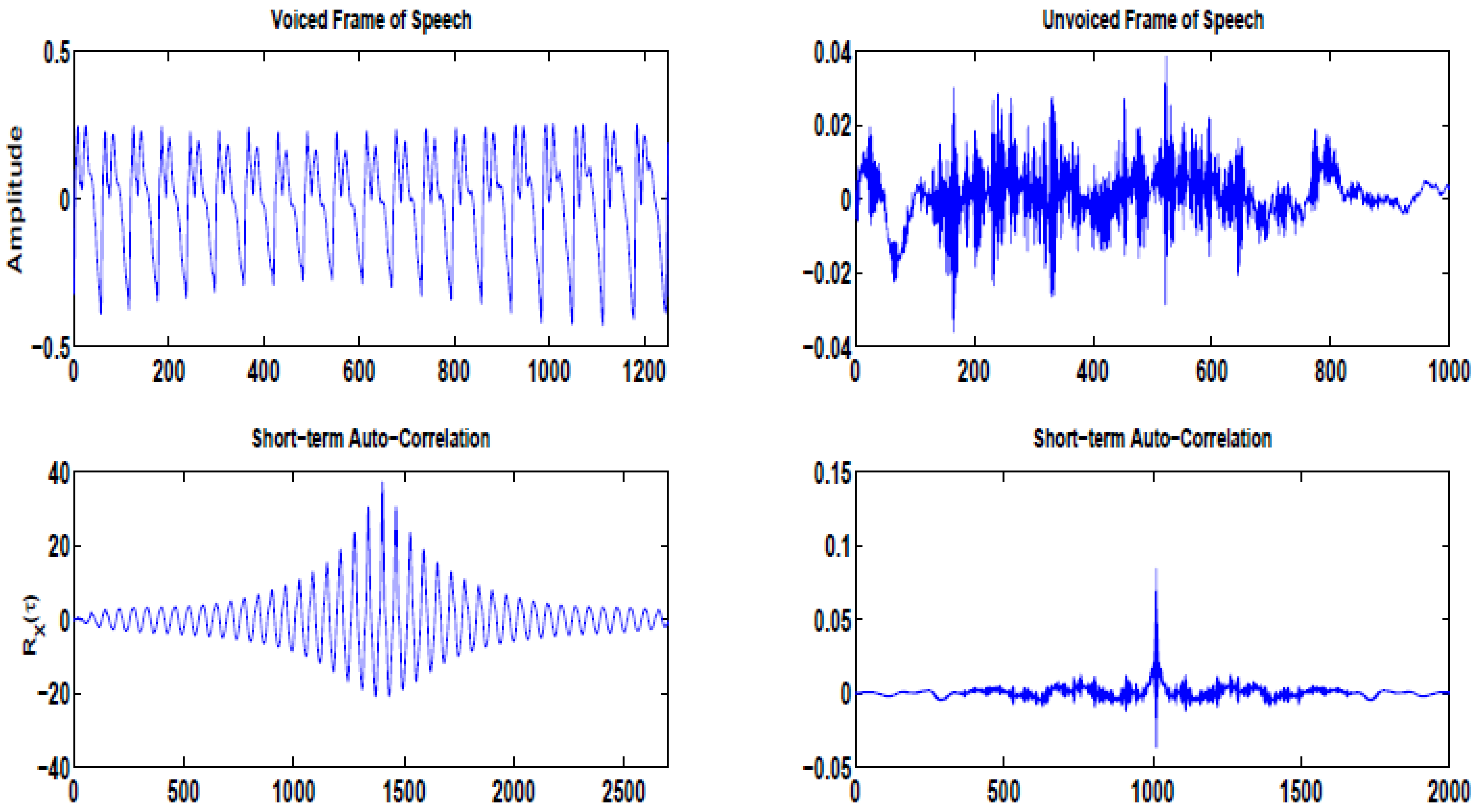
- Important property $\phi[k] = \phi[k + N_p]$; (For a signal is periodic with period *Np* samples)

- Autocorrelation function of the finite-length windowed segment of the speech $\big(x(m)w(\hat{n}-m)\big)$

$$R_{\hat{n}}(k) = \sum_{m=-\infty}^{\infty} \big(x(m)w(\hat{n}-m)\big)\big(x(m+k)w(\hat{n}-k-m)\big)$$

- When k=0, the R(0) computes total energy of the given frame

Short-Time Auto-Correlation: Illustration



Voiced Frame of Speech

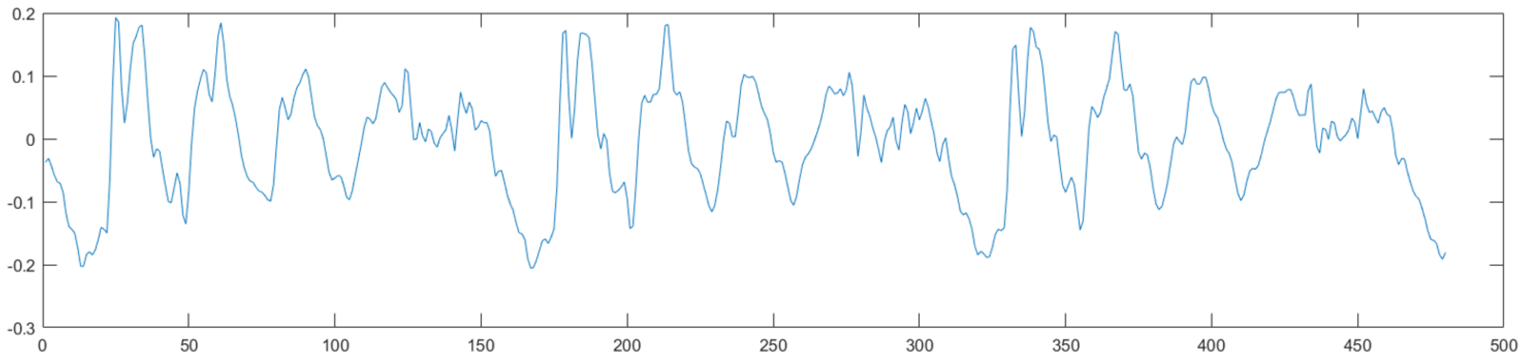
Unvoiced Frame of Speech

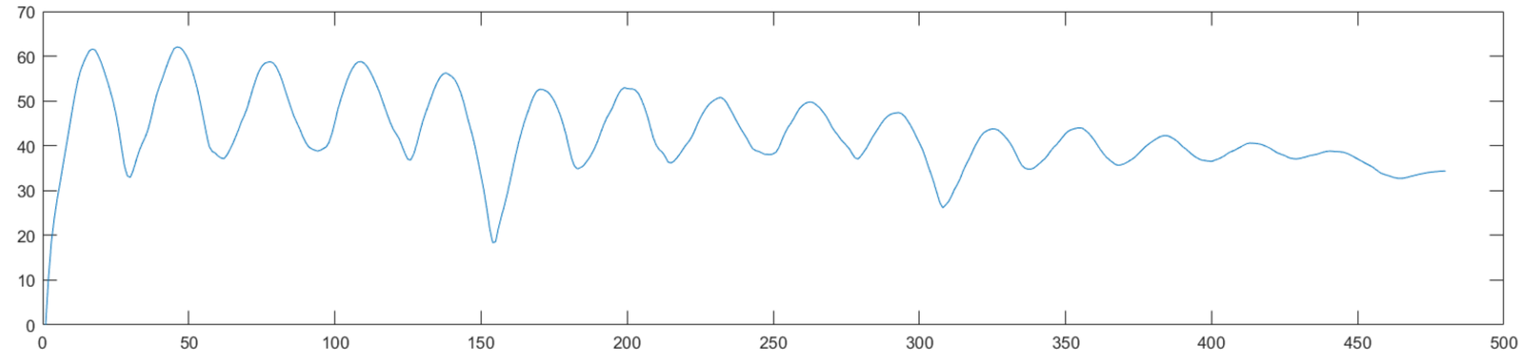
Short-term Auto-Correlation

Short-term Auto-Correlation

Short-Time Average Magnitude Difference Function

$$\gamma_{\hat{n}}(k) = \sum_{m=-\infty}^{\infty} \left| \left(x(\hat{n}+m)w(m)\right) - \left(x(\hat{n}+m-k)w(m-k)\right) \right|$$





Assignment 1.2 (Team-wise)

Assignment 1.1 will be uploaded soon in SharePoint drive.

- Extract time-domain parameters from a speech utterance without using inbuilt commands of librosa or any other similar library.
- Write down inferences about each features in the ipython notebook

(Matlab codes are provided for reference).

Submission link will be provided later

Thank you