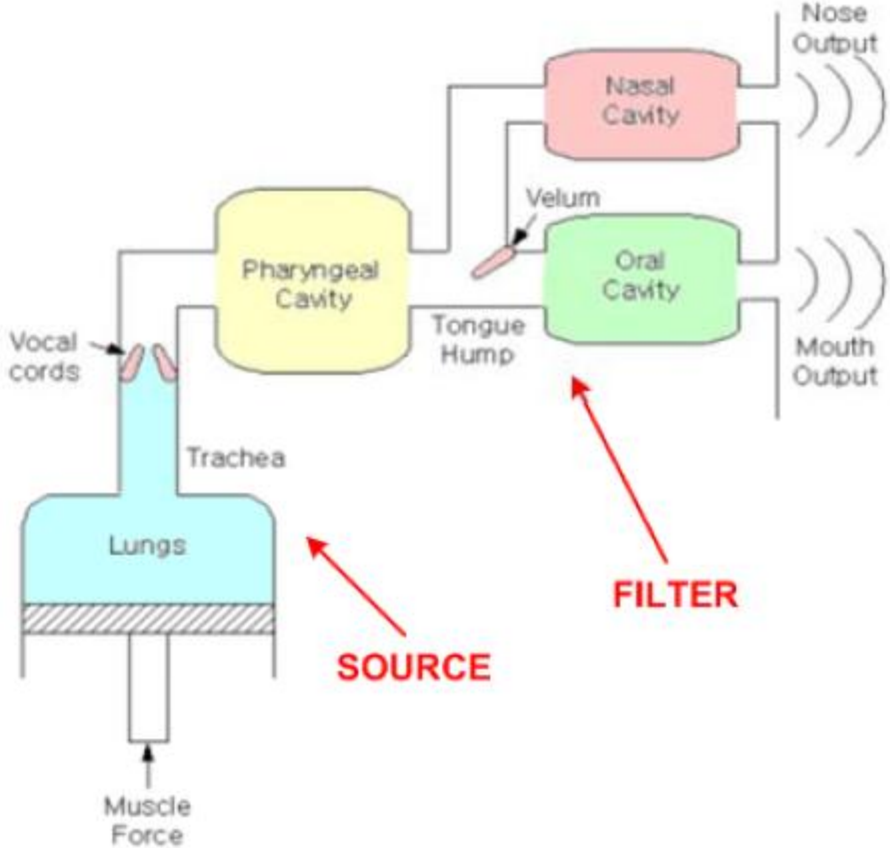
Model for Speech Production

- To develop an accurate model for how speech is produced, it is convenient to develop a digital filter based model of the human speech production mechanism
- Model must accurately represent :
  - The excitation mechanism of speech production system
  - The operation of the vocal tract
  - The lip\nasal radiation process
  - Both voiced & unvoiced speech for 10-20 ms

Source – Filter Model

Schematic diagram of the human speech production apparatus (Rabiner et al. 1993)



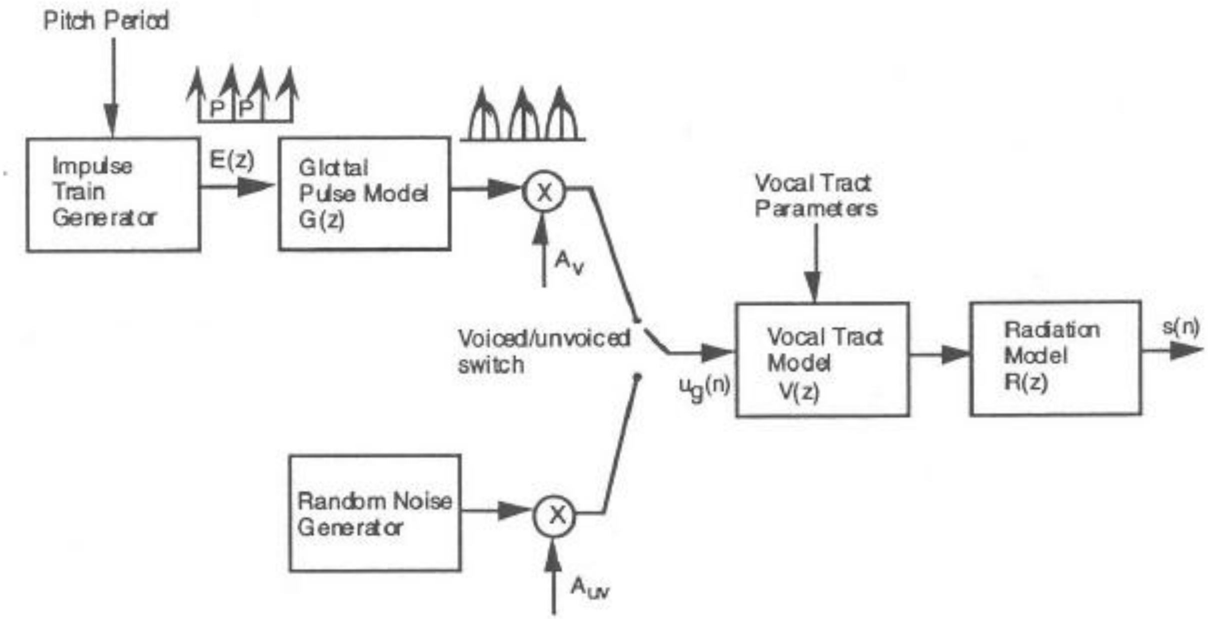
Rabiner, L & Juang, B (1993), *Fundamentals of speech recognition*, Prentice Hall, New Jersey.

Excitation Process

- The excitation process must take into account:-
  - The voiced\unvoiced nature of speech
  - The operation of the glottis
  - The “energy” of the speech signal

  in a given 10-30 ms frame of speech
- The nature of the excitation function of the model will be different dependent on the nature of the speech sounds being produced
  - For voiced speech, the excitation will be a train of glottal pulses spaced at intervals of the pitch period
  - For unvoiced speech, the excitation will be a random noise-like signal

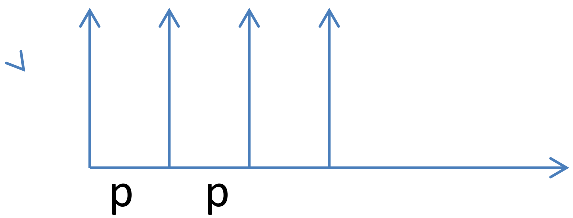
Discrete-Time Model for Speech Production



The model is a linear, time invariant model for the purposes of each 10-20ms frame interval where speech is considered stationary

Excitation Source – Voiced Speech

- Impulse train:



e(n) = δ(n-PK) , k=0,1,2...

E(z) = Z{e(n)}

$\sum_{n=-\infty}^{n=+\infty} e(n) z^{-n}$ = $\sum_{n=0}^{n=+\infty} e(n) z^{-n}$

$$E(z) = 1 + z^{-p} + z^{-2p} + \cdots$$

$$E(z) = \frac{1}{1 - z^{-p}}$$

Glottal Pulse Shaping Model

$$G(z) = \frac{1}{(1 - e^{-cT} z^{-1})^2}$$

Where, c: speed of sound

BUT, cT << 1, so $e^{-cT} \cong 1$

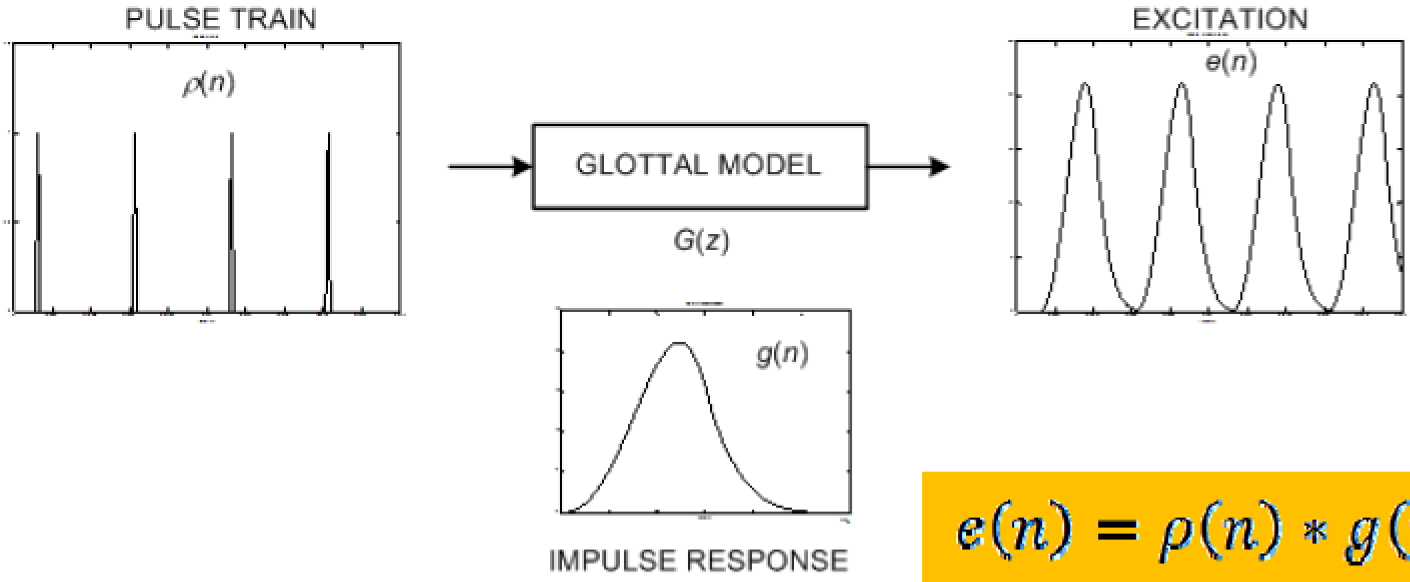
$$G(z) = \frac{1}{(1 - z^{-1})^2}$$

For voiced speech, G(z) = 1 for unvoiced speech

$$G(z) = \frac{1}{1 - z^{-1}} \cdot \frac{1}{1 - z^{-1}}$$

Two first-order LPF cascaded

Glottal Pulses – Voiced Source

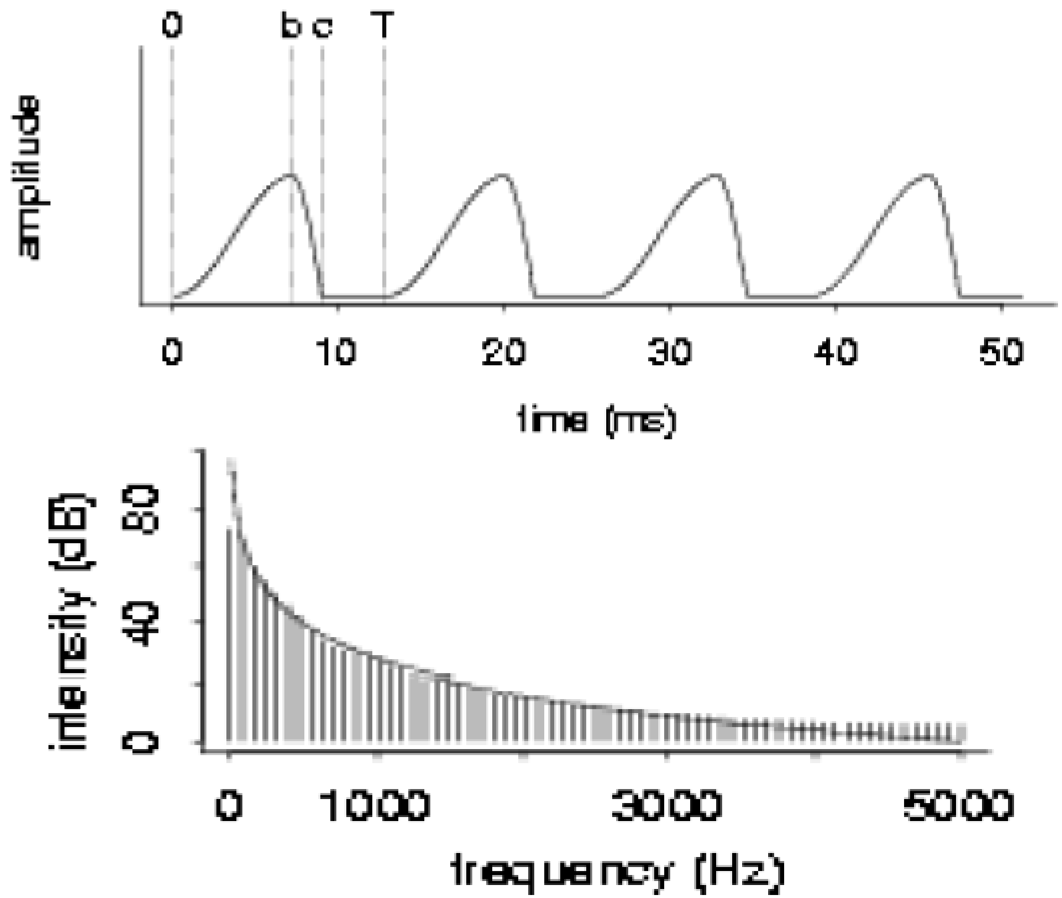


$$e(n) = \rho(n) * g(n)$$

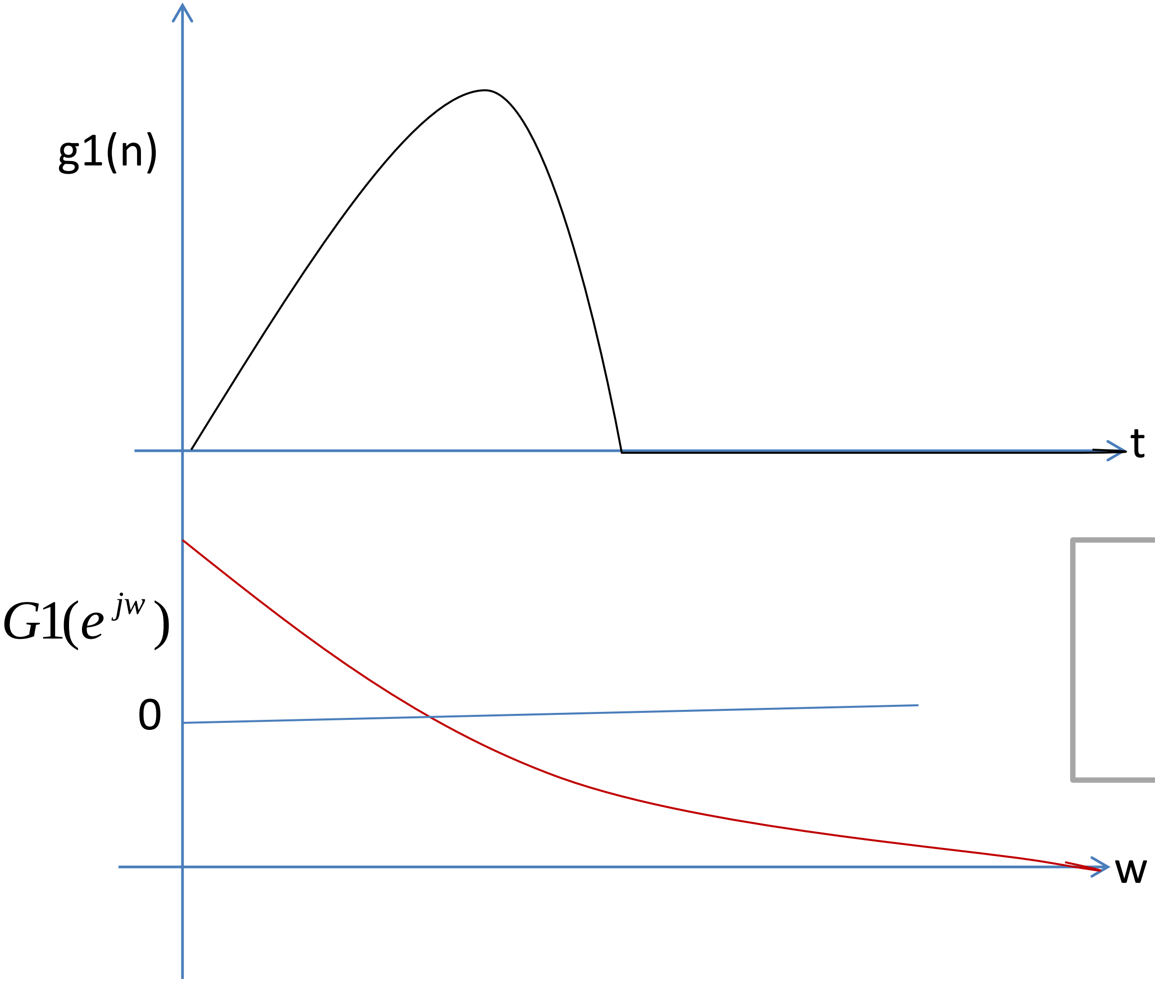
In the case of voiced speech, the glottal excitation can be further considered the result of the convolution of a train of impulses, separated by the pitch period.

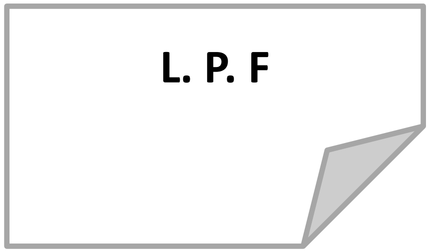
This is a filtering operation, where the impulse response of the filter is the single glottal waveform.

Glottal Pulses – Voiced Source



One Glottal pulse and its spectrum





Exercise: Glottal pulse and spectrum plot

The following expression can be used to model the glottal pulse. Write a Matlab script to plot the pulse and its spectrum.

(Assume N1= 40 and N2=10)

$$g(n) = \begin{cases} 0.5(1 - \cos\left(\frac{\pi n}{N1}\right), 0 \le n \le N1 \\ \cos\left(\frac{\pi(n-N1)}{2N2}\right), N1 \le n \le N1 + N2 \\ 0 \text{ , } otherwise \end{cases}$$

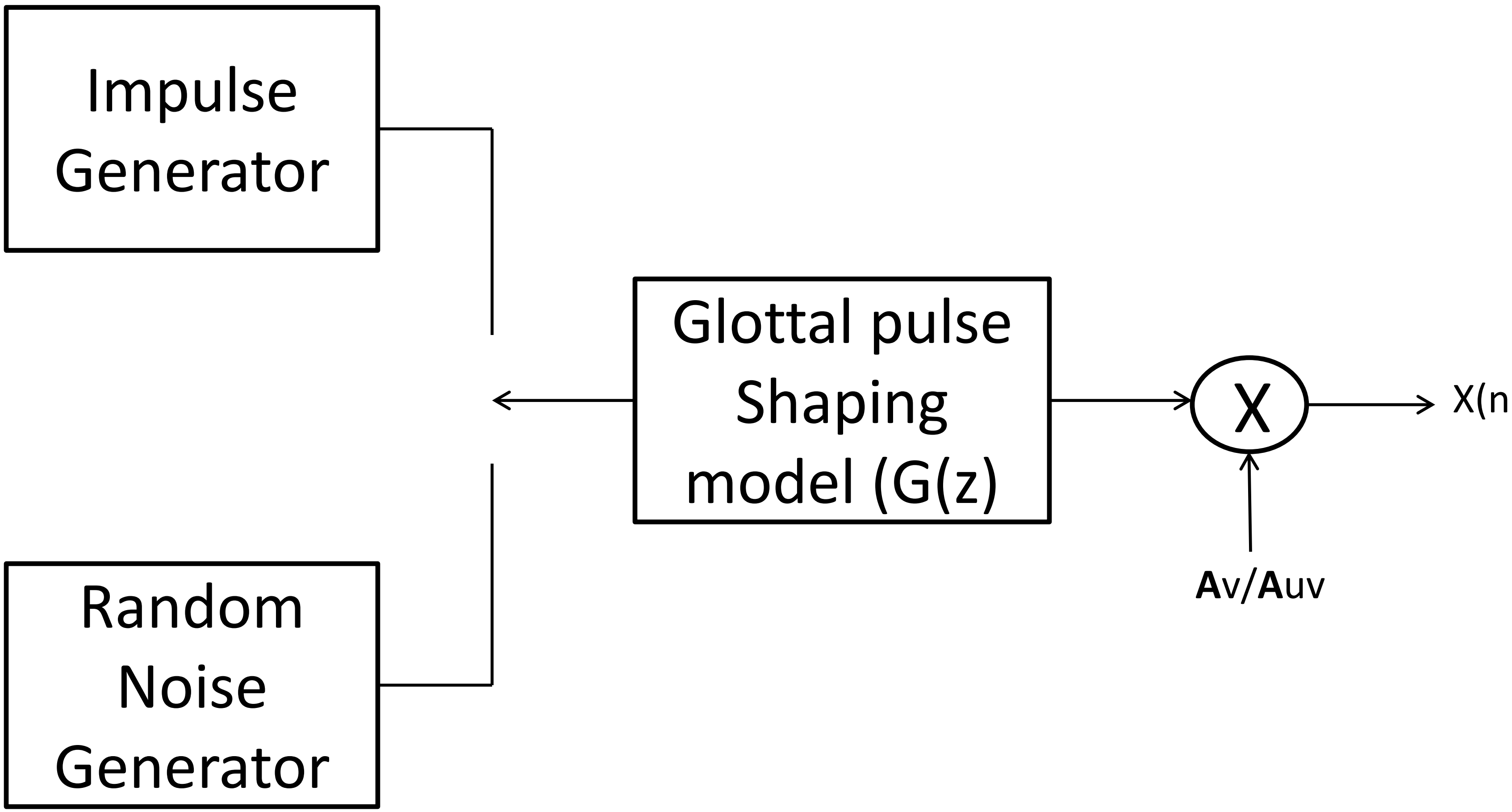
Excitation Process

- The ‘’energy’’ of the sound is modeled by a gain factor.
- Typically, gain factor of the voiced speech (A_v) is about 10 times that of unvoiced speech (A_{uv}).
- Thus the signal coming out of the complete excitation process will be:

$$x(n) = Ae(n) * g(n)$$

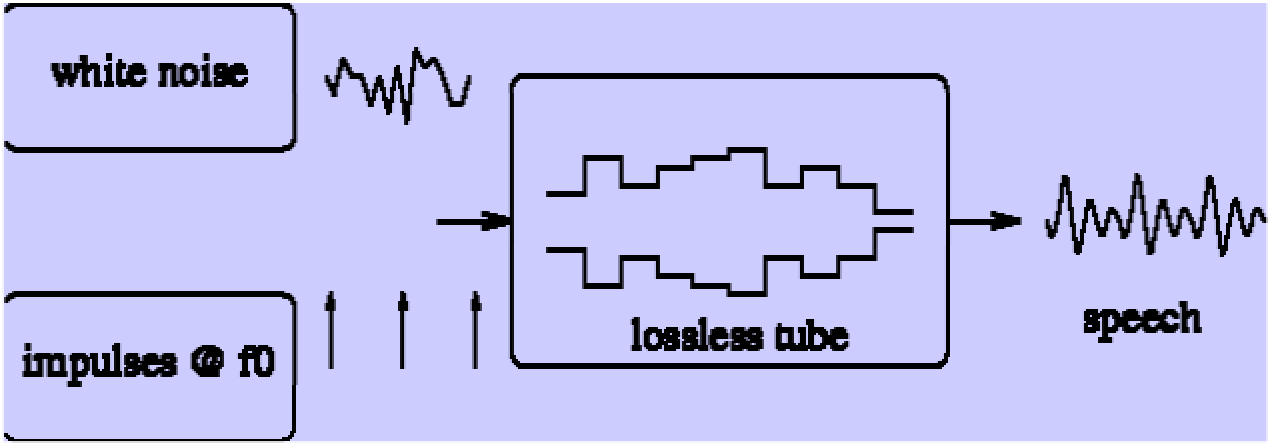
$$X(z) = AE(z)G(z)$$

Excitation Complete Model



Vocal Tract Model – The Filter

- The vocal tract can be modelled acoustically as a series of short cylindrical tubes



- Model consists of N lossless tubes each of length *l* and cross sectional area *A*
- Total length = NL
- Waves propagated down tube are partially reflected and partially junctions

Lossless Tubes Model

- τ is time taken for wave to propagate through single section

  $\boxed{\tau = l/c}$ ….c is speed of sound in air

- It has been shown that to represent the vocal tract by a discrete time system it should be sampled every 2τ seconds

  $$\boxed{fs = 1/2\,\tau \quad = \quad c/2l \quad = \quad Nc/2L}$$

- Thus fs is proportional to number of lossless tubes

- Recall length of vocal tract is about 17cm

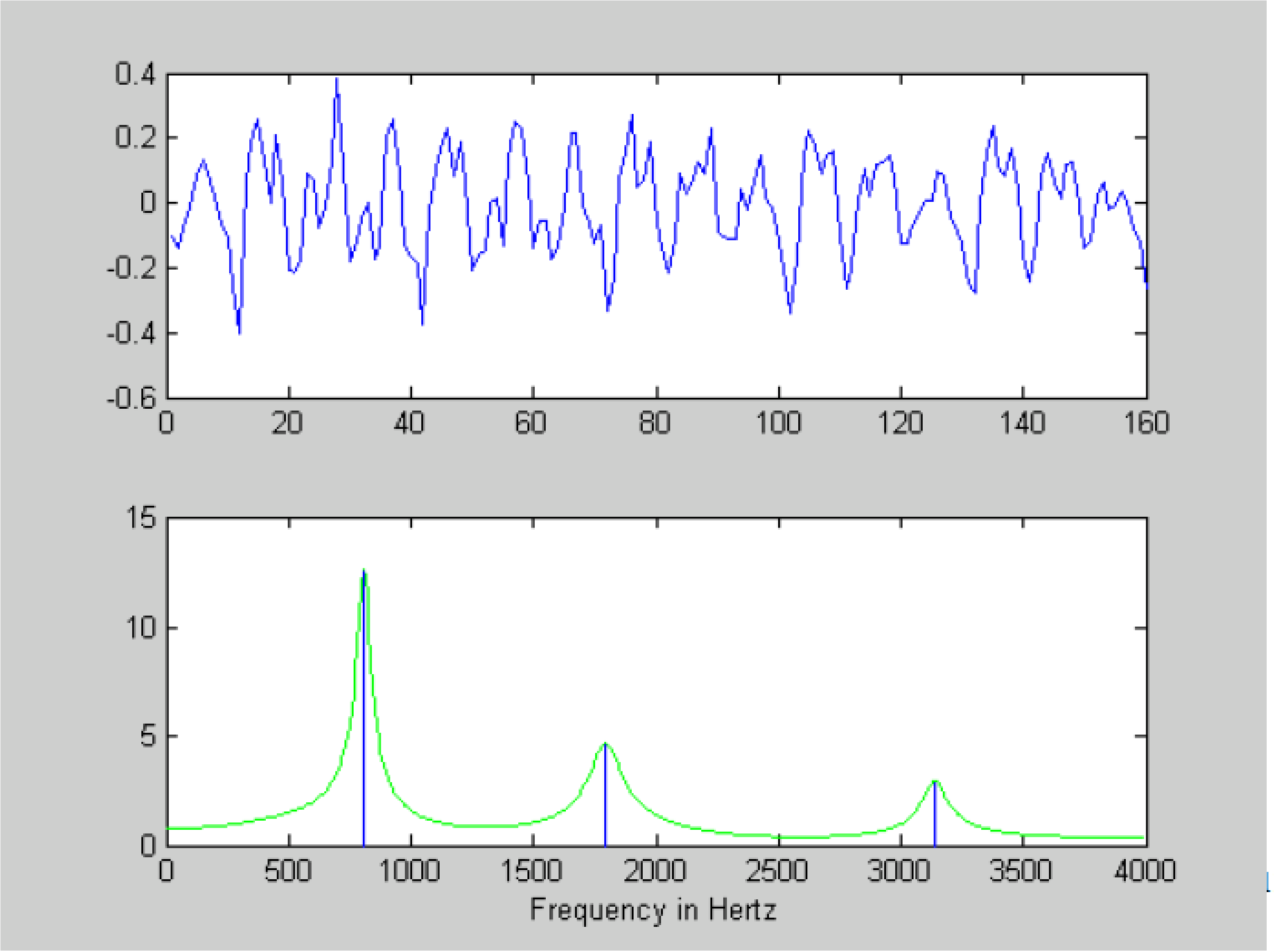
Vocal Tract Model

- This acoustic model can be converted into a time varying digital filter model
- For either voiced or unvoiced speech, the underlying spectrum of the vocal tract will exhibit distinct frequency peaks
- These are known as the FORMANT frequencies of the vocal tract
- Ideally, the vocal tract model should implement at least three or four of the formants

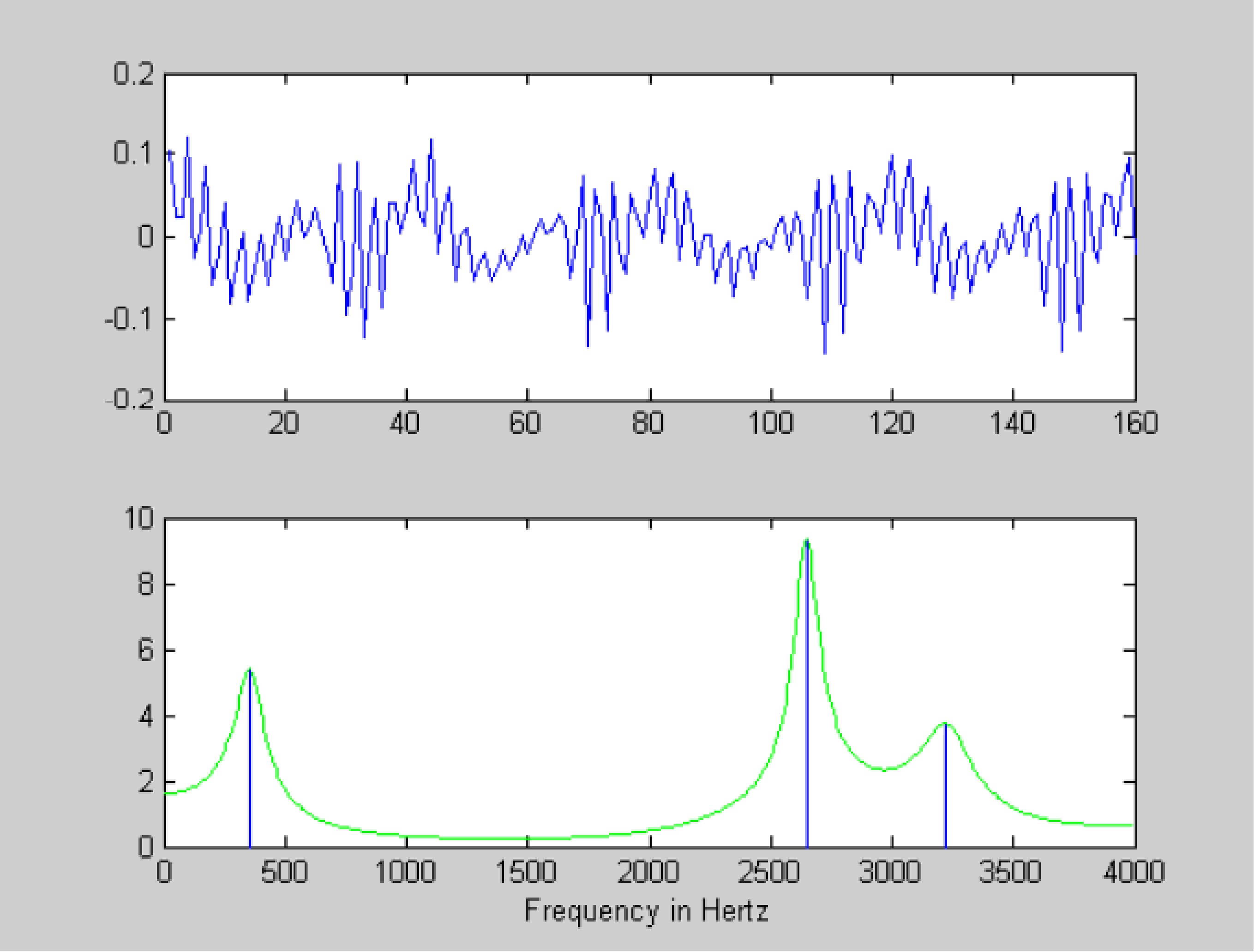
Formant Frequencies

- Speech normally exhibits one formant frequency in every 1 kHz
- For VOICED speech, the magnitude of the lower formant frequencies is successively larger than the magnitude of the higher formant frequencies
- For UNVOICED speech, the magnitude of the higher formant frequencies is successively larger than the magnitude of the lower formant frequencies

Voiced Speech



Unvoiced Speech



Vocal Tract Model – Voiced Speech

- For voiced speech, the vocal tract model can be adequately represented by an "all pole" model
- Typically, two poles are required for each resonance, or formant frequency
- The all-pole model can be viewed as a casacade of 2nd order resonators (2 poles each)
- Thus, the transfer function for the vocal tract will be

$$V(z) = \frac{U_l(z)}{U_g(z)} = \frac{1}{\prod_{k=1}^{K} 1 + b_k z^{-1} + c_k z^{-2}} = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}}$$

Vocal Tract Model – Unvoiced Speech

- Because of the nature of the turbulent air flow which creates unvoiced speech, the vocal tract model requires both poles and zeroes for unvoiced speech
- A single zero in a transfer function can be approximated by TWO poles
- Thus the transfer function for the vocal tract will be:

$$V(z)=\frac{1+\sum_{k=1}^{L} b_k z^{-k}}{1+\sum_{k=1}^{P} a_k z^{-k}} \approx \frac{1}{1+\sum_{k=1}^{P+2L} a_k z^{-k}}$$

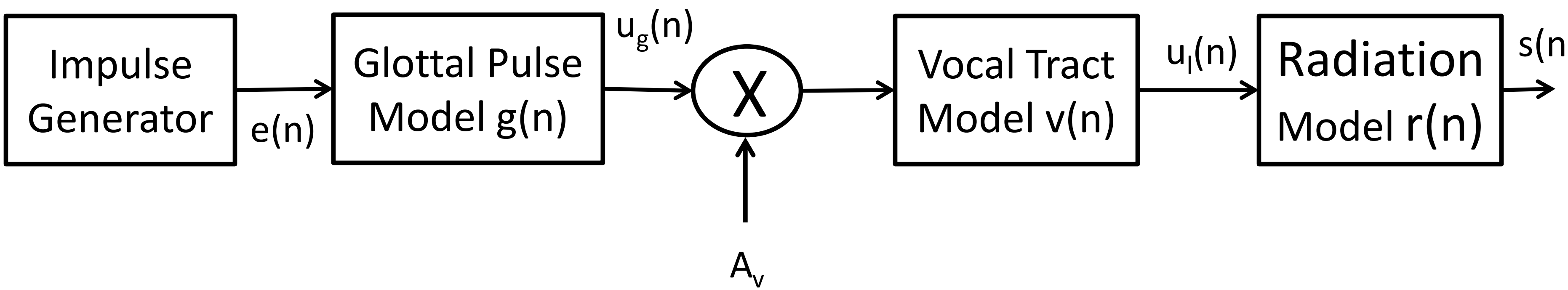
Exercise: 2nd Order Pole Approximation to zeros

- Show that of |a| < 1

$$1 - az^{-1} = \frac{1}{\sum_{n=0}^{n=\infty} a^n z^{-n}}$$

And thus a zero can be approximated as closely as desired by two poles

Discrete-time Model for Voiced speech Production



$$u_g(n) = e(n) * g(n)$$

$$u_l(n) = A_v . u_g(n) * v(n)$$

$$s(n) = A_v . e(n) * g(n) * v(n) * r(n)$$

$$S(z) = A_v E(z) G(z) V(z) R(z)$$

Vocal Tract Model – Unvoiced Speech

- Because of the nature of the turbulent air flow which creates unvoiced speech, the vocal tract model requires both poles and zeroes for unvoiced speech
- A single zero in a transfer function can be approximated by TWO poles
- Thus the transfer function for the vocal tract will be:

$$V(z) = \frac{1 + \sum_{k=1}^{L} b_k z^{-k}}{1 + \sum_{k=1}^{P} a_k z^{-k}} \approx \frac{1}{1 + \sum_{k=1}^{P+2L} a_k z^{-k}}$$

Lip Radiation Model

- The volume velocity at the lips is transformed into an acoustic pressure waveform some distance away from the lips.

- The typical lip radiation model used is that of a simple high pass filter, with the transfer function:

$$R(z)=1-z^{-1}$$

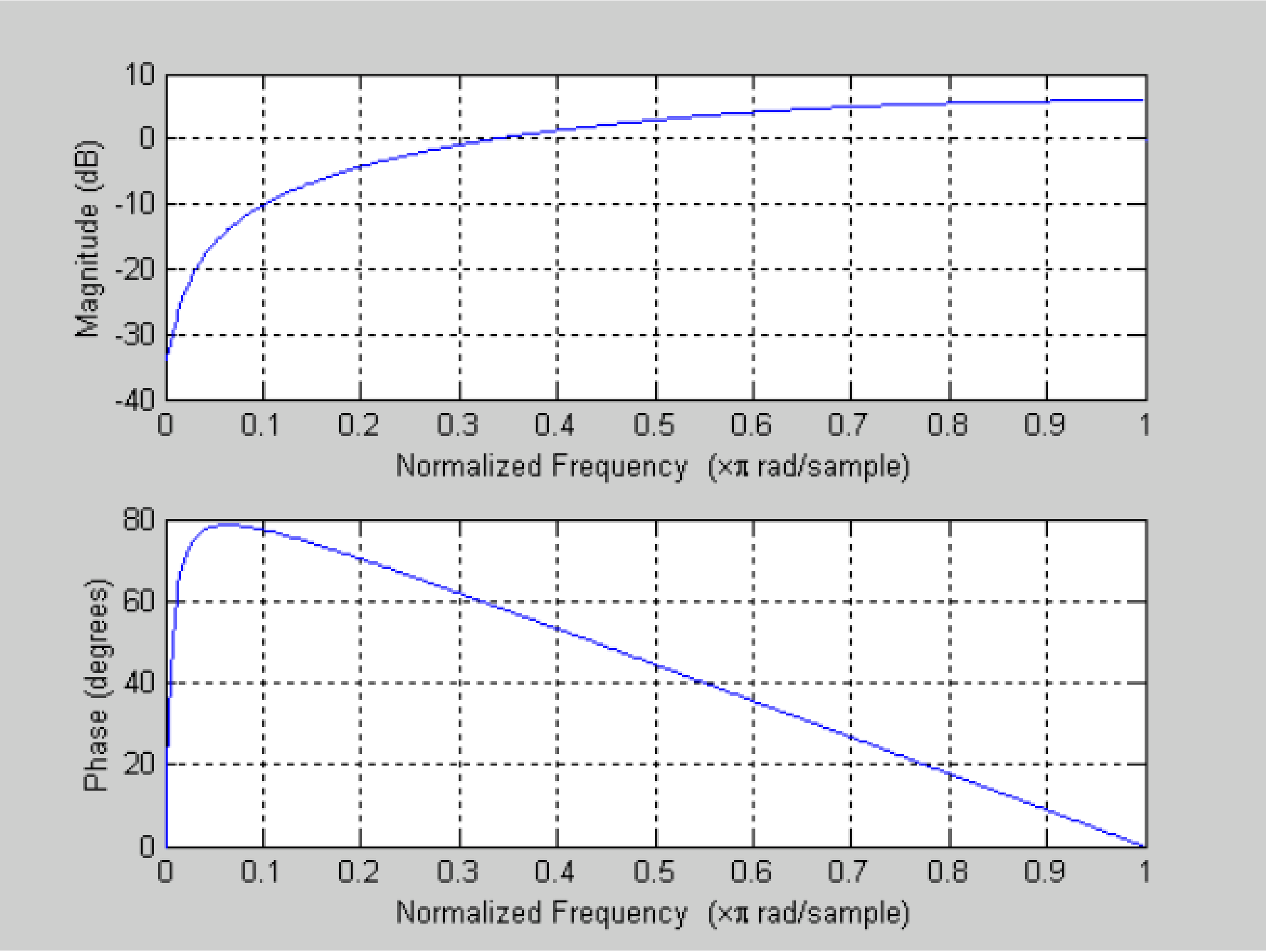
Exercise: Lip Radiation Model

- The following is an approximation to the lip radiation model.

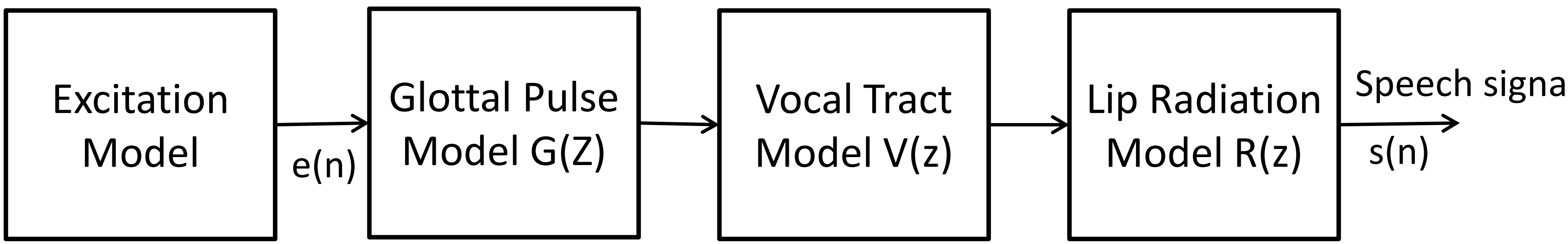
$$R(z)=1-0.98z^{-1}$$

- Use Matlab to plot the frequency response, $|R(\theta)|$ of the model

Frequency Response of Lip Radiation Model



Overall Speech Production Model



S(z) = E(z) G(z) A V(z) R(z)

Transfer Function:

$$\frac{S(z)}{E(z)} = AG(z)V(z)R(z)$$

Overall Transfer Function

- For Voiced Speech:

$$\frac{S(z)}{E(z)} = AvG(z)V(z)R(z)$$

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

$$\frac{S(z)}{E(z)} = Av\frac{1}{1-z^{-1}}.\frac{1}{1+\sum_{k=1}^{p}a_k z^{-k}} = \frac{Av}{1+\sum_{k=1}^{p+1}a_k' z^{-k}}$$

Overall Transfer Function

- For Unvoiced Speech:

$$\frac{S(z)}{E(z)} = A_{uv} G(z) V(z) R(z)$$

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

$$\frac{S(z)}{E(z)} = A_{uv} . \frac{(1 - z^{-1})}{1 + \sum_{k=1}^{p+2L} a_k z^{-k}} = \frac{A_{uv}}{1 + \sum_{k=1}^{p+2L+2} a_k ' z^{-k}}$$

Overall Transfer Function

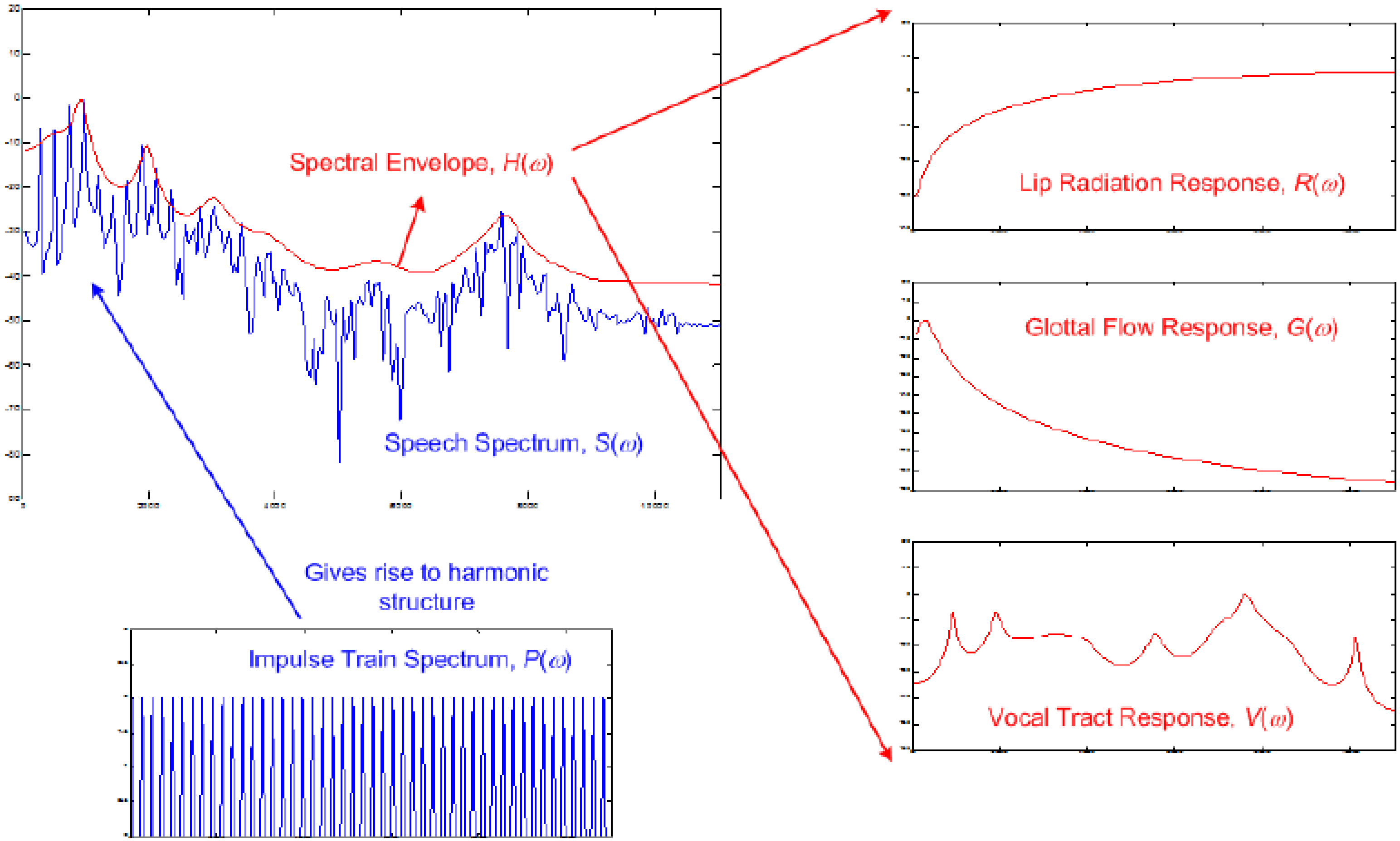
- Clearly, for EITHER form of speech sound, the model exhibits a transfer function of the form:

$$\frac{S(z)}{E(z)} = \frac{G}{1 + \sum_{k=1}^{q} a_k ' z^{-k}}$$

- It is simply a matter of selecting the order of the model (q) such that it is sufficiently complex to represent both voiced and unvoiced speech frames

- Typical values of q used are 10, 12, or 14

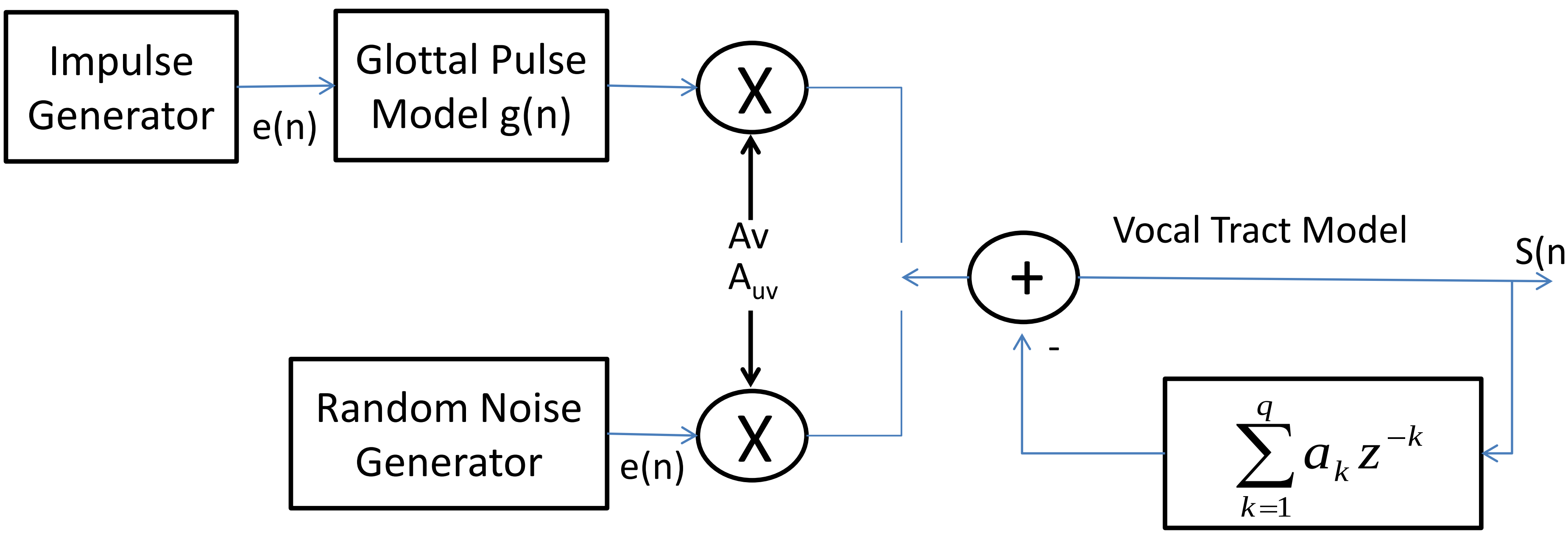
Spectral Structure of Speech



Use of the Vocal Tract Model

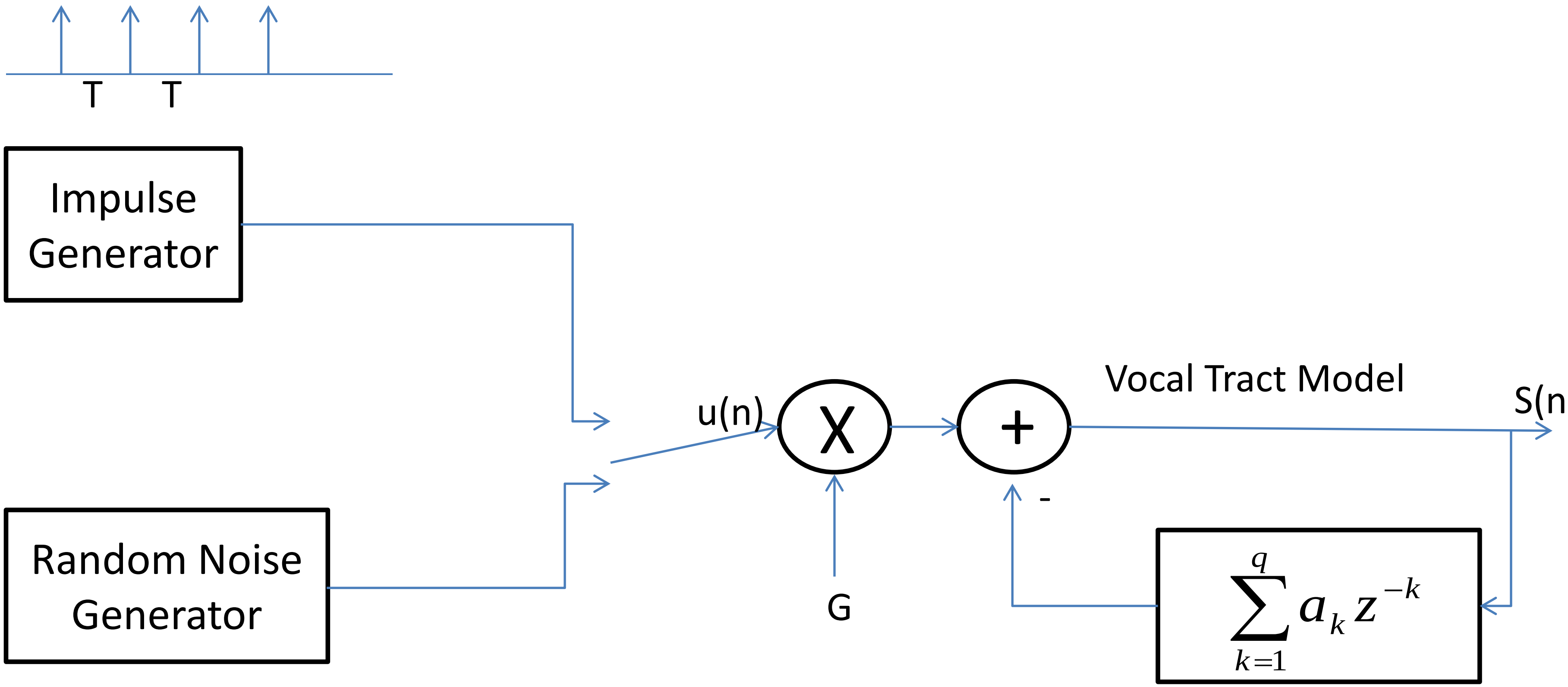
- The model of the vocal tract which has been outlined can be made to be a very accurate model of speech production for short (10-30 ms) frames of speech samples.
- It is widely used in modern low bit rate speech coding algorithms, as well as speech synthesis and speech recognition /speaker identification systems.
- It is necessary to develop a technique which allows the coefficients of the model to be determined for a given frame of speech.
- The most commonly used technique is called Linear Prediction coding (LPC)

Model for Speech analysis



It is possible to combine the components into one all pole model as shown previously

Refinement of this Model



Parameters of this model: $\mathbf{a_k}$ **, G, T, v/uv** classification

Vocal Tract Model

- We have already deduced the transfer function of the vocal tract excitation function to the speech signal

$$\frac{S(z)}{U(z)} = \frac{G}{1 + \sum_{k=1}^{q} a_k z^{-k}}$$

$$s(n) = \sum_{k=1}^{q} a_k s(n-k) + Gu(n)$$

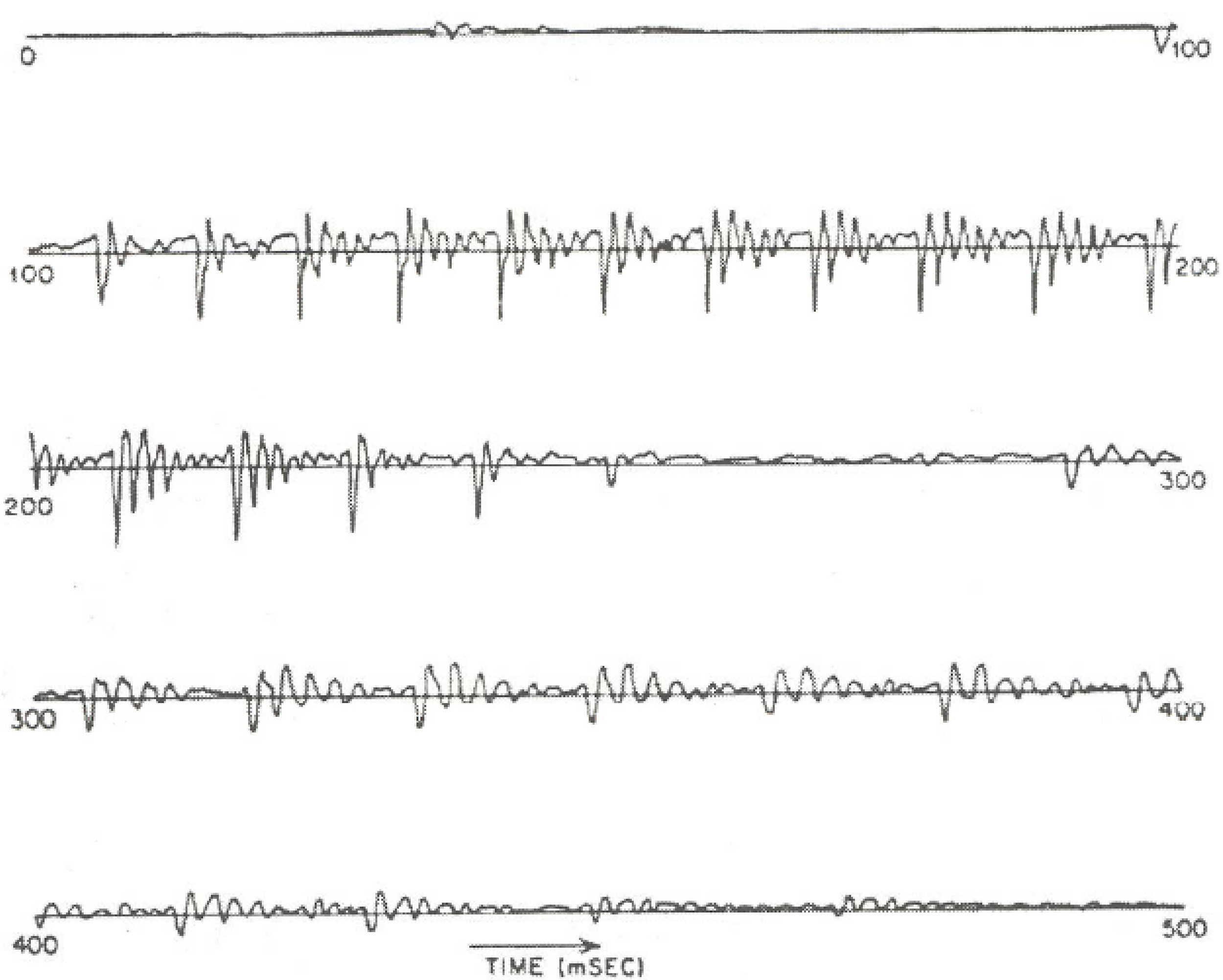
Exercise:

The waveform plot given below is for the word "cattle". Note that each line of the plot corresponds to 10 ms of the signal.

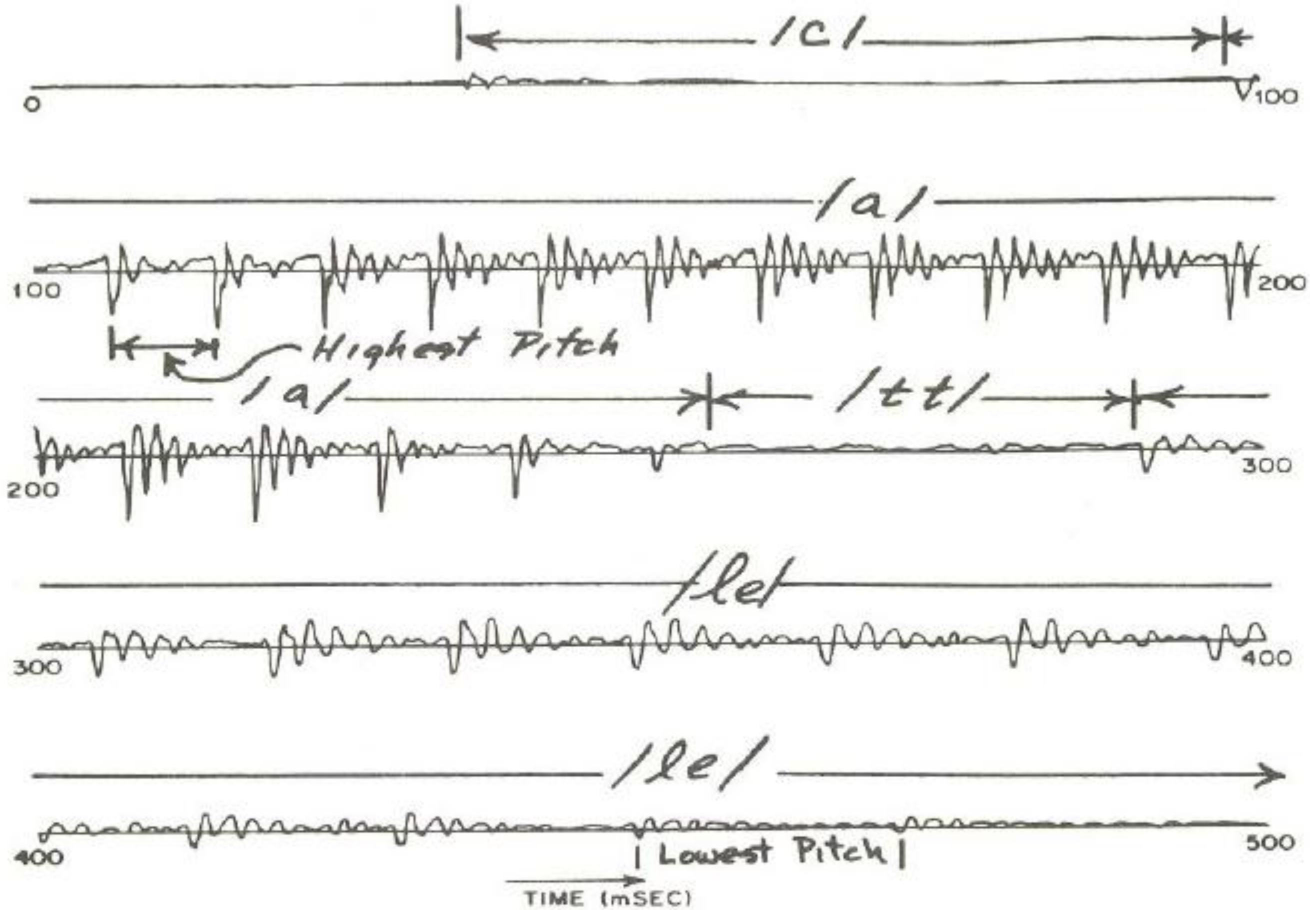
(a) Indicate the boundaries between the phonemes; i.e give the times corresponding to the boundaries /c/a/tt/le/.

(b) Indicate the point where the voice pitch frequency is (i) the highest; and (ii) the lowest. Where are the approximate pitch frequencies at these points?

(c) Is the speaker most probably a male, or a child? How do you know.



Speech waveform of the word ‘Cattle’



The lowest pitch has a period of about 21.5 ms corresponding to the frequency 46 Hz. This low pitch indicates the speaker is probably male

Exercise: The transfer function of the glottal model is given by

$$G(z) = \frac{(1 - e^{-cT})^2}{(1 - e^{-cT} z^{-1})^2}$$

where 'c' is a constant and T is the sampling period (125 μs).

- Obtain the frequency response, G(θ), where θ is the digital frequency.
- Obtain expressions for the magnitude
  - (i) |G(θ)| at DC;
  - (ii) |G(θ)| at half the sampling frequency.

- Calculate the magnitude ratio of (i)/(ii) above in dB. If the magnitude ratio is chosen to be 40 dB, then calculate the value of the constant c.

Example:

The relationship between pressure and volume velocity at the lips is given by

$$P_L(s) = Z_L(s)\, U_L(s)$$

where $P_L(s)$ and $U_L(s)$ are the Laplace transforms of p(t) and u(t) respectively, and

Radiation impedance: $Z_L(s) = \dfrac{sR_r L_r}{R_r + sL_r}$

Radiation resistance: $R_r = \dfrac{128}{9\pi^2}$

Radiation inductance: $L_r = \dfrac{8a}{3\pi c}$

where c is the velocity of sound and a is the radius of the lip opening. In a discrete-time model, we desire a corresponding relationship of the form

$$P_L(z) = Z_L(z)\, U_L(z)$$

where $P_L(z)$ and $U_L(z)$ are z-transforms of $p_L(n)$ and $u_L(n)$, the sampled versions of the bandlimited pressure and volume velocity.

One approach to obtaining R(z) is to use the bilinear transformation, i.e.

$$R(z) = Z_L(s)\Big|_{s=\frac{2}{T}\left\{\frac{1-z^{-1}}{1+z^{-1}}\right\}}$$

(a) For $Z_L(s)$ as given above determine R(z).

Solution: $$R(z) = \frac{\frac{2}{T}\left(\frac{1-z^{-1}}{1+z^{-1}}\right)R_r L_r}{R_r + \frac{2}{T}\left(\frac{1-z^{-1}}{1+z^{-1}}\right)L_r}$$

$$R(z) = \frac{2R_r L_r(1-z^{-1})}{(R_r T + 2L_r) - (2L_r - R_r T)z^{-1}}$$

(b) Write the corresponding difference equation that relates $p_L(n)$ and $u_L(n)$.

$$p_L(n) = \left(\frac{2L_r - R_rT}{2L_r + R_rT}\right) p_L(n-1) + \left(\frac{2L_rR_r}{2L_r + R_rT}\right)\left(u_L(n) - u_L(n-1))\right)$$

(c) Give the locations of thc pole and zero of R(z).

Zero at $z = 1$, pole at $z = \dfrac{2L_r - R_rT}{R_rT + 2L_r}$

(d) If c = 35000 cm/sec, T = 10^{-4} sec^{-1}, and $0.5 \text{ cm} < a < 1.3 \text{ cm}$, what is the range of pole values.

$$R_r = \frac{128}{9\pi^2} = 1.441 \quad \text{and} \quad L_r = \frac{8a}{3\pi c} = 24.25\, a \times 10^{-6}$$

$$= 12.125 \times 10^{-6},\ a = 0.5$$

$$= 31.53 \times 10^{-6},\ a = 1.3\ .$$

When a = 0.5, pole at -0.7119 and a = 1.3, pole at -0.3912

In both cases the pole is pretty far inside the unit circle.

(e) A simple approximation to R(z) obtained above is obtained by neglecting the pole; i.e.

$$\hat{R}(z) = R_0(1 - z^{-1})$$

For a =1 cm and T = 10^{-4}, find R0 such that $\hat{R}(-1) = R(-1) = Z_L(\infty)$.

Solution: $$R(-1) = \frac{2R_r L_r(2)}{R_r T + 2L_r + 2L_r - R_r T} = R_r$$

$$\hat{R}(-1) = 2R_0$$

Therefore $R_0 = R_r/2 = 0.7205$.

Example:

A commonly used approximation to the glottal pulse is

$$g(n) = n\, a^n \qquad n \geq 0 \quad \{ a > 0 \}$$

$$= 0 \qquad n < 0$$

(a) Find the z-transform of g(n).

(b) Sketch the Fourier transform, $|G(\theta)|$, as a function of θ.
(θ = digital frequency; $\theta = \omega T$)

(c) The value a is normally chosen using the following criteria:

$$20 \operatorname{Log}_{10} |G(\theta)|_{\theta=0} - 20 \operatorname{Log}_{10} |G(\theta)|_{\theta=\pi} = 60 \text{ dB}$$

Show that a = 0.9387.

{ Use the fact that the z-transform of $nx(n) = -z\dfrac{dX(z)}{dz}$ }

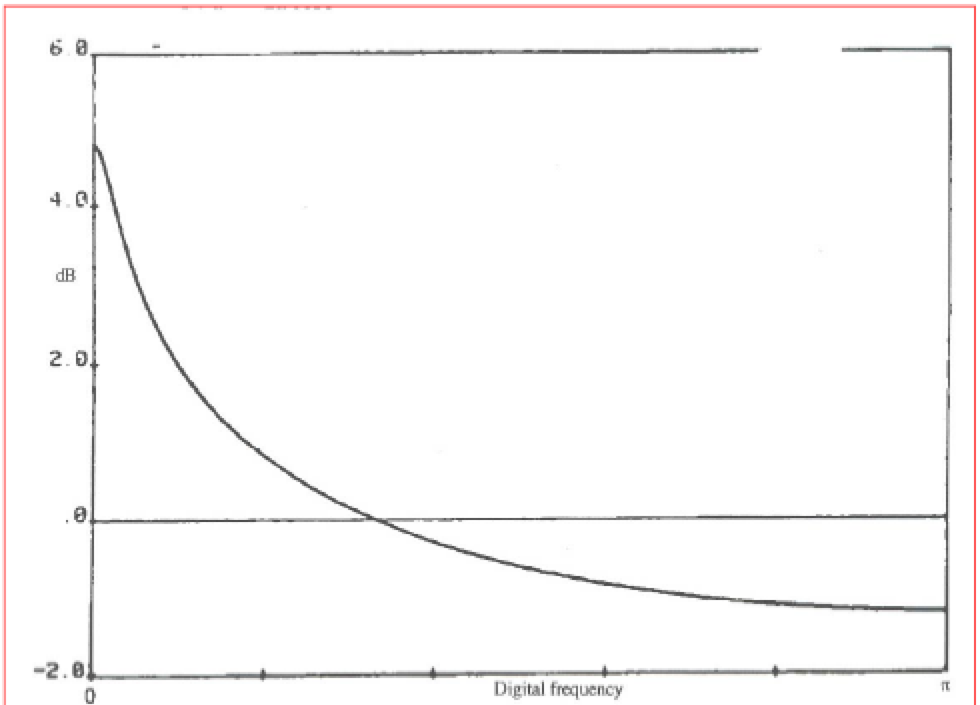
$$\text{(a)} \qquad x(n) = a^{n} \;; \qquad X(z) = \frac{1}{1 - az^{-1}} = \frac{z}{z - a};$$

$$\left\{ \frac{dx(z)}{dz} = \frac{-a}{(z-a)^2} = \frac{-a}{z^2(1 - az^{-1})^2} \right\}$$

$$G(z) = -z\left\{ \frac{-a}{z^2(1 - az^{-1})^2} \right\} = = \frac{az^{-1}}{(1 - az^{-1})^2}$$

(b) $G(\theta) = G(z)\big|_{z=e^{j\theta}} = \dfrac{ae^{-j\theta}}{(1-ae^{-j\theta})^2} = \dfrac{ae^{-j\theta}}{(1-a\cos\theta + ja\sin\theta)^2}$

$$|G(\theta)| = \frac{a}{1+a^2-2a\cos\theta}$$



(c)

$$\left|G(\theta)\right|_{\theta=0} = \frac{a}{(1-a)^2}; \quad \left|G(\theta)\right|_{\theta=\pi} = \frac{a}{(1+a)^2};$$

$$20\log\frac{\frac{a}{(1-a)^2}}{\frac{a}{(1+a)^2}} = 60; \quad \Rightarrow \frac{\frac{a}{(1-a)^2}}{\frac{a}{(1+a)^2}} = 1000; \quad \Rightarrow \frac{1+a}{1-a} = \sqrt{1000}$$

$$a = 0.9387$$