Section 1

Introduction to Speech Processing

Acknowledgement: This lecture is mainly derived from Rabiner, L., and Juang, B.-H., *Fundamentals of Speech Recognition*, Prentice-Hall, New Jersey, 1993

Introduction to Speech Processing

- Speech processing is the application of digital signal processing (DSP) techniques to the processing and\or analysis of speech signals
- Applications of speech processing include
  - Speech Coding
  - Speech Recognition
  - Speaker Verification\Identification
  - Speech Enhancement
  - Speech Synthesis (Text to Speech Conversion)

Process of Speech Production

- Figure 1 shows a schematic diagram of the speech production/speech perception process in human beings

- The speech production process begins when the talker formulates a message in his/her mind to transmit to the listener via speech

- The next step in the process is the conversion of the message into a language code. This corresponds to converting the message into a set of phoneme sequences corresponding to the sounds that make up the words, along with prosody (syntax) markers denoting *duration* of sounds, *loudness* of sounds, and *pitch* associated with the sounds.

Process of Speech Production

- Once the language code is chosen the talker must execute a series of neuromuscular commands to cause the vocal cords to vibrate when appropriate and to shape the vocal tract such that the proper sequence of speech sounds is created and spoken by the talker, thereby producing an acoustic signal as the final output.

- The neuromuscular commands must simultaneously control all aspects of articulatory motion including control of the **lips, jaw, tongue and velum.**

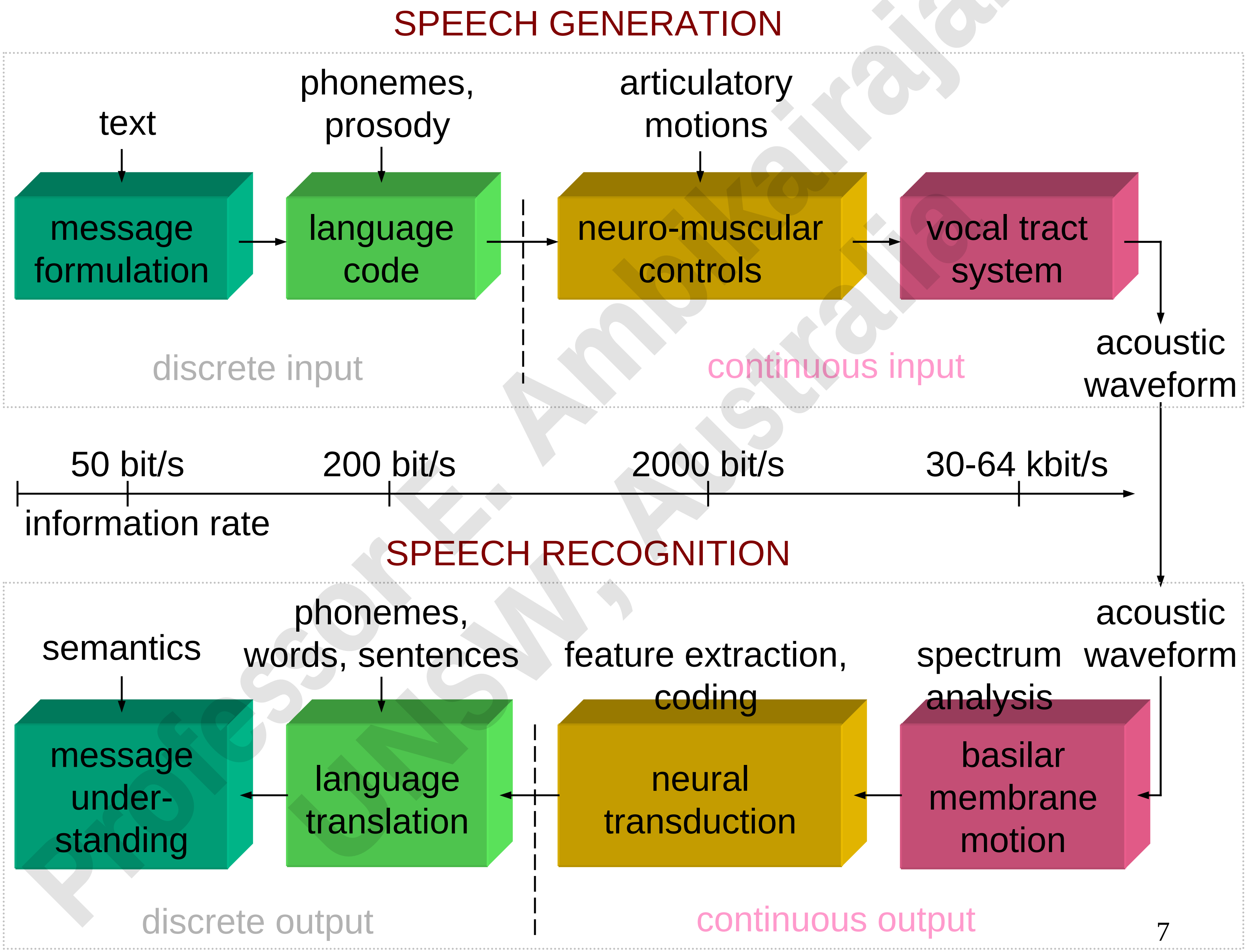
Process of Speech Perception

- Once the speech signal is generated and propagated to the listener, the speech perception process begins.
- A neural transduction process converts the spectral signal at the output of the basilar membrane into activity signals on the auditory nerve, corresponding roughly to a feature extraction process.
- The neural activity along the auditory nerve is converted into a language code at higher centres of processing within the brain, and finally message comprehension (understanding of meaning) is achieved.

Information Rate of the Speech Signal

- The discrete symbol information rate in the raw message text is rather low (about 50 bits per second corresponding to about 8 sounds per second, where each sound is one of the about 50 distinct symbols)

- After the language code conversion, with the inclusion of prosody information, the information rate rises to about 200 bps



SPEECH GENERATION

text

phonemes, prosody

articulatory motions

message formulation

language code

neuro-muscular controls

vocal tract system

acoustic waveform

discrete input

continuous input

50 bit/s

200 bit/s

2000 bit/s

30-64 kbit/s

information rate

SPEECH RECOGNITION

semantics

phonemes, words, sentences

feature extraction, coding

spectrum analysis

acoustic waveform

message under-standing

language translation

neural transduction

basilar membrane motion

discrete output

continuous output

Figure 1.1

Information Rate of the Speech Signal

- In the next stage the representation of the information in the signal becomes continuous with an equivalent rate of about 2000 bps at the neuromuscular control level and about 30,000-50,000 bps at the acoustic signal level.

- The continuous information rate at the basilar membrane is in the range of 30,000-50,000 bps, while at the neural transduction stage it is about 2000 bps.

- The higher level processing within the brain converts the neural signals to a discrete representation, which ultimately is decoded into a low bit rate message.

The mechanism of Speech Production

- In order to apply DSP techniques to speech processing problems it is important to understand the fundamentals of the speech production process.

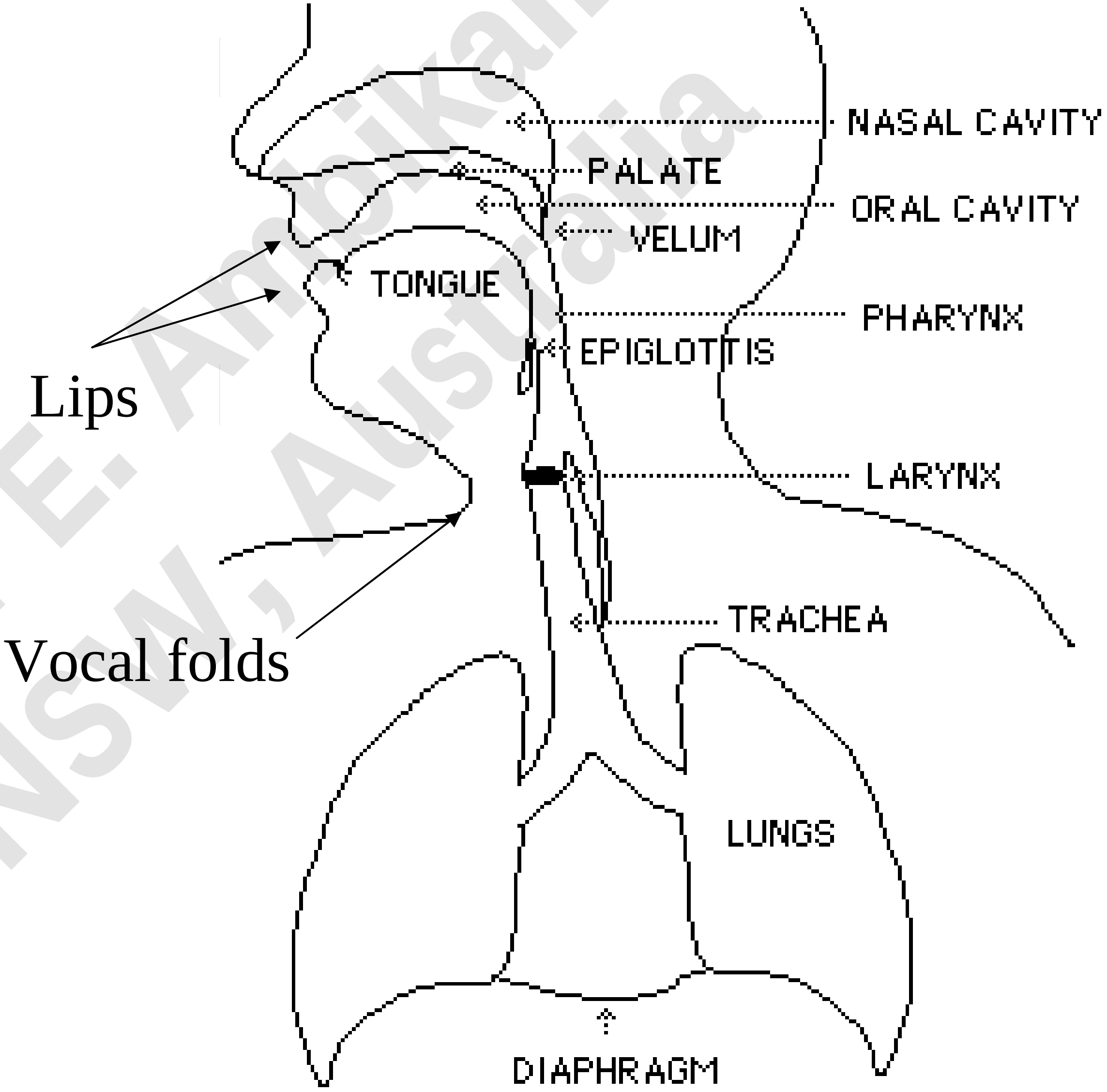
- Speech signals are composed of a sequence of sounds and the sequence of sounds are produced as a result of acoustical excitation of the vocal tract when air is expelled from the lungs (See figure 1.2)

Speech Production Mechanism

- Vocal tract begins at the opening between the vocal cords and ends at the lips

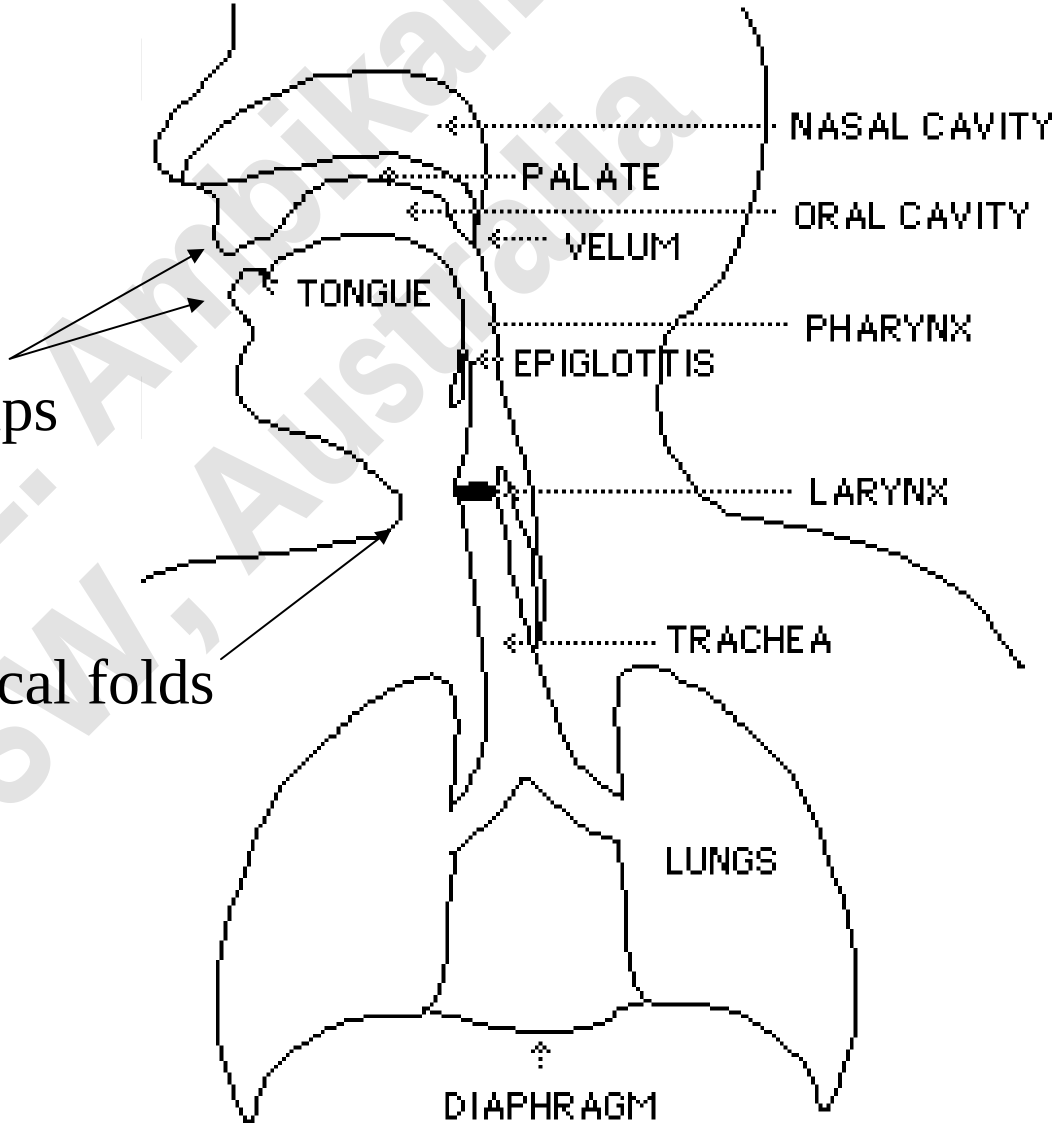
- In the average male, the total length of the vocal tract is about 17 cm.

- The cross-ectional area of the vocal, determined by the positions of the tongue, lips, jaw and velum varies from zero (complete closure) to about 20 cm^2.



Speech Production Mechanism

- The nasal tract begins at the velum and ends at the nostrilss
- When the velum is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech.



Lips

Vocal folds

Classification of Speech Sounds

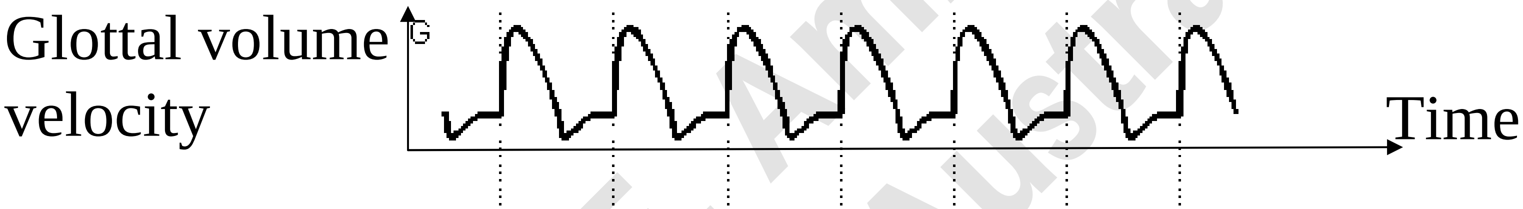
- In speech processing, speech sounds are divided into TWO broad classes which depend on the role of the vocal chords on the speech production mechanism
  - VOICED speech is produced when the vocal chords play an active role (i.e. vibrate) in the production of a sound:

    Examples: Voiced sounds /a/ , /e/, /i/
  - UNVOICED speech is produced when vocal chords are inactive

    Examples: unvoiced sounds /s/, /f/

“Voiced” Speech

- Voiced speech occurs when air flows through the vocal chords into the vocal tract in discrete “puffs” rather than as a continuous flow



- The vocal chords vibrate at a particular frequency, which is called the fundamental frequency of the sound
  - 50 : 200 Hz for male speakers
  - 150:300 Hz for female speakers
  - 200:400 Hz child speakers

“Unvoiced” Speech

- For unvoiced speech, the vocal chords are held open and air flows continuously through them
- The vocal tract, however, is narrowed resulting in a turbulent flow of air along the tract
- Examples include the unvoiced fricatives /f/ & /s/
- Characterised by high frequency components

Other Sound Classes

- **Nasal Sounds**
  - – Vocal tract coupled acoustically with nasal cavity through velar opening
  - – Sound radiated from nostrils as well as lips
  - – Examples include m, n, ing
- **Plosive Sounds**
  - – Characterised by complete closure/constriction towards front of the vocal tract
  - – Build up of pressure behind closure, sudden release
  - – Examples include p, t, k

Resonant Frequencies of Vocal Tract

- Vocal Tract is a non-uniform acoustic tube that is terminated at one end by the vocal chords and at the other end by the lips
- The Cross-sectional area of the vocal tract determined by the positions of the tongue, lips, jaw and velum.depends on lips, tongue, jaw and velum
- The spectrum of vocal tract response consists of a number of resonant frequencies of the vocal tract.
- These frequencies are called Formants
- Three to four formants present below 4kHz of speech

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 (see Fig 1.3_
- For UNVOICED speech, the magnitude of the higher formant frequencies is successively larger than the magnitude of the lower formant frequencies (see Fig 1.3)

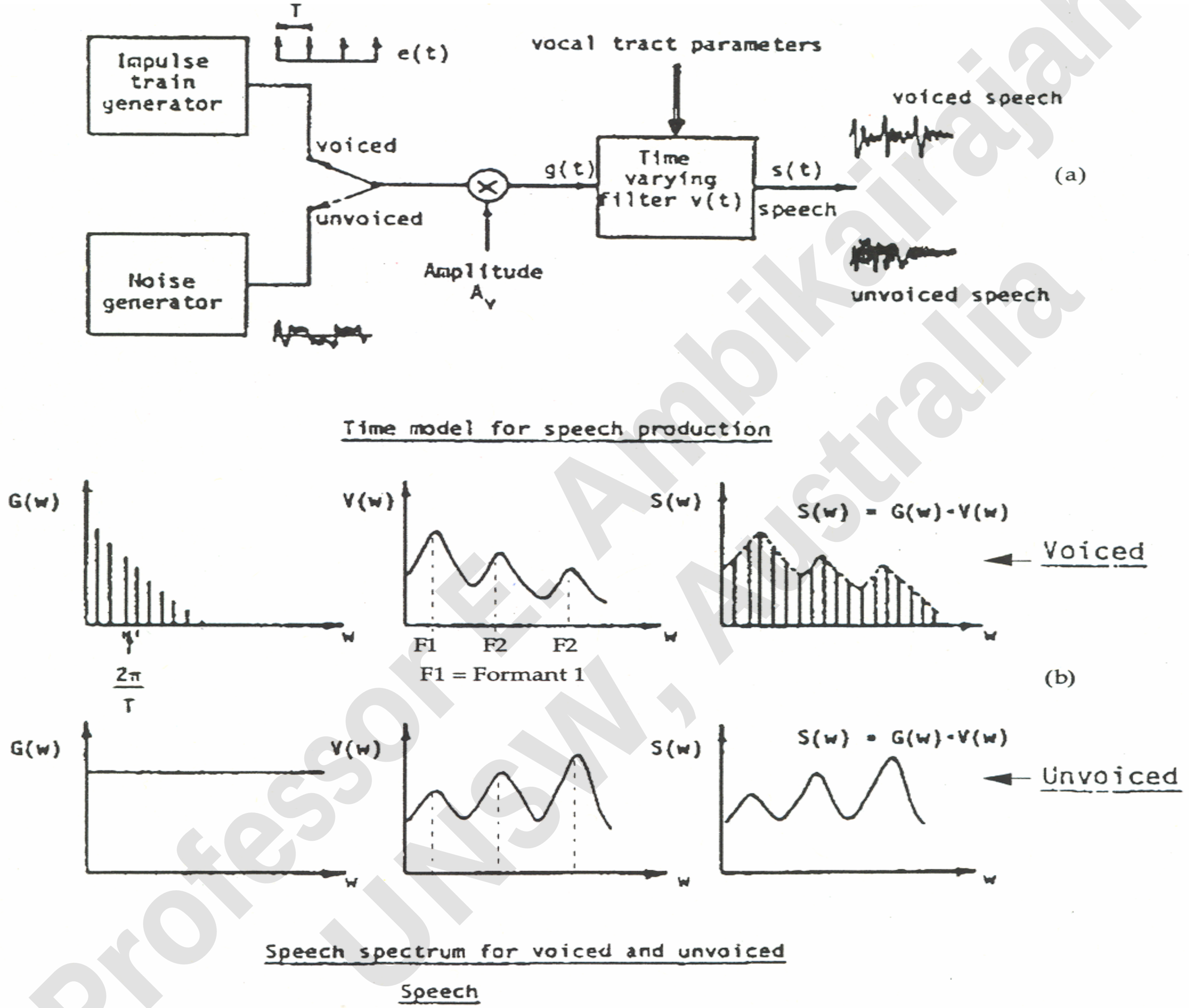


Figure 1.3: Time model for speech production (b) Speech spectrum

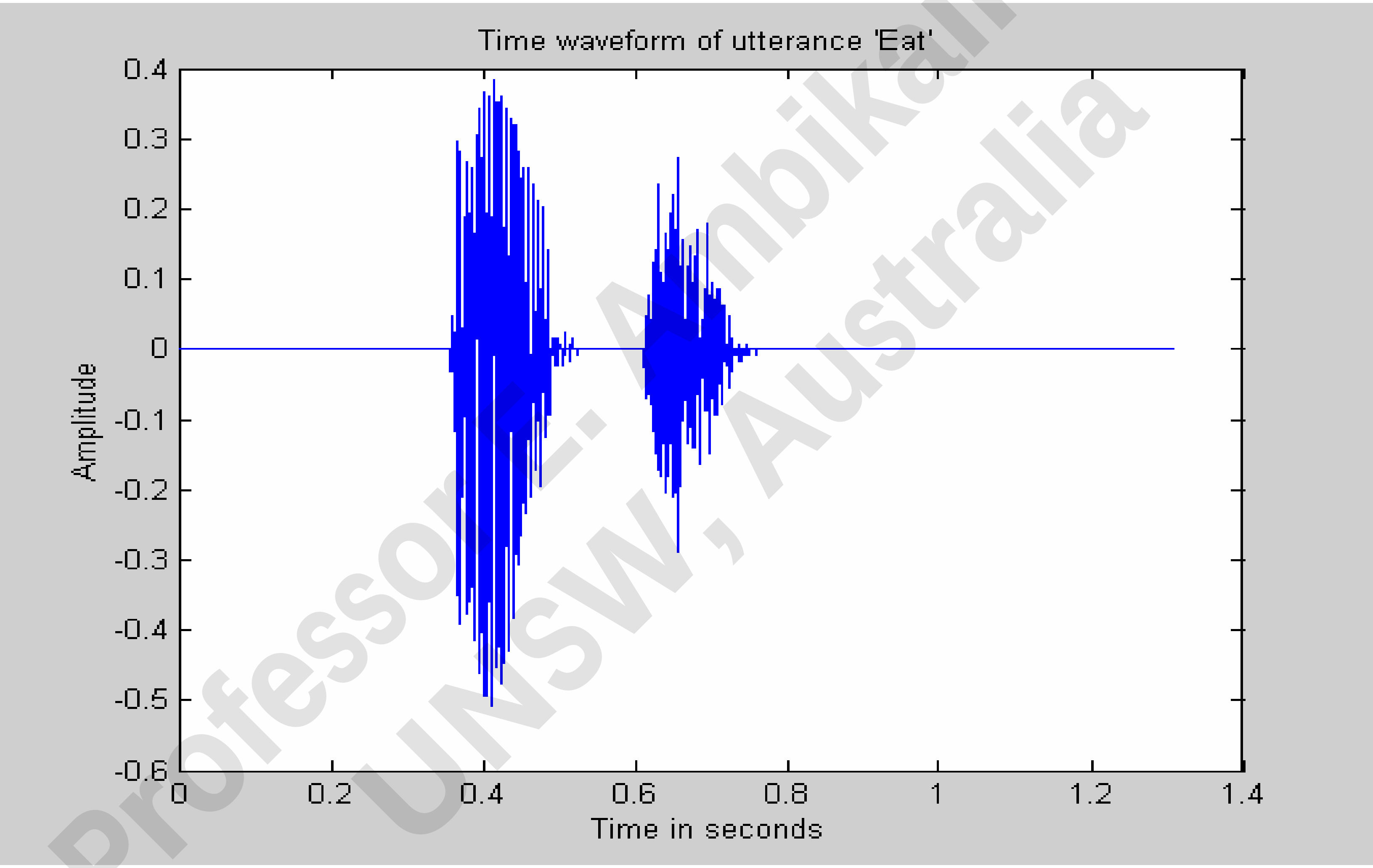
Basic Assumptions of Speech Processing

- The basic assumption of almost all speech processing systems is that the source of excitation and the vocal tract system are independent.
- Therefore, it is a reasonable approximation to model the source of excitation and the vocal tract system separately as shown (Figure 1.3)
- The vocal tract changes shape rather slowly in continuous speech and it is reasonable to assume that the vocal tract has a fixed characteristics over a time interval of the order of 10 ms.
- Thus once every 10 ms, on average, the vocal tract configuration is varied producing new vocal tract parameters (resonant frequencies)

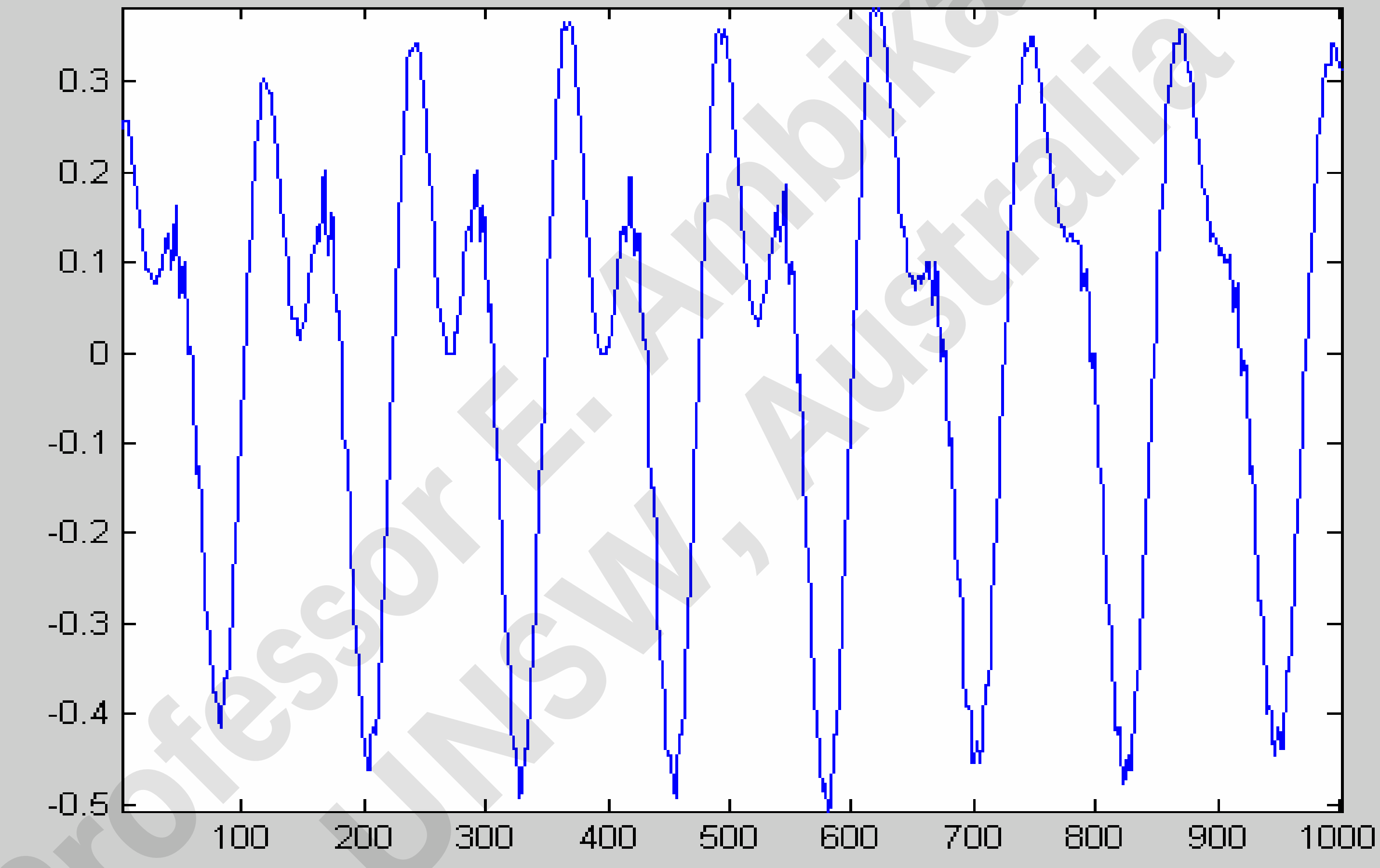
Speech Sounds

- Phonemes: smallest segments of speech sounds /d/ and /b/ are distinct phonemes e.g. d*ark* and b*ark*
- It is important to realise, that phonemes are abstract linguistic units and may not be directly observed in the speech signal
- Different speakers producing the same string of phonemes convey the same information yet sound different as a result of differences in dialect and vocal tract length and shape.
- There are about 40 phonemes in English
- See Table A for IPA (International Phonetic Alphabet) symbol for each phoneme together with sample words in which they occur.

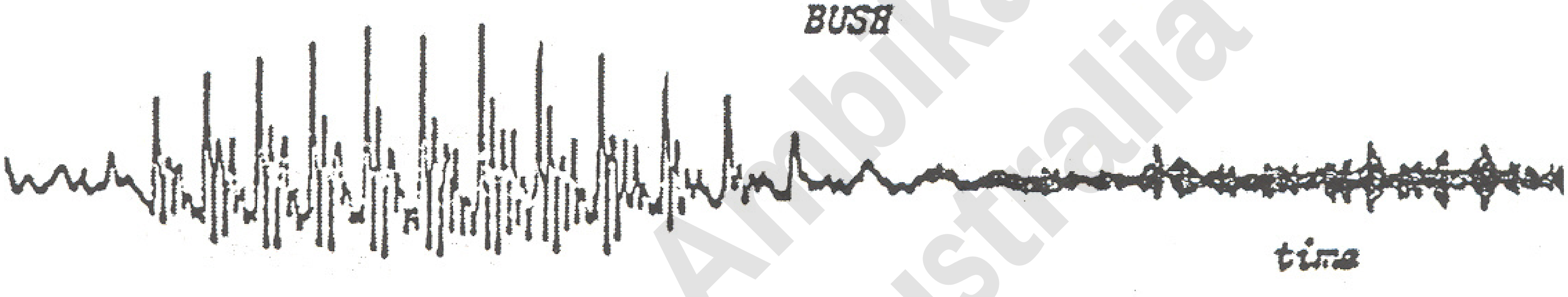
Acoustic Waveforms



Frame of waveform



Example: The acoustic waveform of the word ‘Bush’



Amplitude/Time plot of 'Bush'

The speech signal is a slowly time varying signal in the sense that when examined over sufficiently short period of time, its characteristics are fairly stationary.

Speech Production Model

Model for Speech Production

- To develop an accurate model for how speech is produced, it is necessary to develop a digital filter based model of the human speech production mechanism
- Model must accurately represent (Figure 1.4):
  - 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

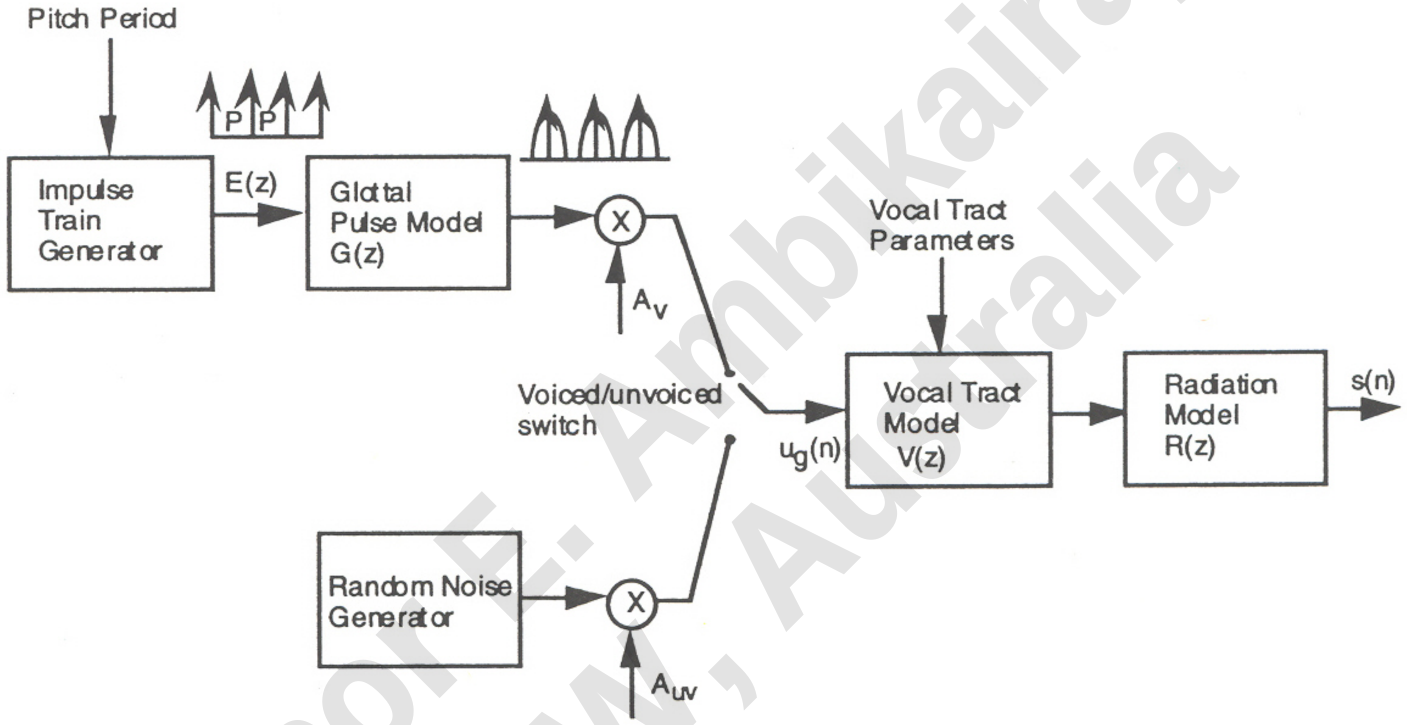


Figure 1.4: Discrete –Time Model for Speech Production

Excitation Process

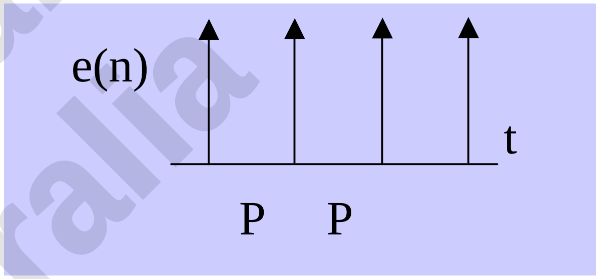
- The excitation process must take into account:-
  - The voiced\unvoiced nature of speech
  - The operation of the glottis
  - The "energy" of the speech signal

  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 unit impulses spaced at intervals of the pitch period ($e[n]=\delta[n-Pk]$ $k=0,1,2\ldots$)
  - For unvoiced speech, the excitation will be a random noise-like signal ($e[n]=\text{random}[n]$)

Excitation Source – Voiced Speech

- Impulse train:

  e(n)=δ(n-Pk) k=0,1,2…



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

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

$$E(z) = 1 + z^{-P} + z^{-2P} + \ldots$$

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

Excitation Process

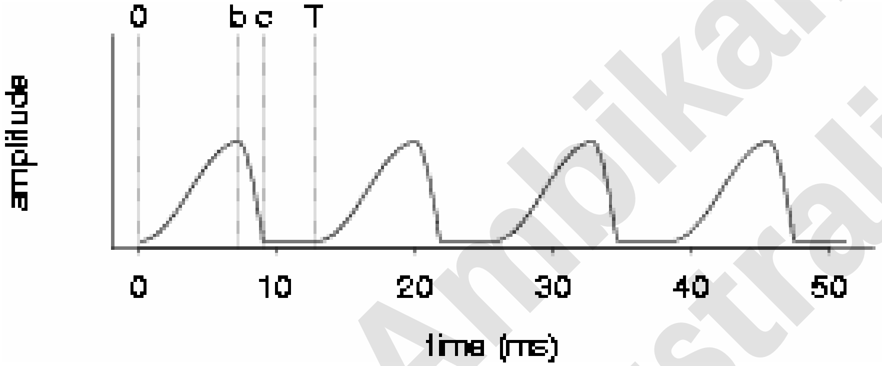
- The next stage in the excitation process will be a model of the pulse shaping operation of the glottis
- This is only used for VOICED speech
- Typically used transfer function for the glottal model are:

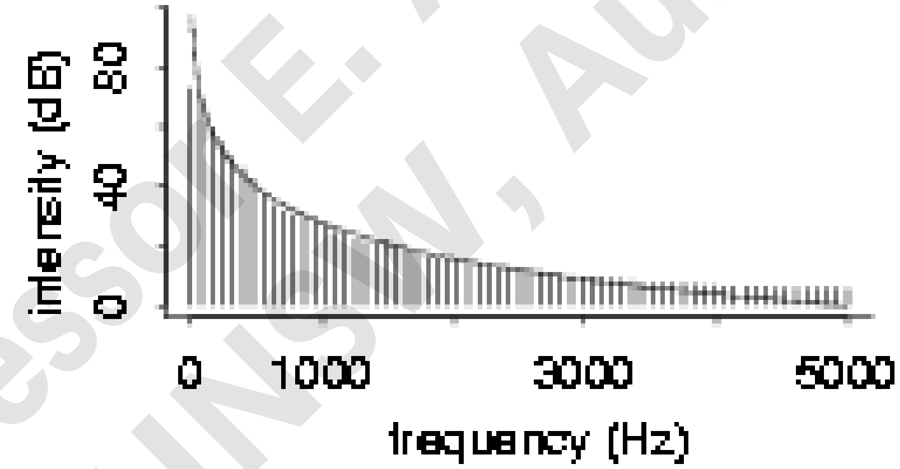
$$G(z) = \frac{1}{\left(1 - e^{-cT} z^{-1}\right)^2} \quad \text{where c : speed of sound}$$

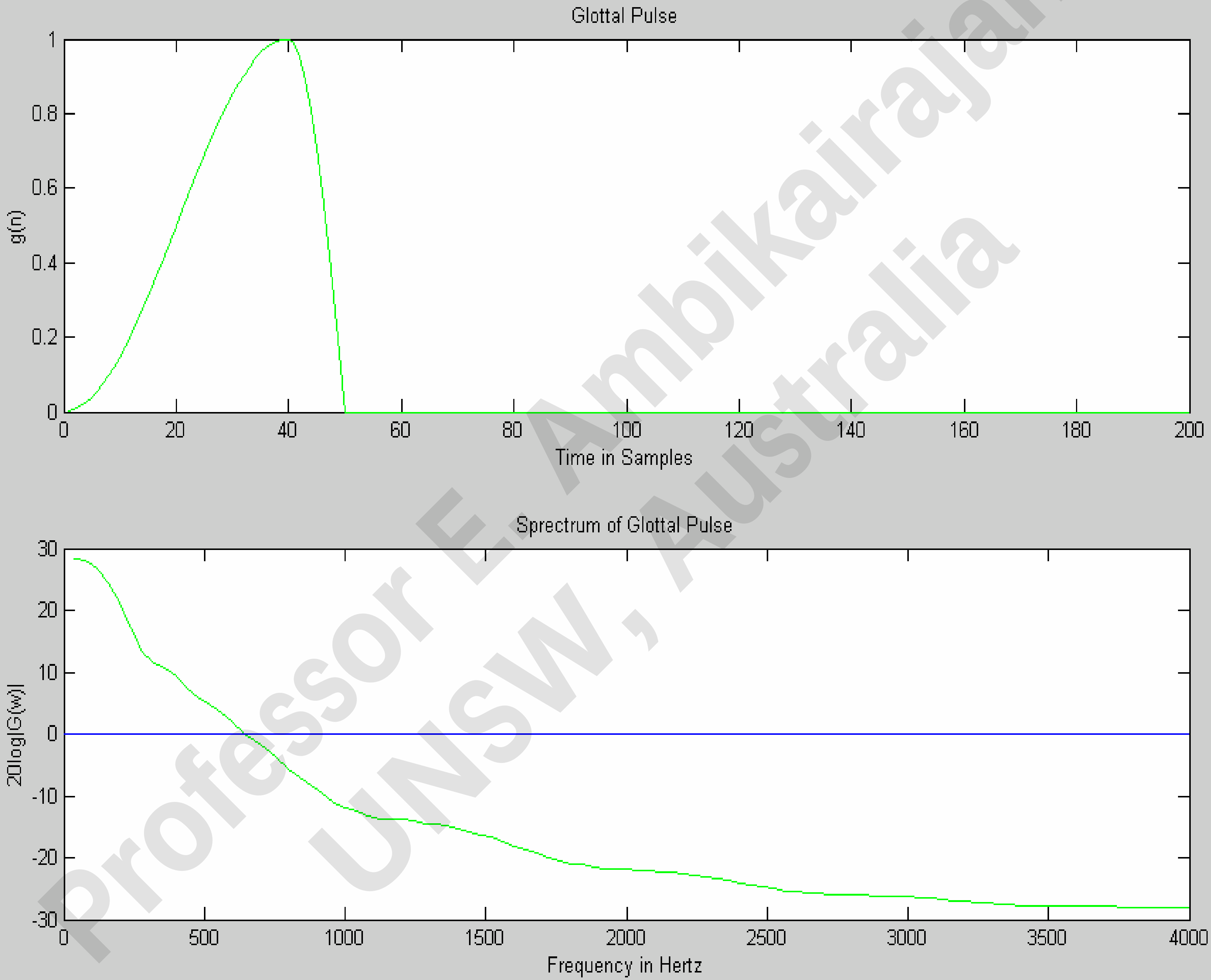
$$\text{But cT} << 1, \therefore e^{-cT} \approx 1$$

$$\therefore G(z) \cong \frac{1}{\left(1 - z^{-1}\right)^2} \quad \text{for voiced speech, G(z)=1 for unvoiced speech}$$

Glottal Pulse and Spectrum






Glottal Pulse
g(n)
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
120
140
160
180
200
Time in Samples
Sprectrum of Glottal Pulse
20log|G(w)|
30
20
10
0
-10
-20
-30
0
500
1000
1500
2000
2500
3000
3500
4000
Frequency in Hertz

Exercise: Glottal Pulse & Spectrum Plot

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

  (N_1 =40 and N_2 = 10)

$$g(n) = \begin{cases} \frac{1}{2}[1-\cos(\pi n/N_1)] & 0 \le n \le N_1 \\ \cos(\pi(n-N_1)/2N_2) & \mathrm{N}_1 \le n \le N_1 + N_2 \\ 0 & otherwise \end{cases}$$

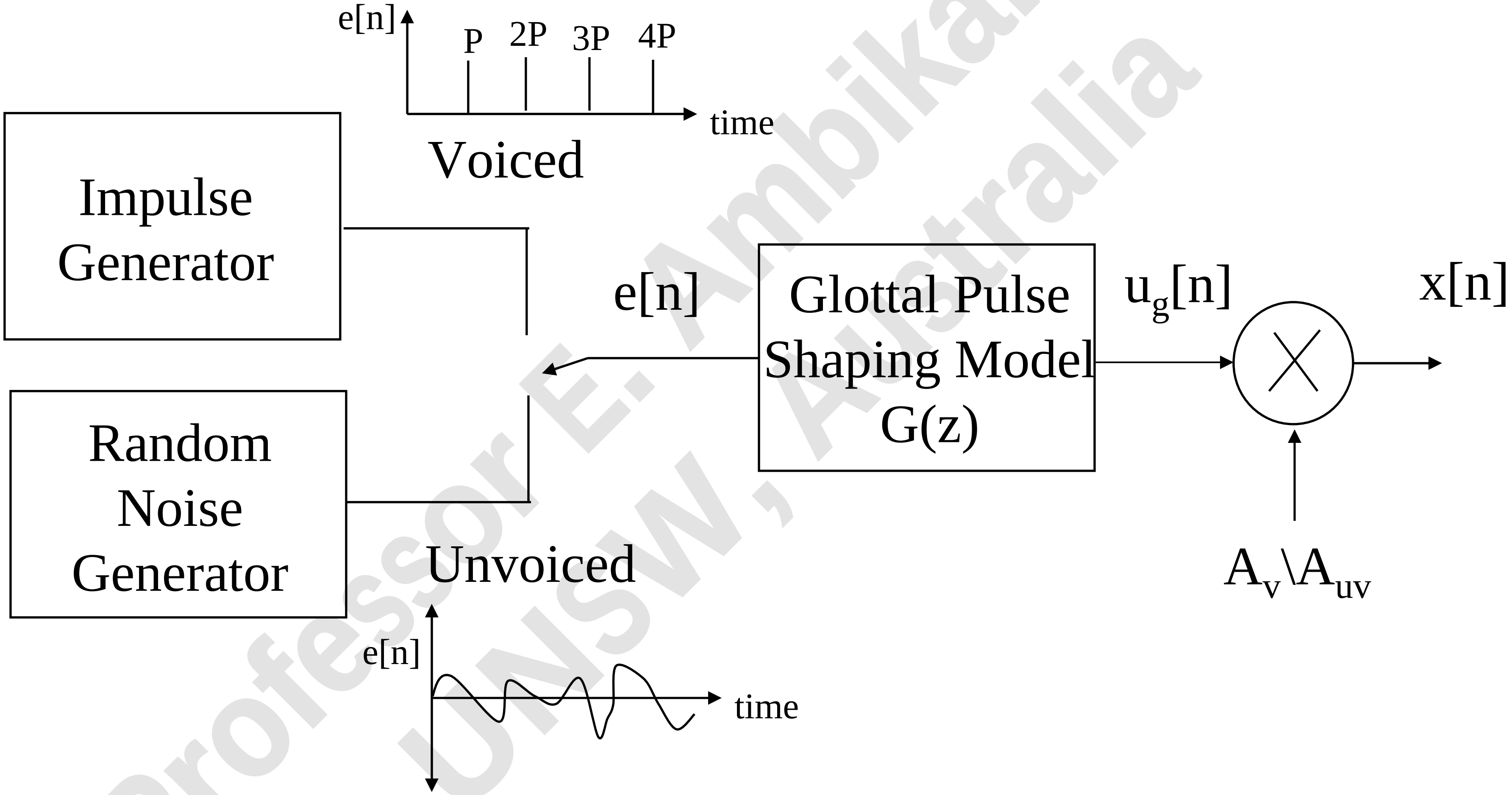
Excitation Process

- Finally, the "energy" of the sound is modelled by a gain factor
  - Typically the gain factor for voiced speech (A_v) will be in the region of 10 times that of unvoiced speech (A_{uv})
- Thus the signal coming out of the complete excitation process will be:

$$x[n]=Ae[n]*g[n], \text{ or}$$

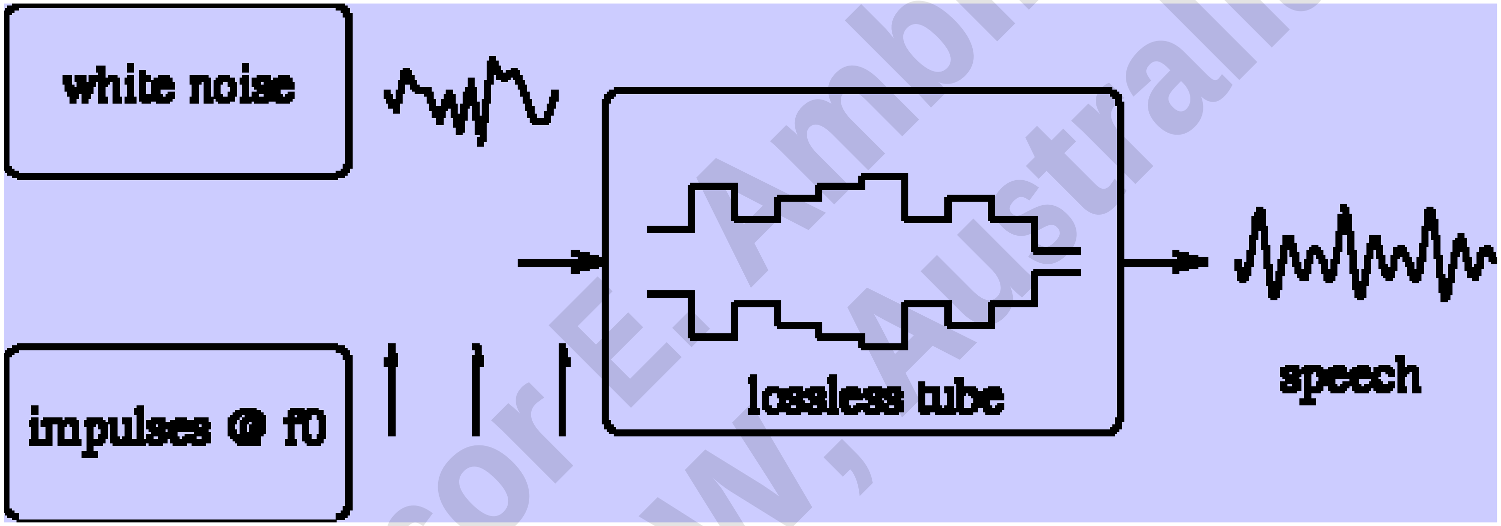
$$X(z)=AE(z)G(z)$$

Discrete Time Model of Excitation Process



Vocal Tract Model

- 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

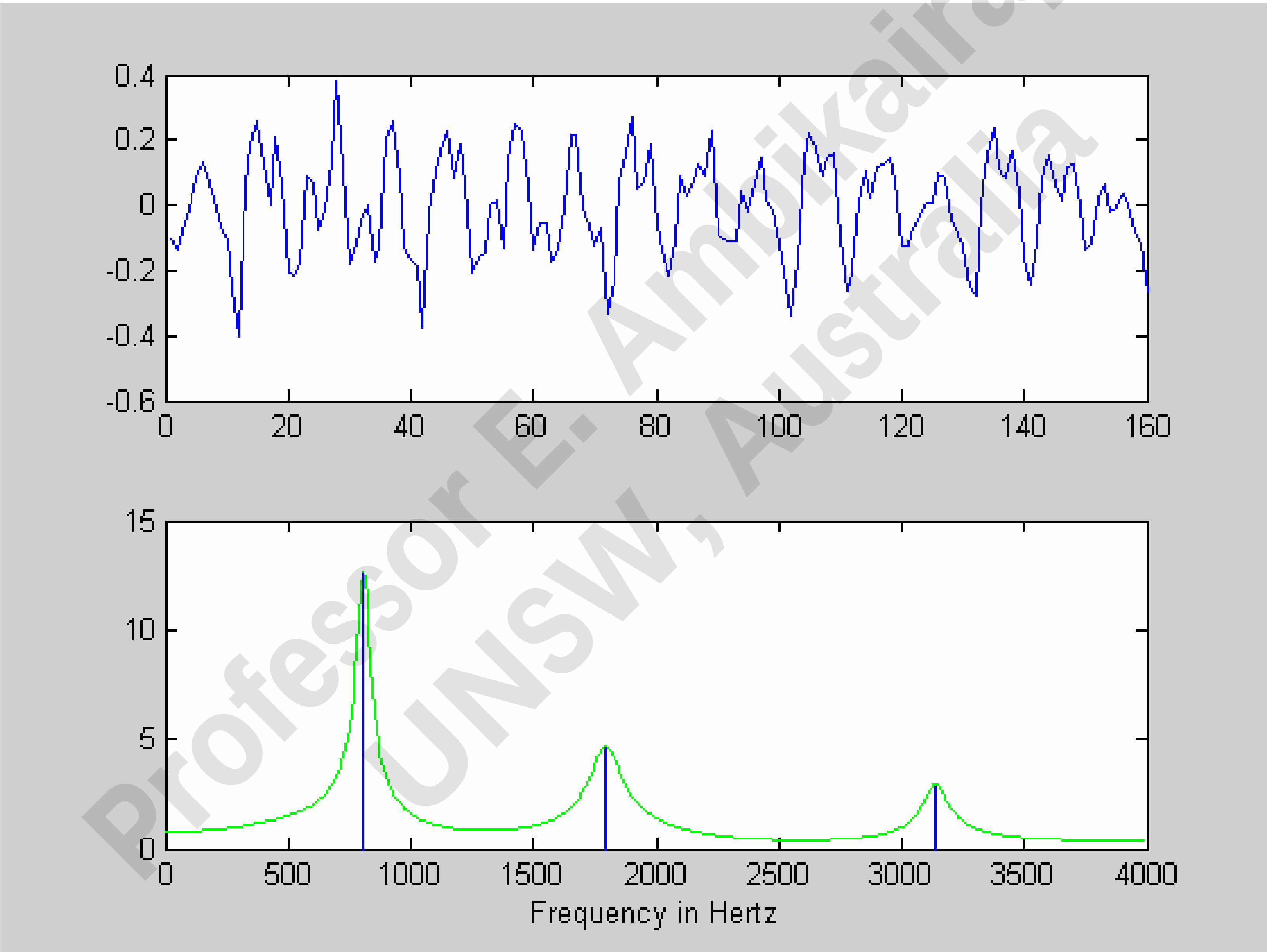
Vocal Tract Model

- 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

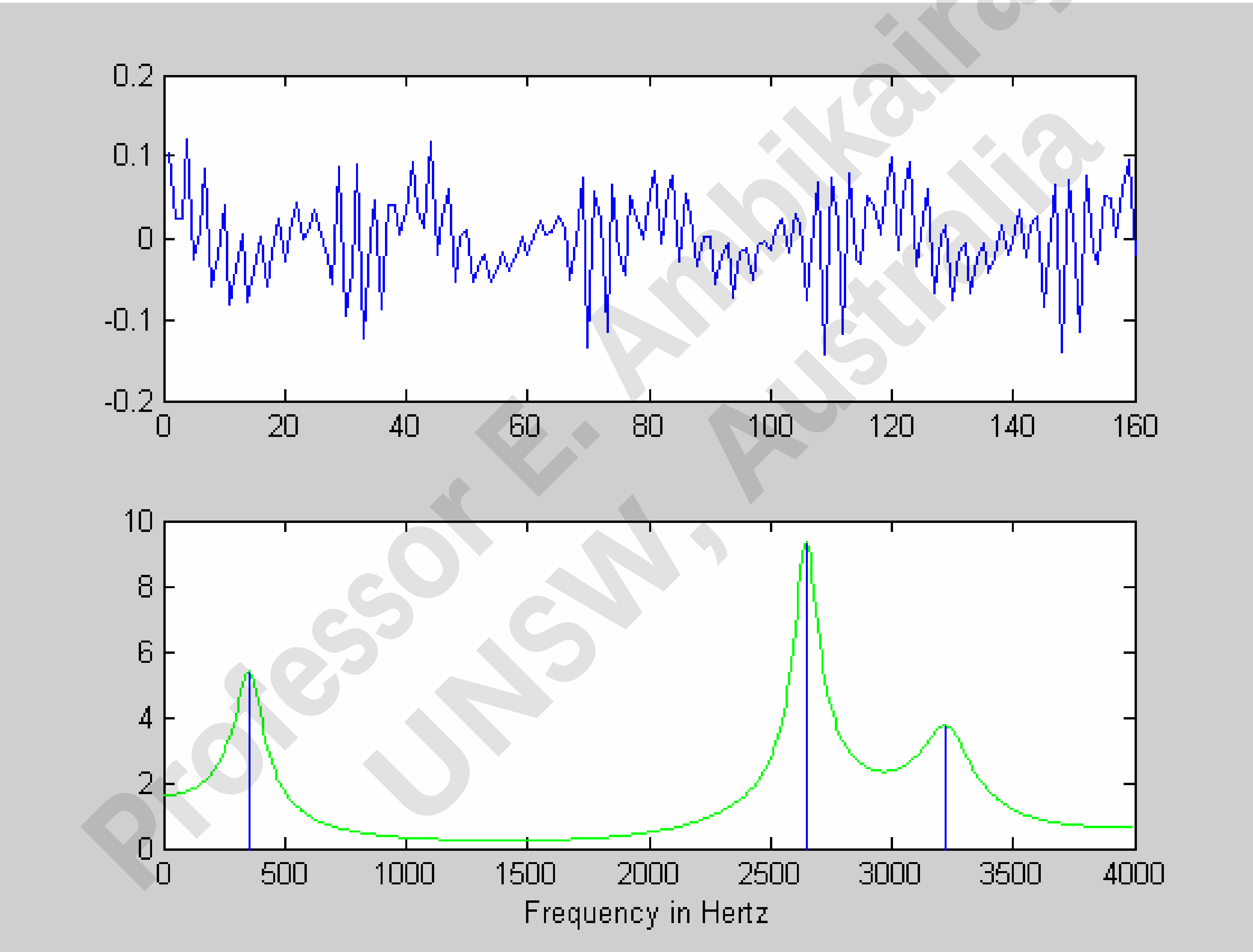
Formant Frequencies

- 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

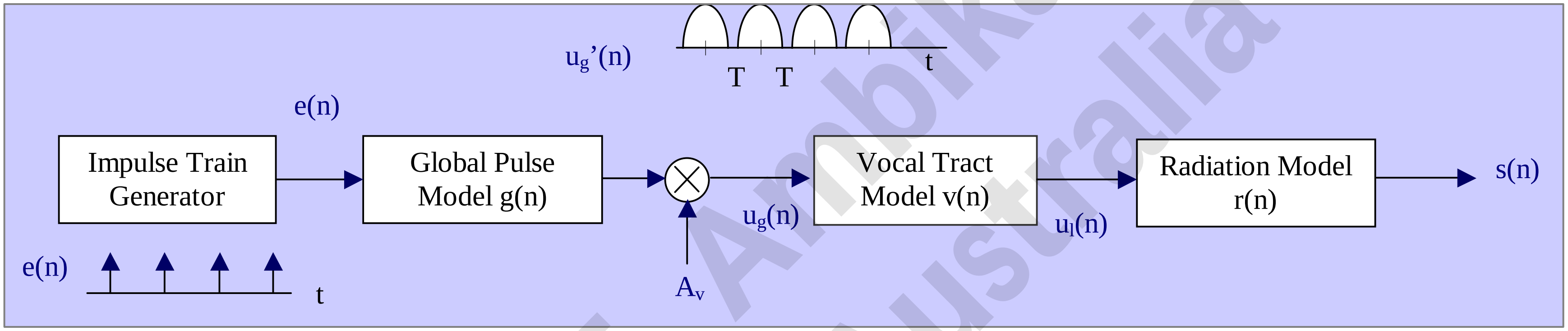


Vocal Tract Model – Voiced Speech

- 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

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

Discrete Time Model for Voiced Speech Production



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

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

$$s(n) = u_l(n) * r(n)$$

$$\therefore \; s(n) = A_V \left[\left(e(n) * g(n)\right) * v(n)\right] * r(n)$$

$$\therefore \; \frac{S(z)}{E(z)} = A_V G(z) V(z) R(z)$$

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:

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

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

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:

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

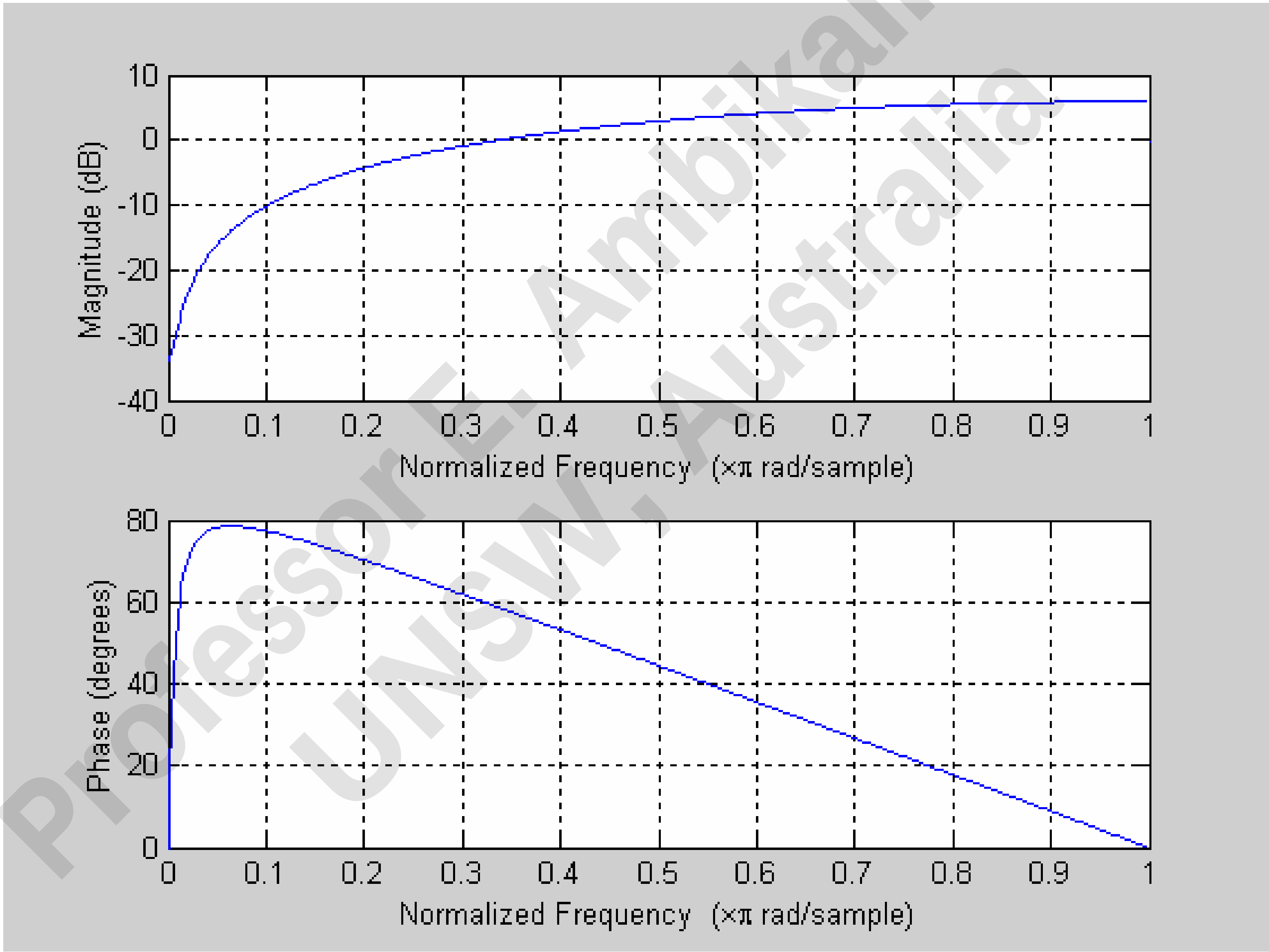
Exercise: Lip Radiation Model

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

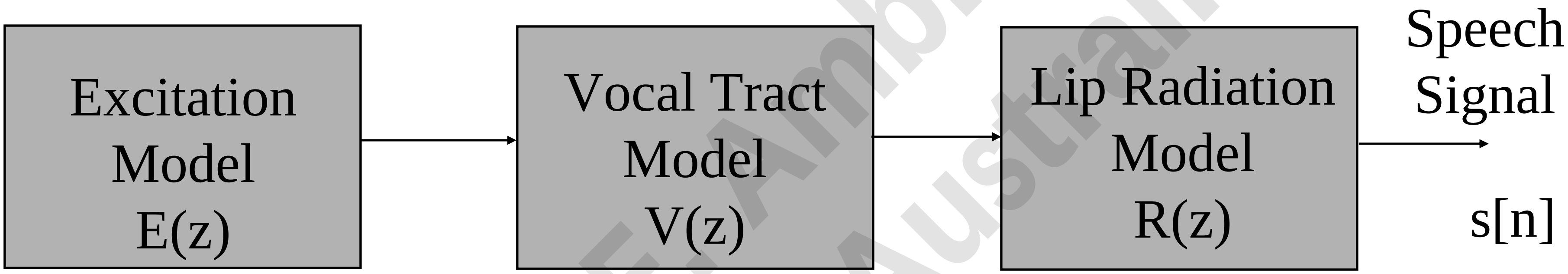
$$R(z)=1-0.98z^{-1}$$

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

Frequency Response of Lip Radiation Model



Overall Speech Production Model



Excitation Model E(z)

Vocal Tract Model V(z)

Lip Radiation Model R(z)

Speech Signal

s[n]

S(z)=E(z)G(z)AV(z)R(z)

Transfer Function:

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

Overall Transfer Function

- For Voiced Speech:

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

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

$$\frac{S(z)}{E(z)} = A_v \frac{1}{\left(1-z^{-1}\right)} \frac{1}{1+\sum_{k=1}^{P} a_k z^{-k}} = \frac{A_v}{1+\sum_{k=1}^{P+1} a'_k \, z^{-k}}$$

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

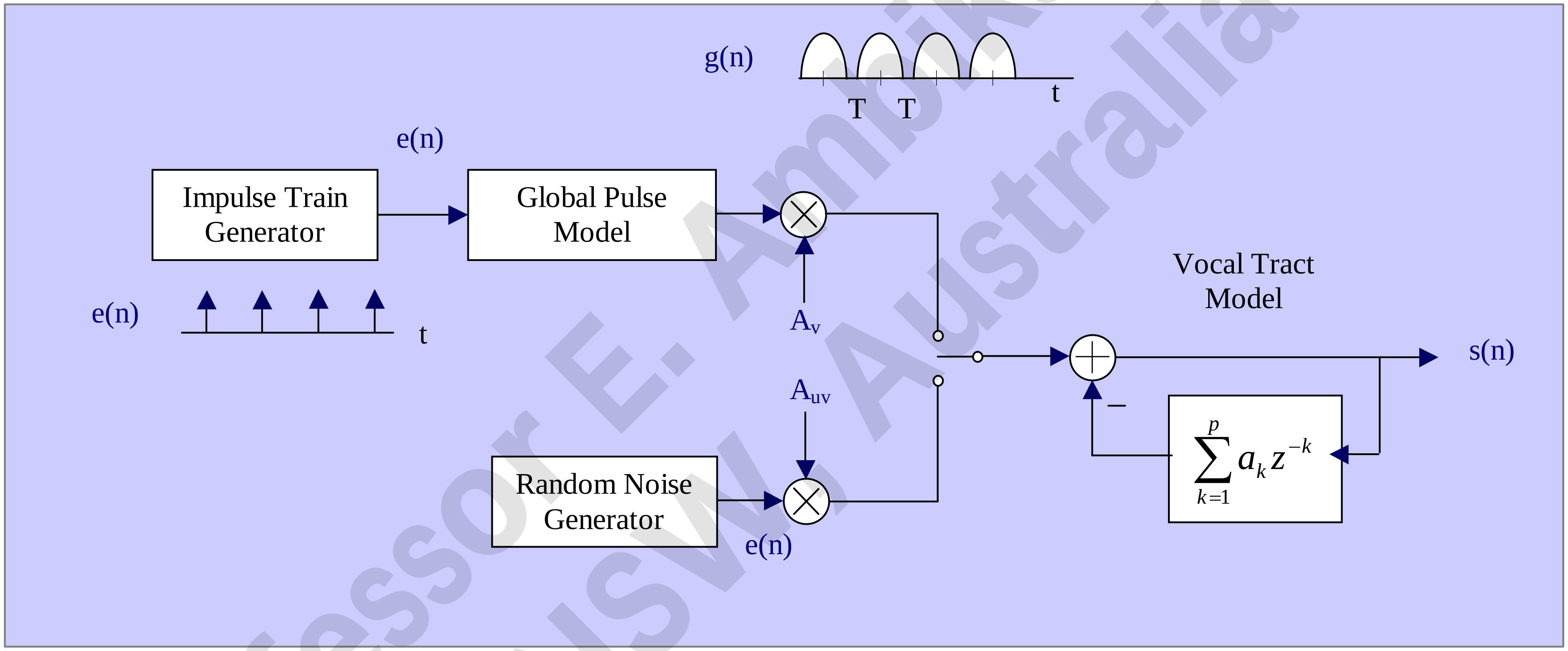
$$\frac{S(z)}{E(z)} = \frac{G}{1+\sum_{k=1}^{q} a'_k z^{-k}}$$

- 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

Use of the Vocal Tract Model

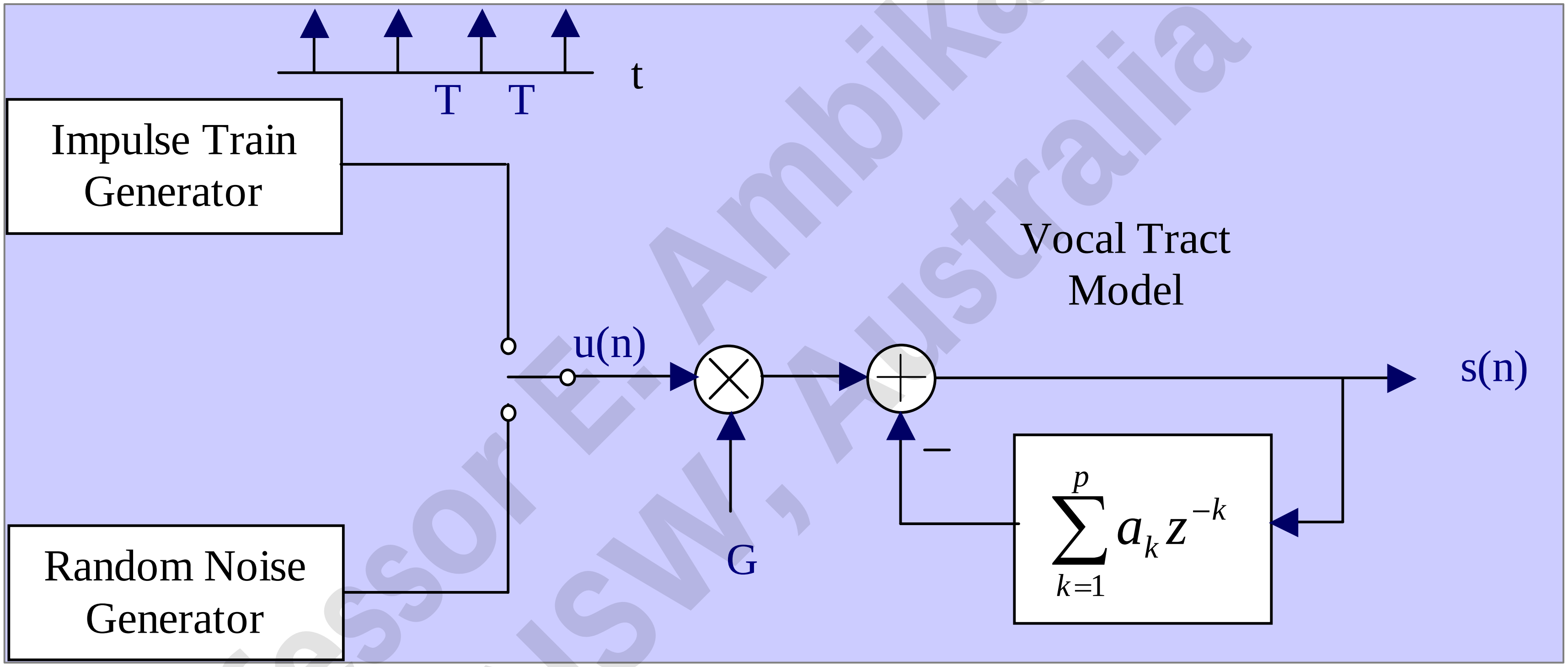
- The model of the vocal tract which has been outlined can be made to be a very accurate model of speech production for short (10-30 ms) frames of speech samples
- It is widely used in modern low bit rate speech coding algorithms, as well as speech synthesis and speech recognition\speaker identification 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 Predictive 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: $a_{k,}$ G, T, v/uv classification

Vocal Tract Model

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

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

$$s[n] = \sum_{k=1}^{q} a_k s[n-k] \quad + Gu[n]$$

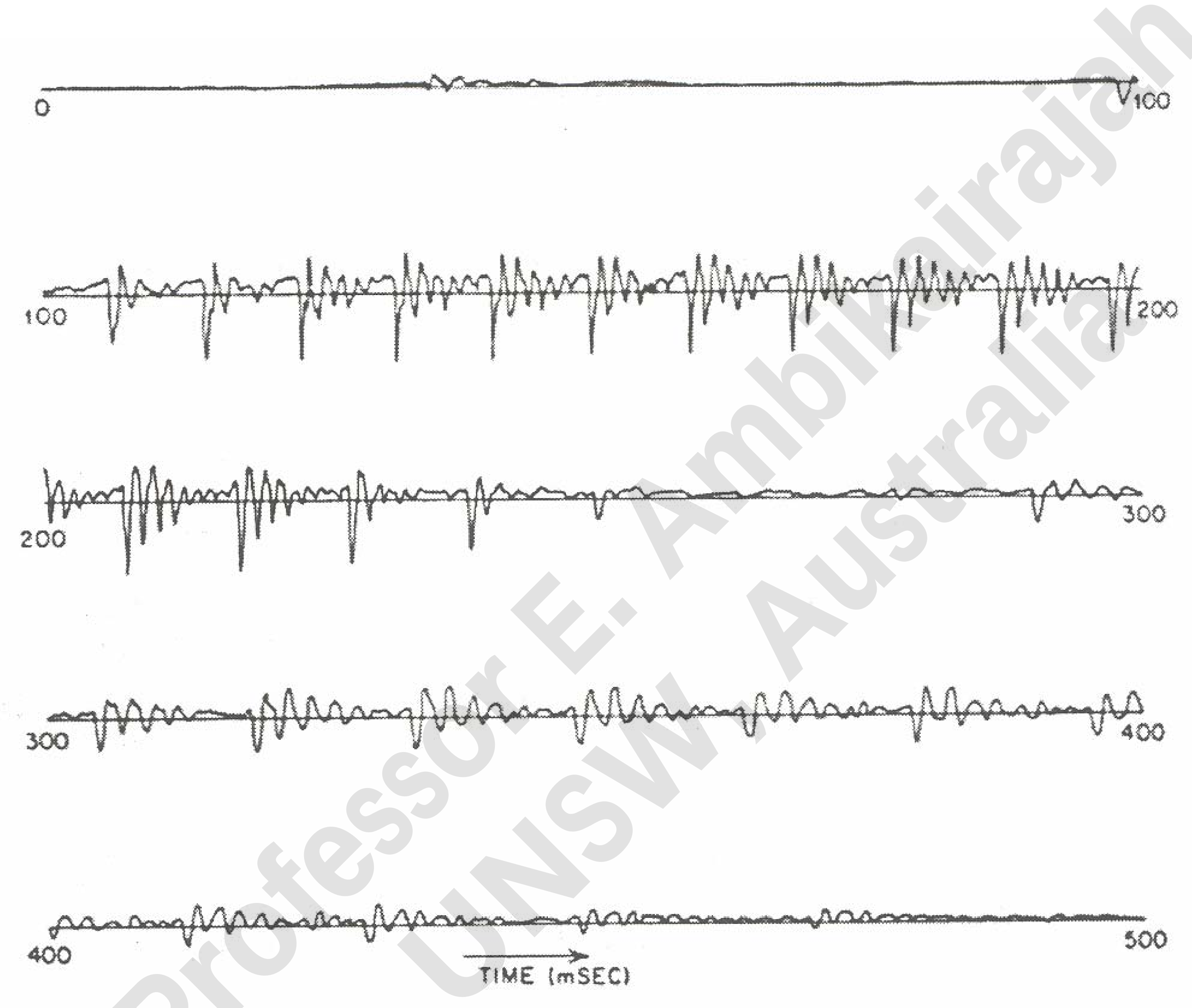
Exercise:

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

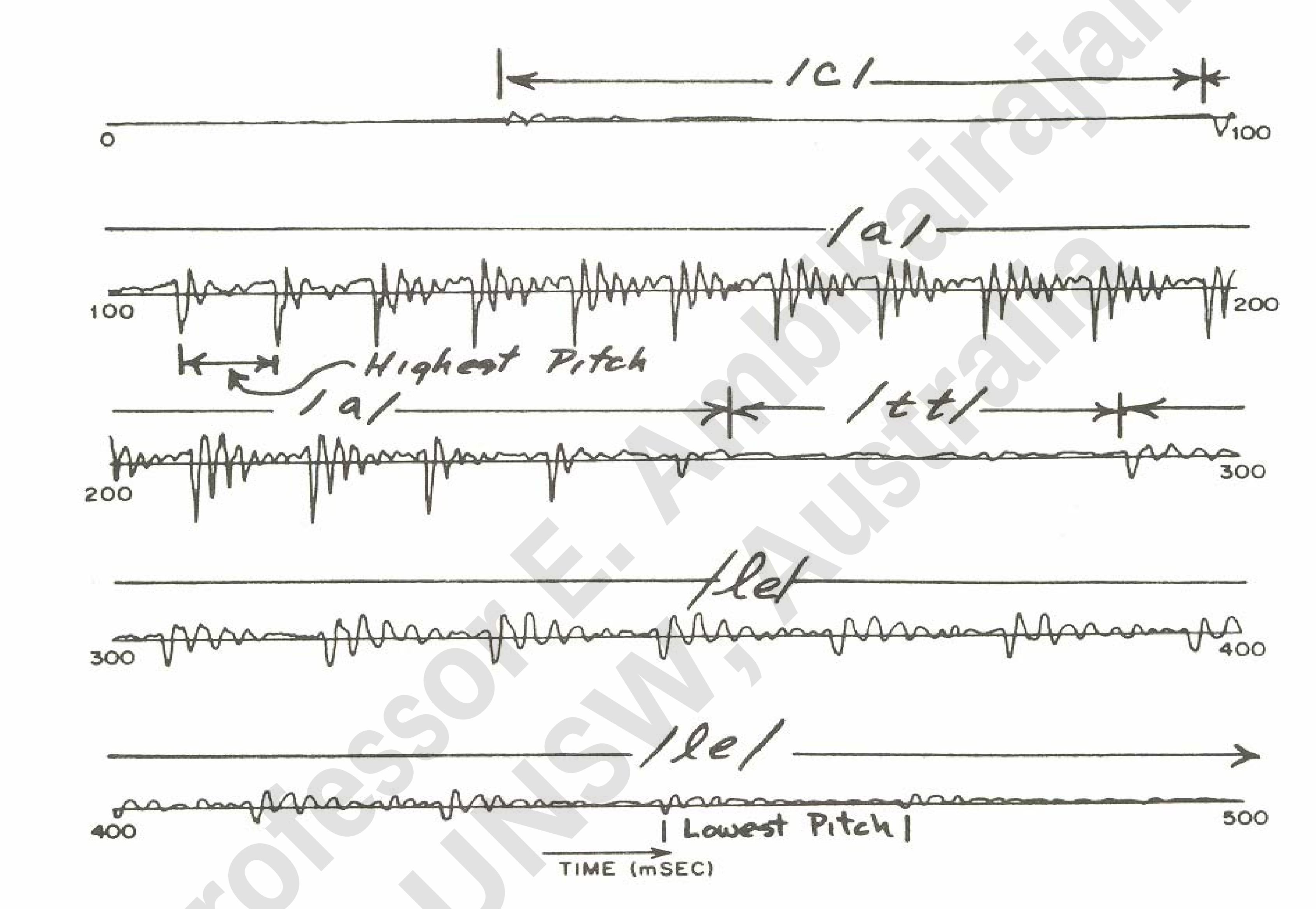
(a) Indicate the boundaries between the phonemes; i.e give the times corresponding to the boundaries /c/a/tt/le/.

(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 $u_L(n)$.

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

(c) Give the locations of the 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.

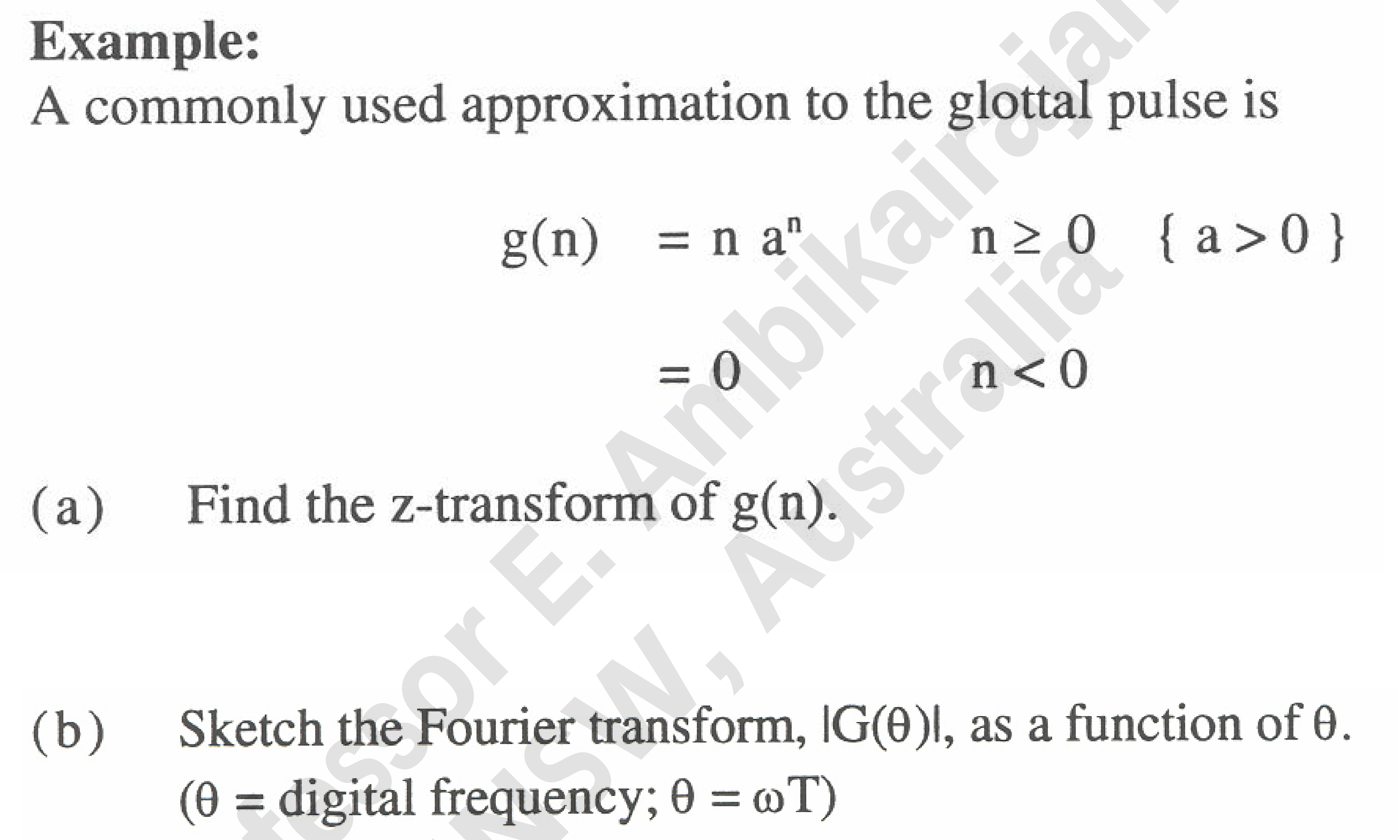
$$\hat{R}(z) = R_0(1 - z^{-1})$$

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

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

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

Therefore R_0 = $R_r/2$ = 0.7205.



Example:

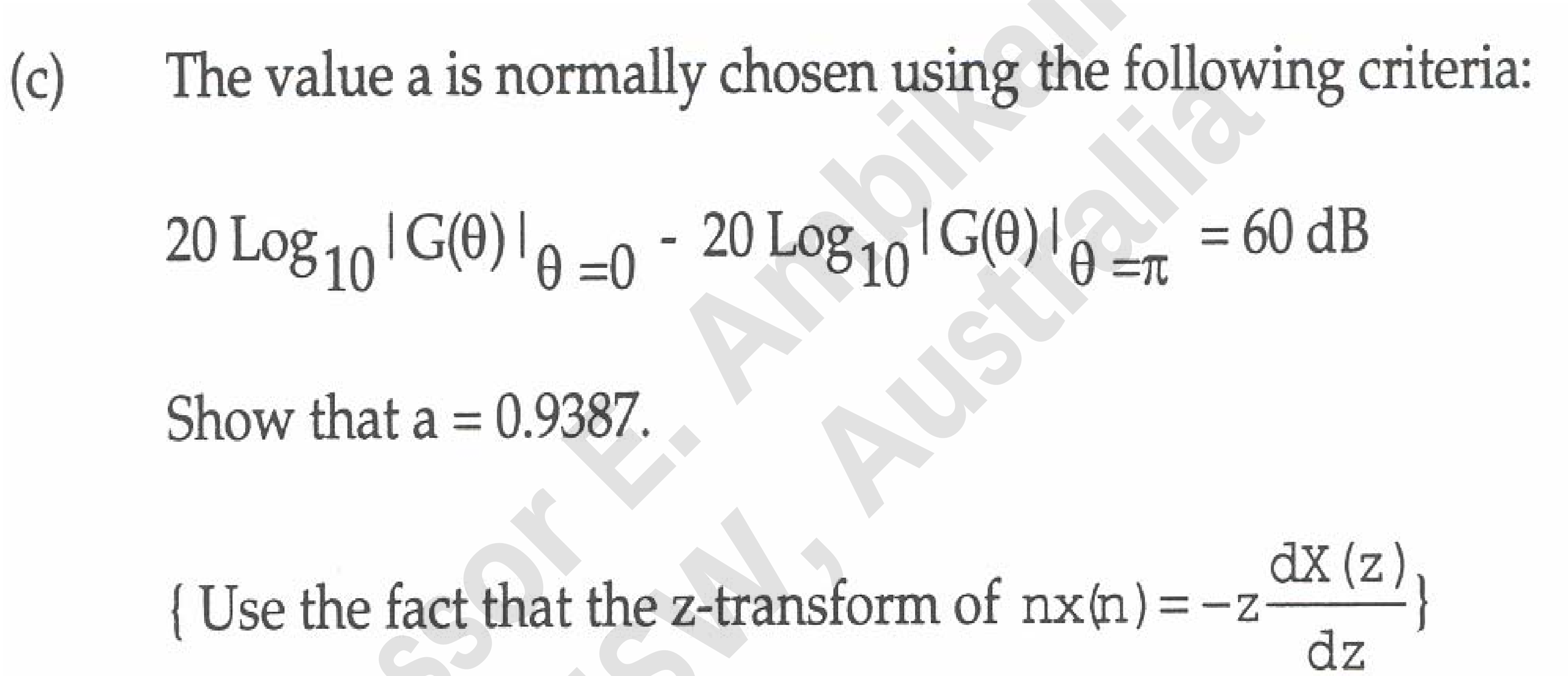
A commonly used approximation to the glottal pulse is

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

$$= 0 \qquad n < 0$$

(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:

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

Show that a = 0.9387.

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

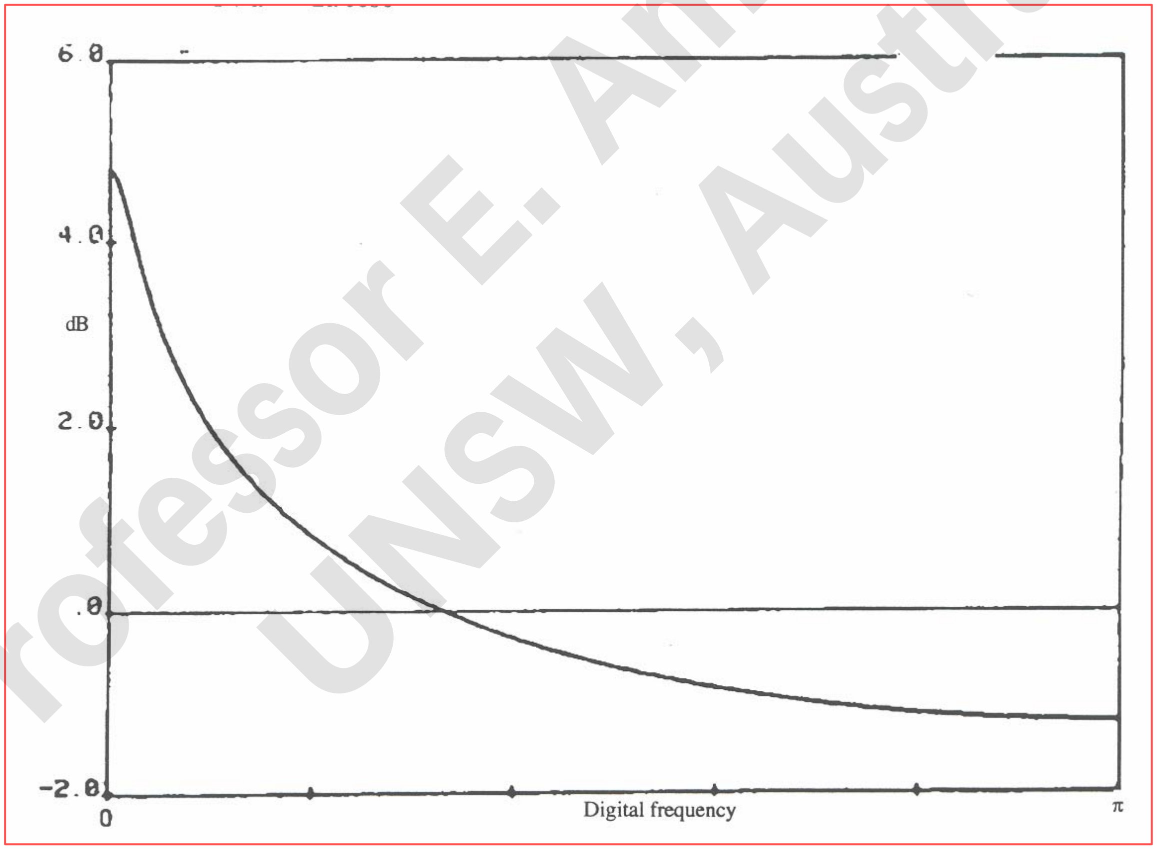
(a) $x(n) = a^{n}$; $X(z) = \dfrac{1}{1-az^{-1}} = \dfrac{z}{z-a}$;

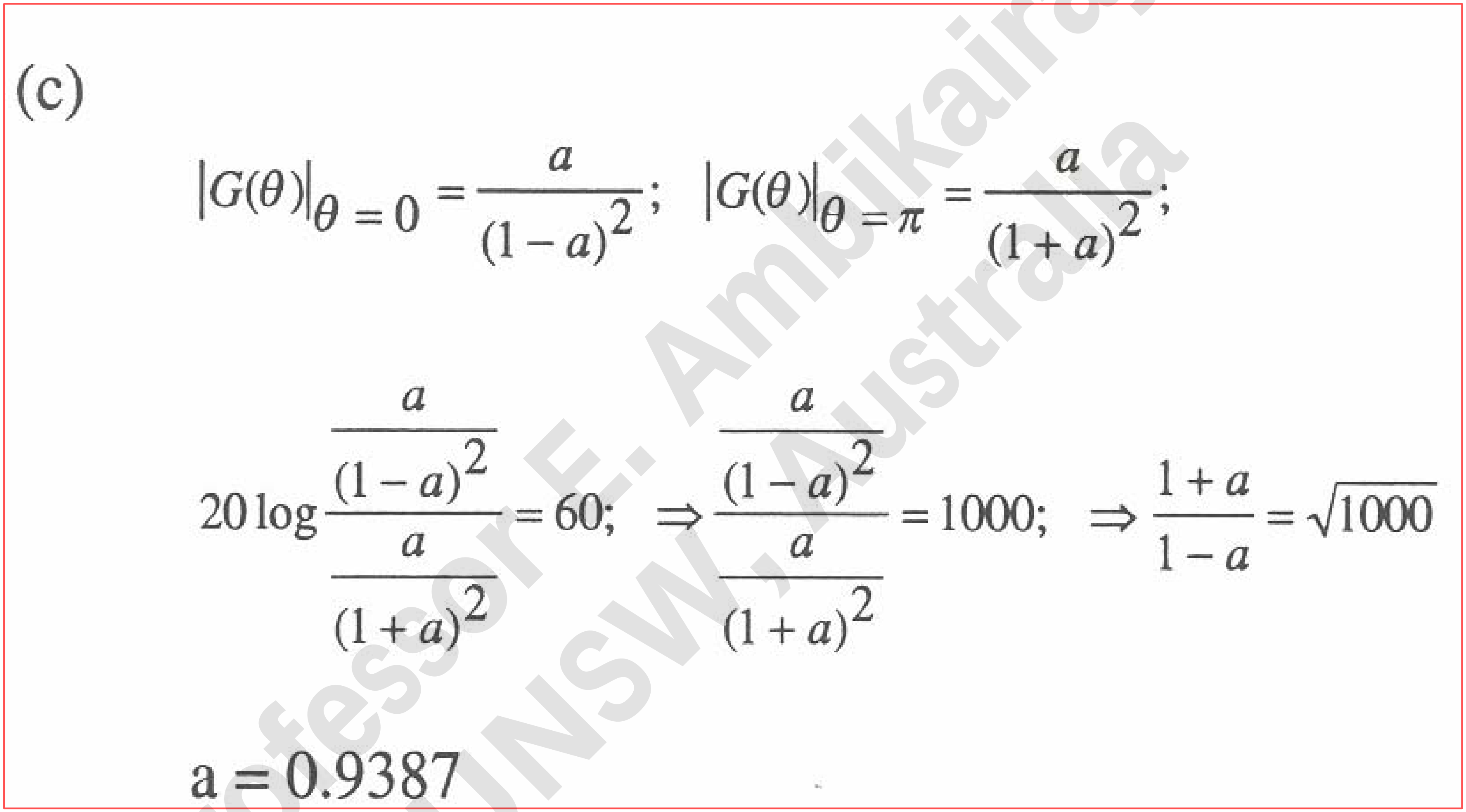
$$\left\{ \frac{dx(z)}{dz} = \frac{-a}{(z-a)^{2}} = \frac{-a}{z^{2}(1-az^{-1})^{2}} \right\}$$

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

(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)

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

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

$a = 0.9387$