ELEC9344:Speech & Audio Processing

Chapter 4

Speech Coding I

Speech Coding

- Speech Coding is a technique used to compress a speech signal for digital storage or transmission purposes and then decompress the stored or transmitted data, so as to allow a perceptually faithful reproduction of the original signal.
- Sophisticated speech coding algorithms exist, so that speech signals may be compressed at various bit-rates.
- It is known that the lower the bit-rate the lesser the speech quality of the recovered speech.
- However, there is a constant quest to achieve a better speech quality at lower bit-rates.

Speech Coders

- There are basically three classes of speech coders:
  - Waveform Coders
  - Source Coders
  - Hybrid coders

- **Waveform Coders** accept continuous analogue speech signals and encode them into digital signals prior to transmission. Decoders at the receivers reverse the encoding process to recover the speech signals. In the absence of transmission errors, the recovered speech waveforms have a close resemblance to the original speech waveforms.

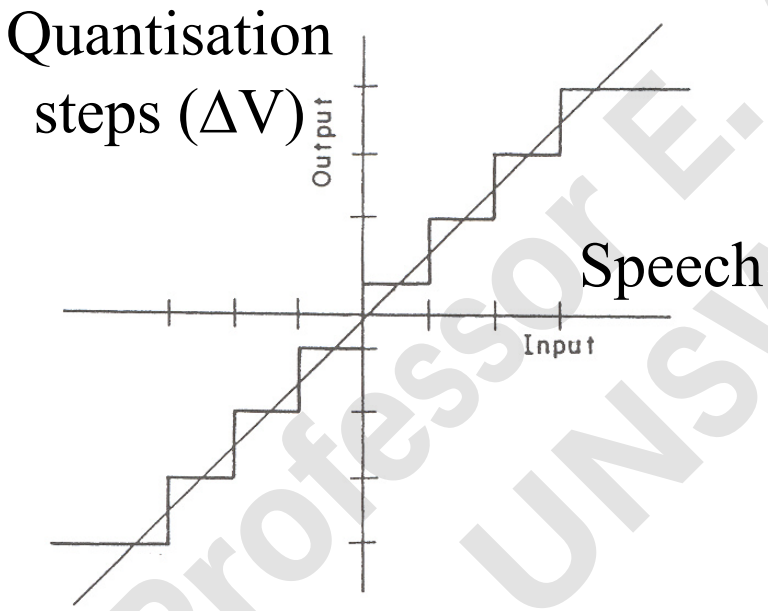
Waveform Encoding Techniques

- The waveform encoding techniques are:
  - PCM (Pulse Code Modulation)
  - DPCM (Differential Pulse Code Modulation)
  - ADPCM (Adaptive Differential Pulse Code Modulation)
  - DM (Delta Modulation)
  - ADM(CVSD) [Adaptive Delta Modulation or Continuously Variable-Slope Delta Modulation]

- The above time-domain coding algorithms outlined above have their counterparts in the frequency domain:
  - SBC (Sub-band coding)
  - ATC (Adaptive Transform Coding)

PCM

- The simplest waveform coding method is linear pulse code modulation. The analogue signals are quantised linearly (uniform quantisation- all steps are equal size). No compression involved.



Linear Quantisation

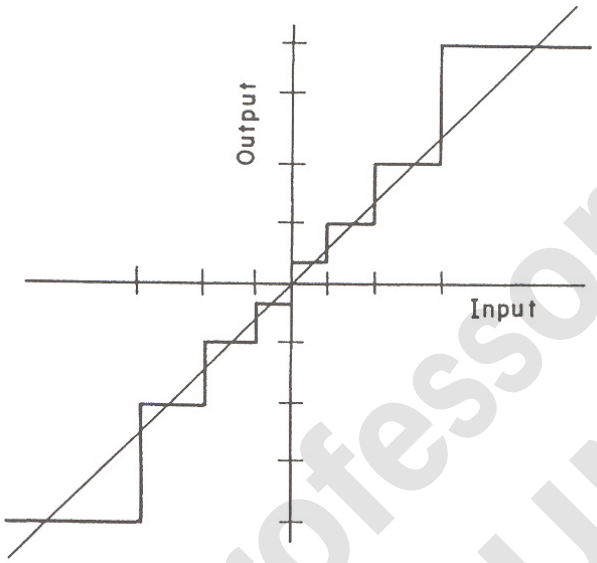
We can show that the Signal to Quantisation Noise Ratio (SQNR) is given by:

$$SQNR = 6.02B + 1.76 \quad dB$$

- The SQNR increases approximately 6dB for each bit.
- Assumes Uniform PDF

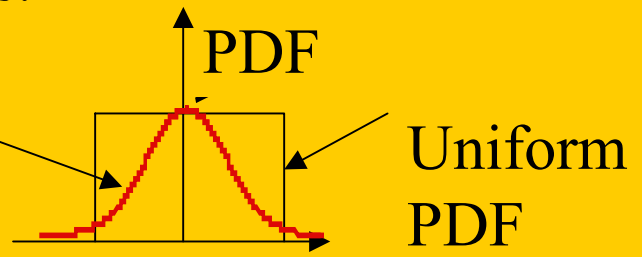
Non-Uniform PCM

- We know that the speech signals are heavily concentrated in the low amplitudes and hence it is a much better strategy to use nonuniform quantiser in which the steps are densest at the low levels



Nonlinear Quantisation

Speech PDF is not uniform but Gaussian and lower level speech occurs more frequently than higher level speech. Therefore more quantisation noise is tolerable for high levels of speech than at lower levels.



DPCM

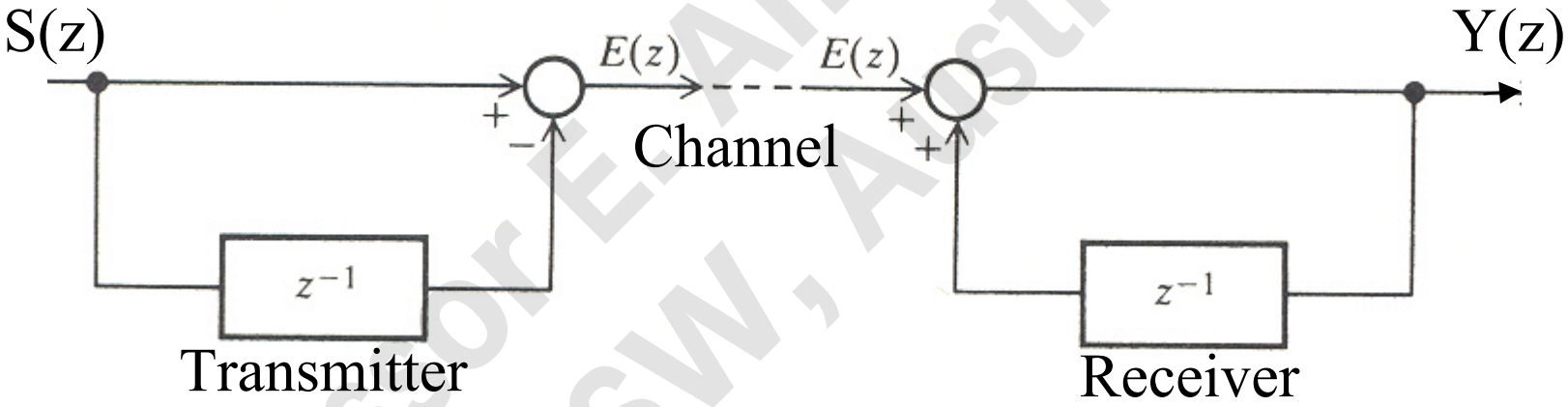
- By taking the advantage of considerable amount of redundancy available in the speech signal, the bitrate may be reduced.
- For example, if we take the difference of the speech signal (s(n)- s(n-1)), this will have a lower variance (and hence lower power) than signal itself.
- If the power is reduced, then we should be able to encode the signal with fewer bits and thereby reduce the bit rate required.

$$e(n) = s(n)-s(n-1)$$

$$E(z) = S(z)[1-z^{-1}]$$

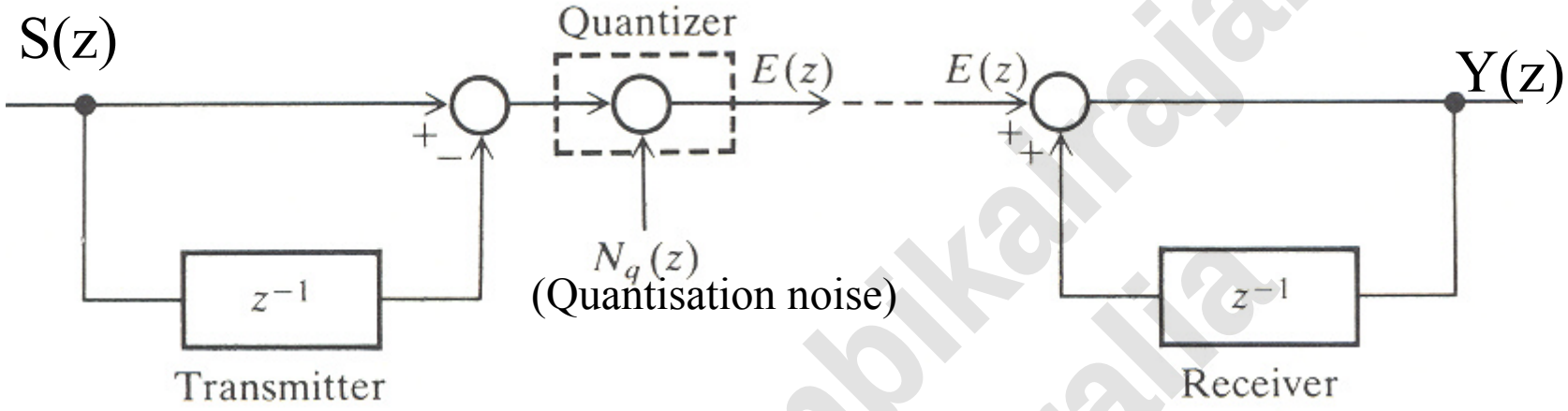
DPCM

- The Transmitter computes e(n) and this difference is transmitted through the channel to the receiver (see below), where it is integrated in the feedback loop and the original signal is restored



$$E(z) = S(z)(1 - z^{-1})$$

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

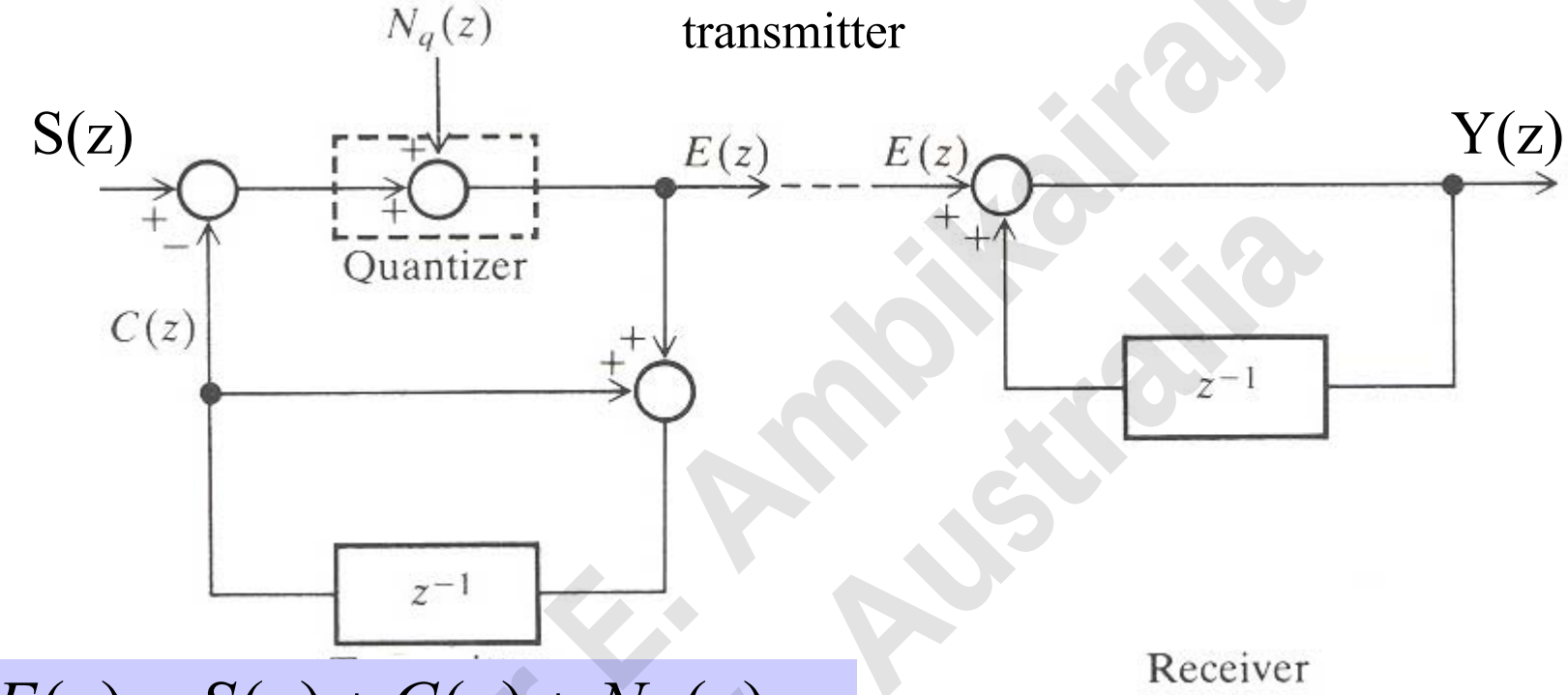


DPCM with quantisation of the difference signal. The quantisation noise is modelled by additive noise

$$Y(z) = S(z) + \frac{N_q(z)}{1 - z^{-1}}$$

Thus the quantisation noise is integrated at the receiver, and this build up of noise is found to be more disturbing as compared to the same noise had it been merely added to the original speech signal.

Differential quantisation with the quantiser inside a feedback loop at the transmitter



$$E(z) = S(z) + C(z) + N_q(z)$$

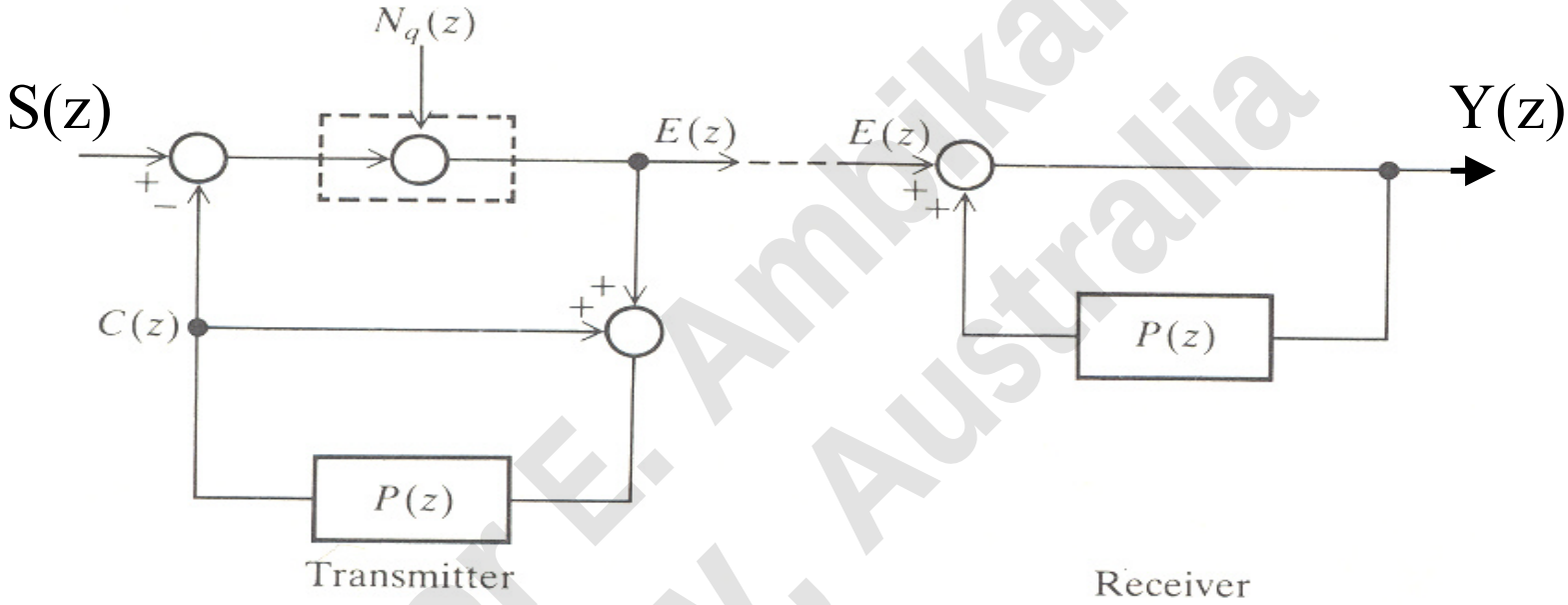
$$E(z) = [S(z) + N_q(z)](1 - z^{-1})$$

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

$$Y(z) = S(z) + N_q(z)$$

← The quantisation noise is being differenced along with s(n)

Next refinement is to recognise that simple differencing does not minimise the error of the output. Replace the simple differencer by a full p^{th} order linear predictor (P(z))



$$E(z)=[S(z)+N_q(z)](1-P(z))$$

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

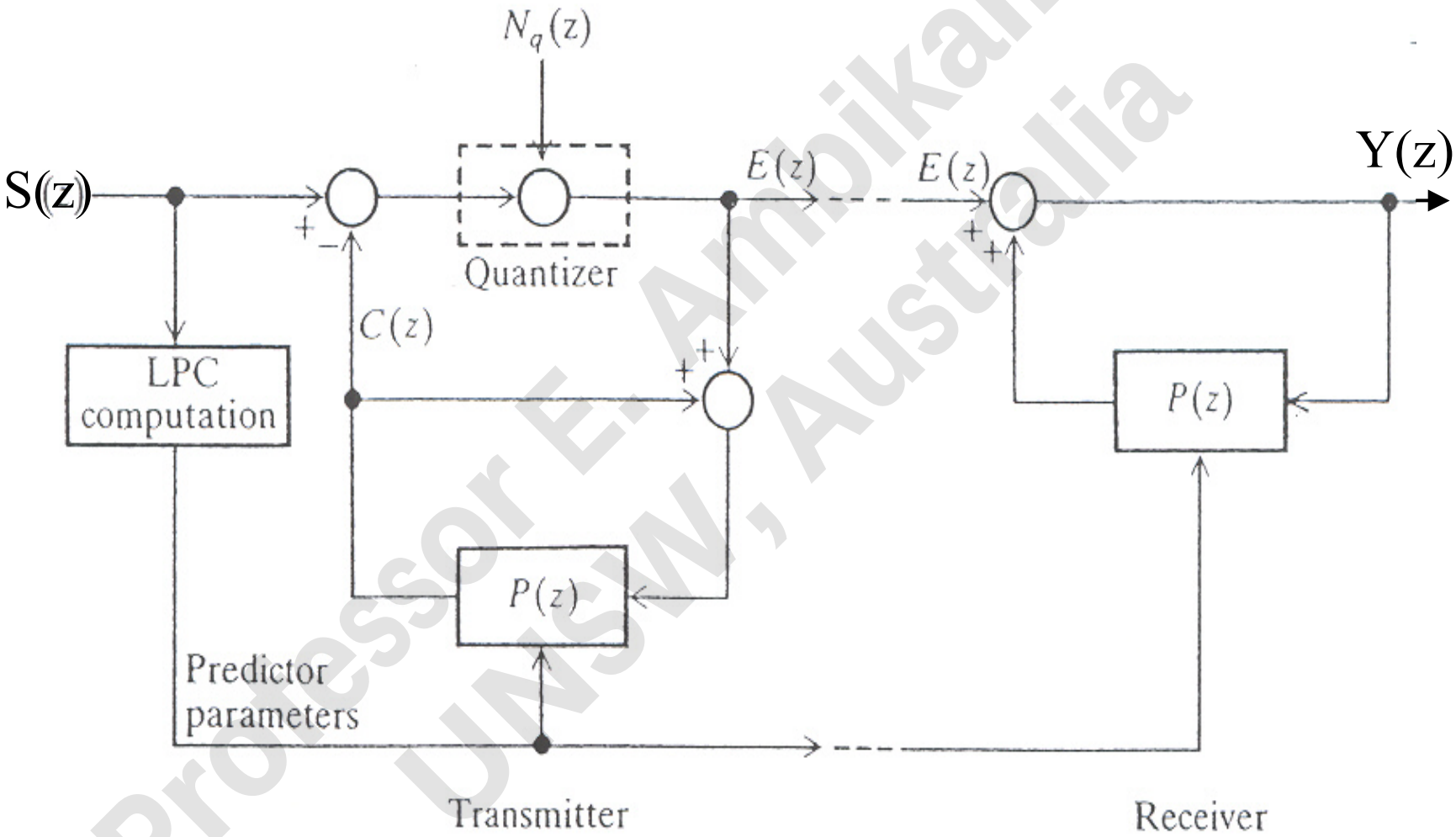
$$Y(z)=S(z)+N_q(z)$$

Using such a predictor, a reduction in variance of 6dB relative to PCM has been reported (p =2). If the reduction in variance is 6dB means, we can use one less bit in quantising the residual than would be required if the original signal were sent.

ADPCM

- The compression ratio of DPCM can be further improved by adaptive prediction. Adaptive prediction in its simplest form is simply linear prediction.
- The speech data is segmented into frames of typically 10 to 20 ms and the predictor coefficients are transmitted along with the residual (see next slide).
- At the receiver, an inverse filter controlled by the predictor coefficients reconstructs the speech.
- It is reported that the S/N improvement of more than 12 dB can be achieved for a 10^{th} order linear predictor.

ADPCM



Summary of Waveform Coders

- A Law or μ Law PCM: 8 bit compressed PCM, 8 kHz sampling rate; 64 kbps
- DPCM: 32 kbps
- ADPCM: 24-32 kbps

Source Coders

- Source Coders such as Linear Predictive Coding (LPC) vocoders, which are based on the speech production model, operate at bit rates as low as 2 kbits/s. However, the synthetic quality of the vocoded speech is not broadly appropriate for commercial telephone applications.

- LPC in its basic form has been mainly used in secure military communications where speech must be carried out at very low bit rates. A US standard algorithm known as LPC-10 is widely used.

LPC-10 Algorithm

- The input speech is sampled at 8 kHz and digitised as 12 bit two's complement words. The incoming speech is partioned into 180 sample frames (22.5 ms), resulting in a framerate of 44.44 frames/sec.
- A tenth order linear predictor is used to compute the vocal tract filter parameters (reflection coefficients)
- Pitch and voicing are determined using AMDF algorithm.

LPC-10 Algorithm

- The bit assignments used are: 5 bits each for k_1 to k_2, 4 bits each for k_5 through k_8, 3 bits for k_9 and 2 bits for k_{10}.
- In the transmitted bit stream, 41 bits are used for the reflection coefficients, 7 bits for pitch and voicing, and 5 bits for amplitude.
- One bit is used for synchronisation, giving a total of 54 bits per frame.
- Since there are 44.44 frames/s, this gives an overall bit rate of 2.4 kbits/s.

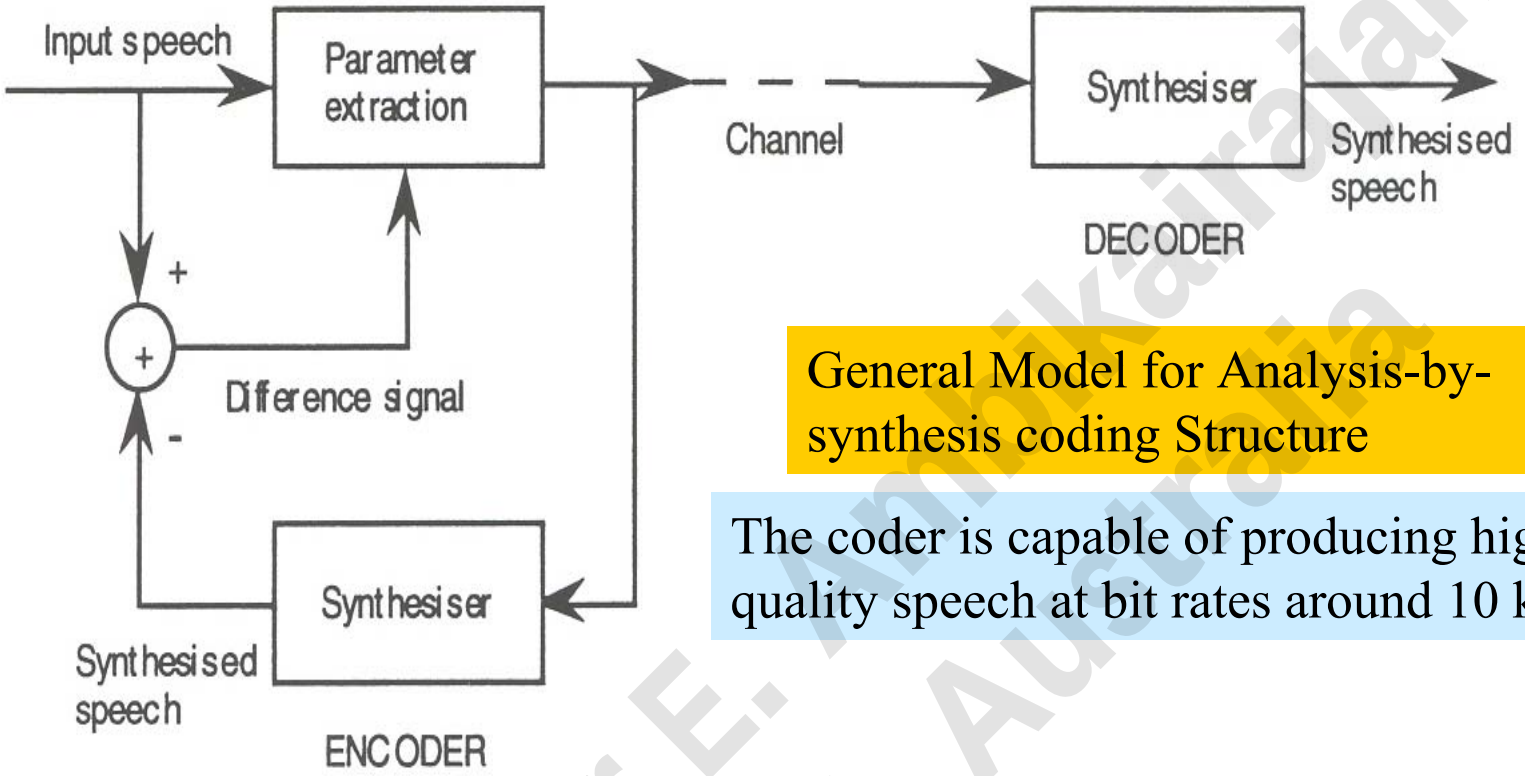
Limitation of LPC Vocoding

- The main limitation of LPC vocoding is the assumption that speech signals are either voiced or unvoiced, hence the source of excitation of the synthesis all-pole filter is either a train of pulses (for voiced speech), or random noise (for unvoiced speech).
- In fact there are more than two modes in which the vocal tract is excited and often these modes are mixed.
- Even when the speech waveform is voiced, it is a gross simplification to assume that there is only one point of excitation in the entire pitch period.

Hybrid Coders

- Hybrid coders combine features from both source coders and waveform coders. Several hybrid coders employ an analysis-by-synthesis process in order to derive code parameters.
- A block diagram of an analysis-by-synthesis coding system is shown below.
- A local decoder is present inside the encoder and the aim of the analysis-by-synthesis speech coding structure is to derive some codec parameters so that the difference between the input and synthesised signal is minimised according to some suitable error minimisation criterion.

LPC vocoder



General Model for Analysis-by-synthesis coding Structure

The coder is capable of producing high quality speech at bit rates around 10 kbits/s

In this new generation of analysis-by-synthesis models, no prior knowledge of a voiced/unvoiced decision or pitch period is needed. The excitation is modelled by a number of pulses, usually 4 per 5ms, whose amplitudes and positions are determined by the perceptually weighted error between the original and synthesised speech.

The multipulse approach assumes that both the pulse positions and amplitudes are initially unknown, they are then determined inside the minimisation loop one pulse at a time.

Hybrid Coders……

- When the bit rate is reduced below 9.6 kbit/s, the multi-pulse excited coders fail to maintain good speech quality.
- This is due to large number of bits needed to encode excitation pulses and the quality deteriorates when these pulses are coarsely quantised.
- Therefore, if analysis-by-synthesis structure is to be used for producing good quality speech at bit rates below 8 kbits/s, more suitable approaches to the definition of the excitation signal must be sought.
- The Code-Excited Linear Prediction (CELP) coder provided an improved excitation signal.

Speech Coding Standards

- For speech coding to be useful in telecommunication applications, it has to be standardised (it must conform to the same algorithm and bit format).
- Speech-coding standards are established by various standards organisations: For example, ITU-T (International Telecommunication Union, Telecoomuncation Standardisation Sector (formally CCITT), ETSI (European Telecommunications Standards Institute) etc

ITU standard

- G711/712: 64 kbits/s A/μ Law PCM
- G726 : 32 kbits/s ADPCM
- G728: 16 kbits/s LD-CELP (Low Delay CELP)
- G729: 8 kbits/s CS-ACELP (Conjugate Structure Algebraic CELP)
- The next standardisation issue involves a 4 kbit/s speech coding algorithm. The main terms of reference for the ITU 4 kbits/s speech coding standard are given below.

Parameter	Requirement
Speech quality in error free condition	Not worse than that of G. 726 at 32 kbit/s
BER = 10^{-3} (Random Errors) The Bit Error Ratio is defined at the speech decoder input after channel decoding.	Not worse than that of 16 kbit/s LD-CELP (CCITT Rec. G.728) under similar conditions
Detected Frame Erasures (3% random). The Detected frame erasures require that the decoder be informed that the incoming encoded bit-stream is in error and that the associated frame may be rejected.	No more than 0,5 MOS degradation from 16 kbit/s LD-CELP under error-free condition.
One way Coder/decoder delay: - algorithmic delay -total CODEC delay	 ≤20 ms with ≤10 ms objective ≤40 ms with ≤20 ms objective
Speech quality dependency on the input signal level between -36 dB and -16 dB with respect to the overload point.	Not worse than that of CCITT Rec. G.726 at 32 kbit/sec (as low as possible)
Capability to transmit signalling and information tones.	DTMF, CCITT No. 5&6, CCITT R2,Q.35, Q.23, V.25. The tones have to be transmitted with as little distortion as possible.
Tandeming capability of speech	2 asynchronous with a total distortion ≤ 4 asynchronous G.726 at 32 kbit/s.
Implementation	Fixed-Point

Low Bit rate Speech Coding

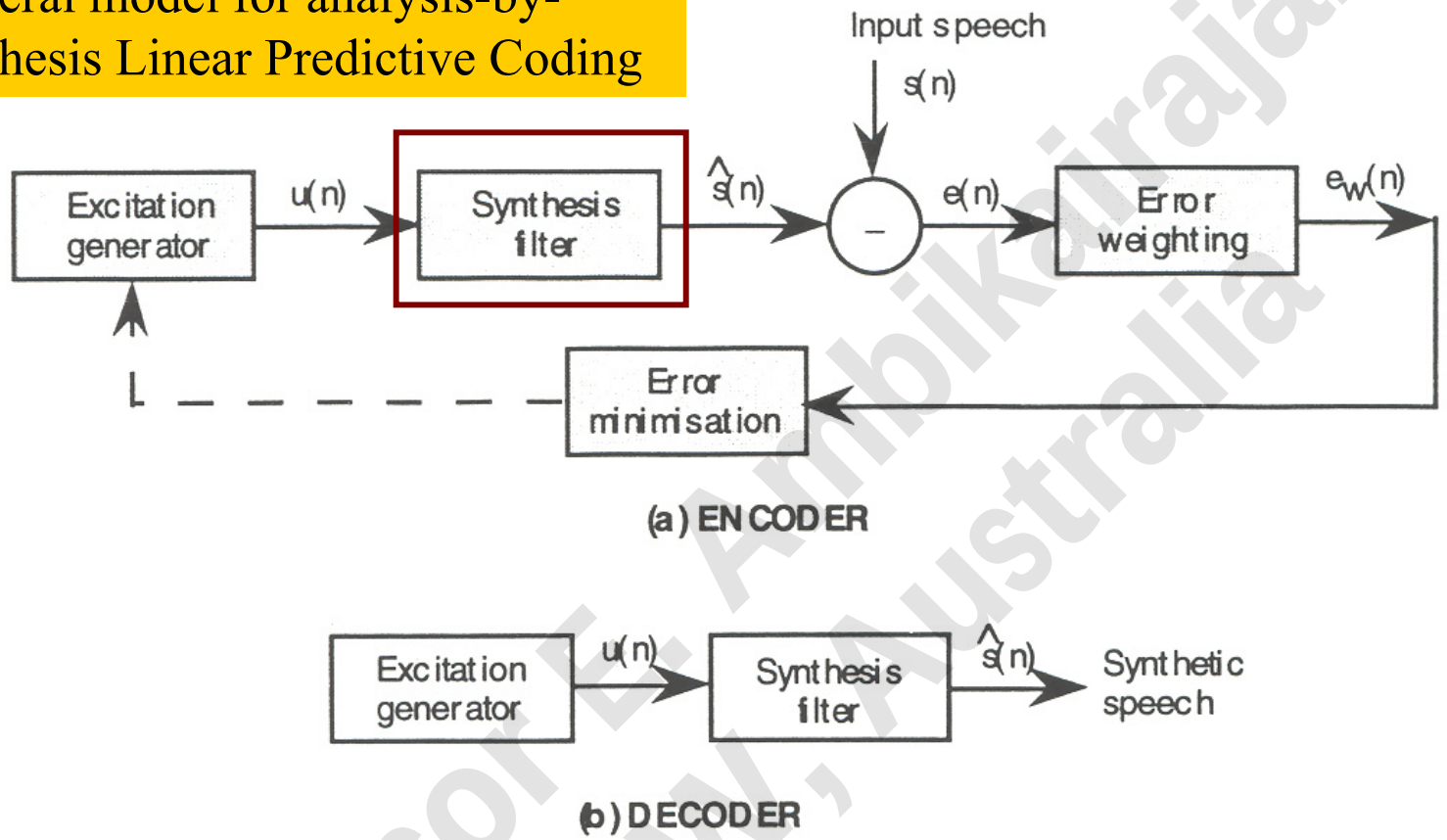
- It is important to develop low bit-rate speech coding systems for voice storage such as voice mail and voice driven directory enquiries
- Therefore , the speech research on speech coding methods whose bit rate is less than 4 kbits/s but still provide toll quality speech has recently become very active.
- The ultimate goal is to design a low-delay, low bit rate codec (in real time) (around 2.4 kbits/s) with good perceptual quality for digital mobile communications and other applications.

Analysis-by-Synthesis Speech Coding

- As the Code-Excited Linear Prediction (CELP) coder is based on the analysis-by-analysis technique, first the theory behind the analysis-by-synthesis method is explained and then it is extended to CELP coding.

- The basic structure of the general model for analysis-by-synthesis predictive coding of speech is shown below.

General model for analysis-by-synthesis Linear Predictive Coding



- The model consists of four main parts:
  - Excitation Generator
  - Synthesis Filter
  - Error Weighting filter
  - Error minimisation

The Short-Term Predictor (Synthesis Filter)

- The synthesis filter is an all-pole time varying filter for modelling the short-time spectral envelope of the speech waveform.
- It is often called a short-term prediction filter because its coefficients are computed by predicting a speech sample from a few previous (8 to 16 samples) samples.
- The set of coefficients is called Linear predictive Coding (LPC) parameters.

- The spectral envelope of a speech segment of length L samples can be approximated by the transfer function of an all-pole digital filter of the form .

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

- The coefficients a_k are computed using the method of linear prediction . The number of coefficients p is called the predictor order.

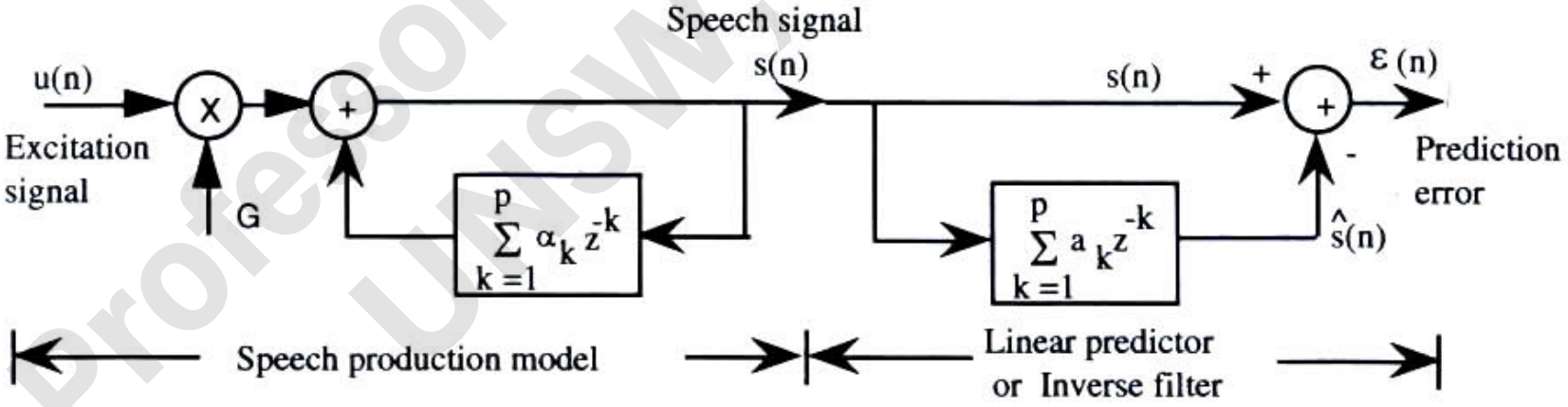
- The basic idea behind the linear predictive analysis is that a speech sample can be approximated as a linear combination of 'p' past speech samples i.e.

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k)$$

- Where s(n) is the speech sample and $\hat{s}(n)$ is the predicted speech sample at sampling instant n. The prediction error, $\varepsilon(n)$ is defined as

$$\varepsilon(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$$

- The prediction error signal, $\varepsilon(n)$ is shown in the diagram:



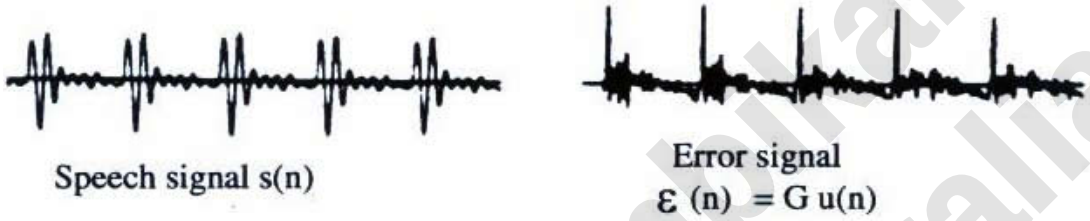
Linear Predictor

- It can be seen from the above figure when $a_k = \alpha_k$, then $\varepsilon(n) = G\ u(n)$ and the linear prediction is called an inverse filter. The transfer function of the inverse filter is given by

$$A(z) = \frac{\mathcal{E}(z)}{S(z)} = 1 - \sum_{k=1}^{p} a_k z^{-k}$$

- For voiced speech this means that $\varepsilon(n)$ would consist of a train of impulses; $\varepsilon(n)$ would be small most of the time except at the beginning of the pitch period .

- An examples of a signal s(n) and a prediction error for the vowel ‘a’ is shown below.

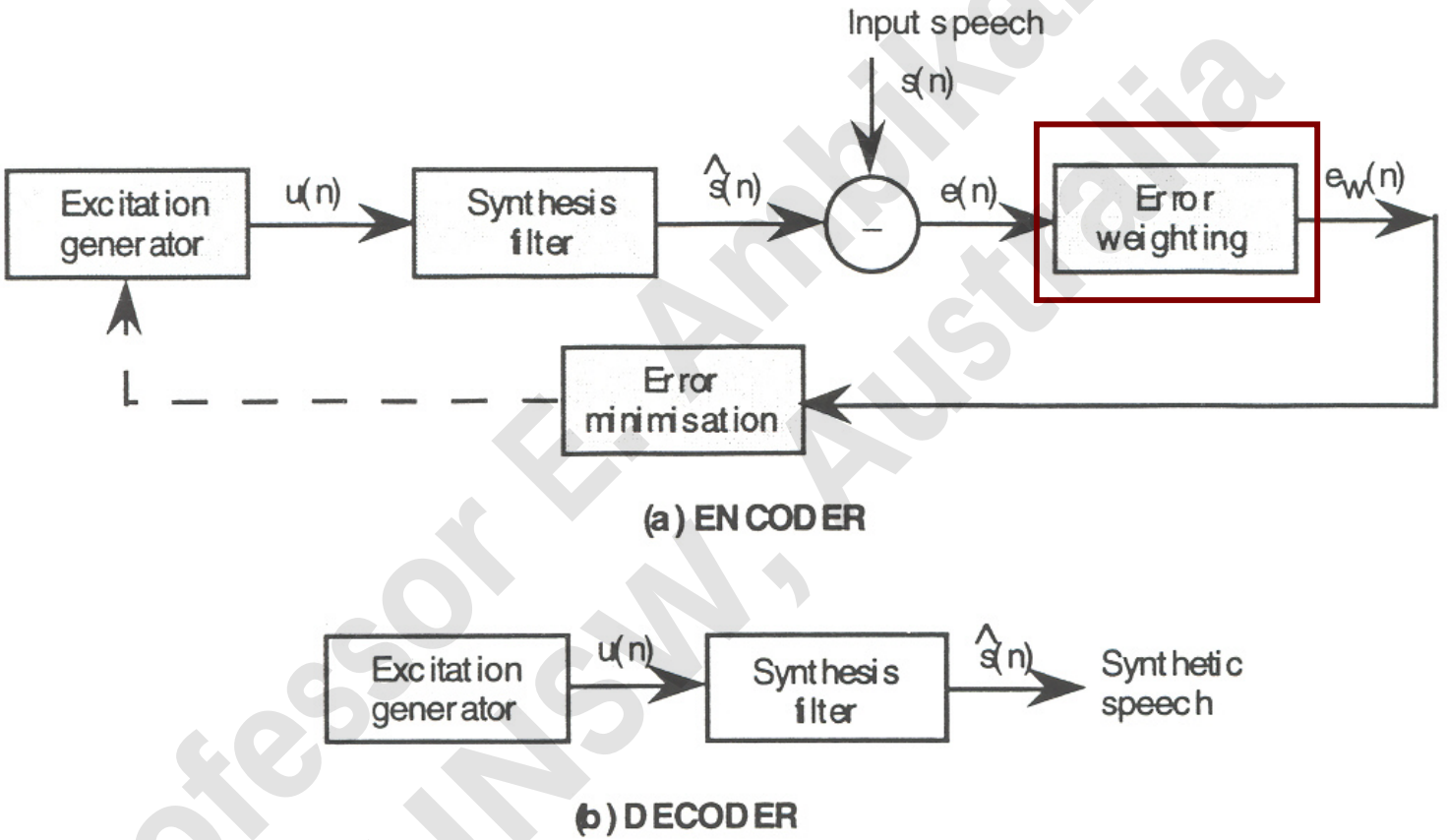


Prediction Error

- Because of the time-varying nature of the speech, the prediction coefficients should be estimated from short segments of speech signal (10-20 ms).

- The basic approach is to find a set of predictor coefficients a_k that will minimize the mean-squared prediction error over a short segment of speech waveform.

General model for analysis-by-synthesis Linear Predictive Coding



The Error Weighting Filter

- The function of the perceptual error weighting filter, W(z), in the above diagram is to reduce a large part of perceived noise in the coder which comes from the frequency region where signal is low.

- The theory of auditory masking suggests that noise in the formant regions would be particularly or totally masked by the speech signal.

- Therefore to reduce perceived noise , the error spectrum E(z) is shaped (see figure previous slide) so that the frequency component in the noise around the formant regions are allowed to have higher energy relative to the components in the inter-formant regions.

- The coefficients for the error weighting filter are derived from the LPC coefficients.
- The weighting filter W(z) can be expressed as,

$$W(z) = \frac{A(z)}{A(z/\gamma_D)} = \frac{1 - \sum_{k=1}^{p} a_k z^{-k}}{1 - \sum_{k=1}^{p} a_k \left(\frac{z^{-k}}{\gamma_D^{-k}}\right)} = \frac{1 - \sum_{k=1}^{p} a_k z^{-k}}{1 - \sum_{k=1}^{p} a_k z^{-k} \gamma_D^{k}}$$

where $A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}$, and $0 < \gamma_D < 1$.

The value γ_D is determined by the degree to which one wishes to deemphasize the formant regions in the error spectrum E(z).

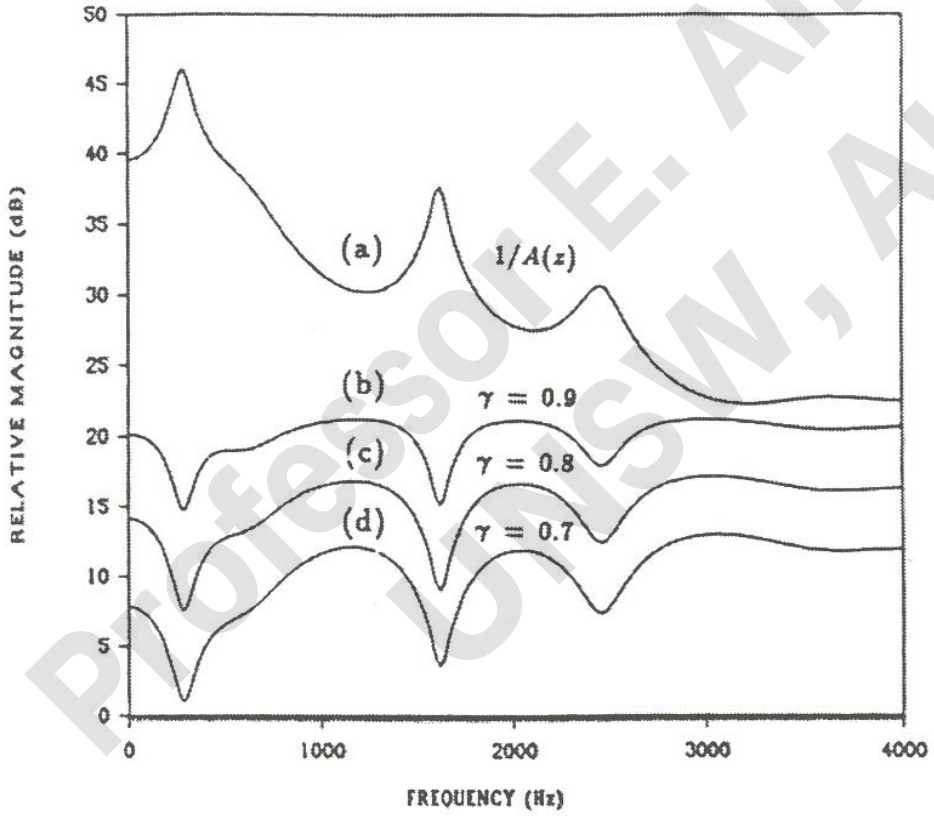
- Note that decreasing $\gamma_D = 1$ increases the bandwidth of the poles of W(z).
- If $\gamma_D = 1$ and W(z)=1 which is equivalent to no weighting .
- A good choice is to use a value of γ_D between 0.8 and 0.9.
- In this coder γ_D is chosen as 0.9.
- Note that decreasing γ_D increases the bandwidth of the poles of $\left\{\frac{A(z)}{A'(z)}\right\}$
- The increase in the bandwidth $\Delta\omega$ is given by

$$\Delta\omega = -(f_s/\pi\ )\ \ln(\gamma_D)$$

- Where f_s the sampling frequency

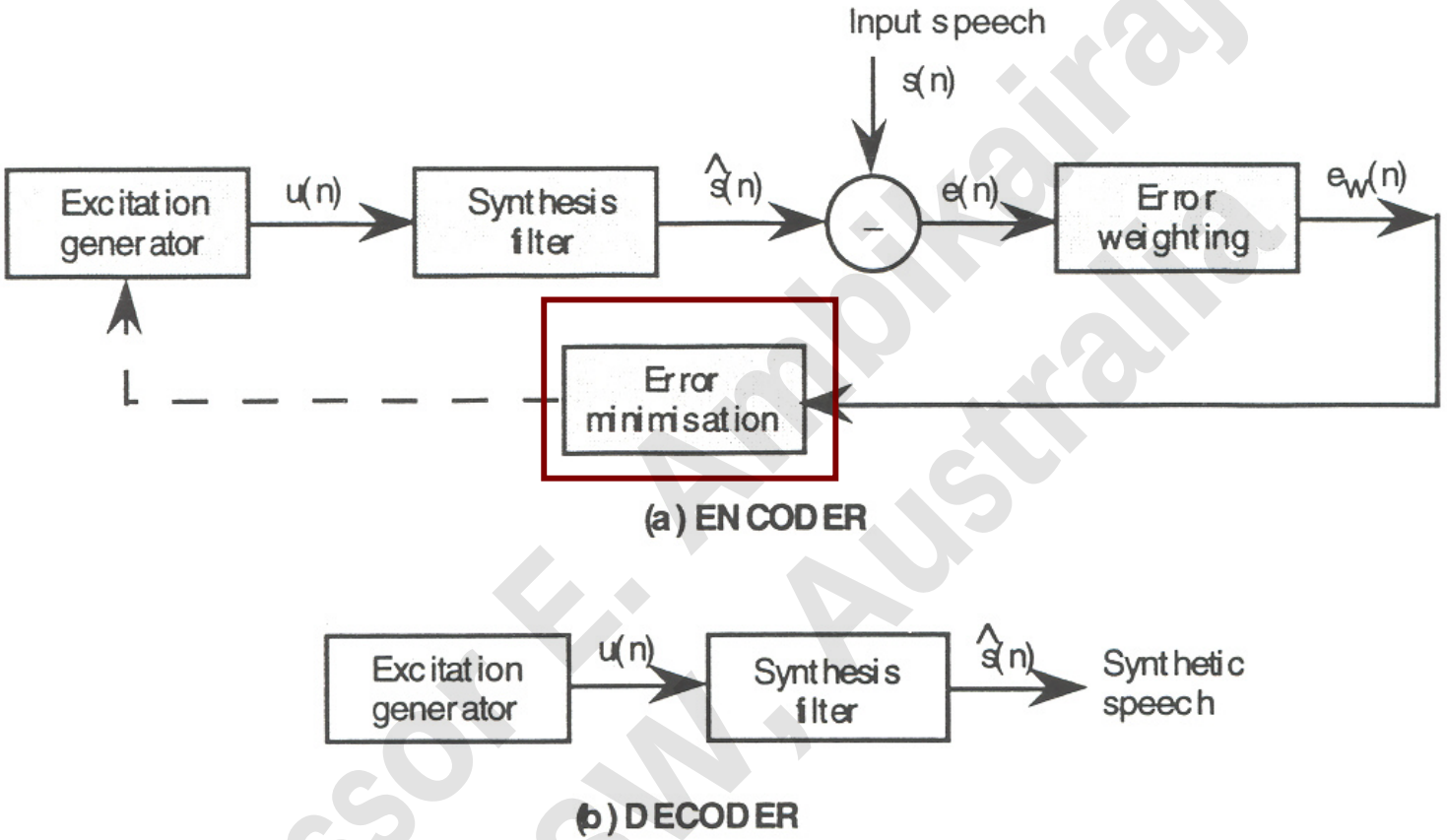
- Figure below shows an example of the spectrum of 1/A(z) and $A(z) \Big/ A\left(\frac{z}{\gamma_D}\right)$ for different values of γ_D

- An increase in bandwidth of the poles is noticeable when γ_D changed from 0.9 to 0.8



Spectrum of 1/A(z) and A(z)/A(z/γ) for different values of γ

The Error Minimization



➢The excitation generator produces the excitation sequence which is fed to the synthesis filter to produce the reconstructed speech.

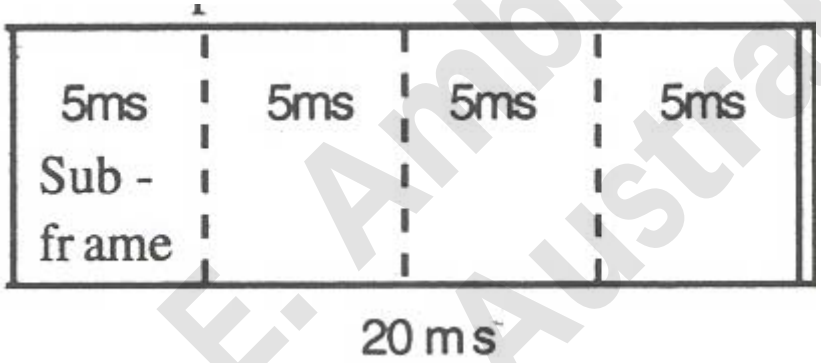
The Error Minimization

- The most common error minimization criterion is the mean squared error.
- In the model (see previous slide) subjectively meaningful error minimization criterion is used, where the error e(n) is passed through a perceptual weighting filter.
- The excitation sequence is optimized by minimizing the perceptually weighted error between the original and synthesized speech.

The Encoder

- In the encoding procedure, the synthesis filter parameters (LPC coeffiecients are determined from speech samples (20 ms of speech is a frame = 160 samples) outside the optimisation loop.
- The optimum excitation sequence for this synthesis filter is then determined by minimising the weighted error criterion.
- The excitation optimisation interval is usually in the range of 4 to 7.5 ms which is less than the LPC coefficients update frame size.

- Thus the speech frame is divided into sub frames (normally four 5 ms sub frames = 4x40 samples), where the excitation is determined individually for each sub frame.



- A new set of LPC coefficients is transmitted every 20ms, although interpolation of LPC parameters between adjacent frame can be used to obtain different set of parameters for every sub frame.

- This interpolation enables the updating of the filter parameters every 5 ms while transmitting them to the decoder every 20 ms. i.e. without requiring the higher bit rate associated with shorter updating frames.

- The set of predictor coefficients, $a_{k,}$, cannot be used for interpolation ,because the interpolated parameters in this case do not guarantee a stable synthesis filter.

- The interpolation is, therefore, performed using a transformed set of parameters where the filter stability can be easily guaranteed by using the log-area ratios (LAR), Line Spectral Paris (LSP) or Line Spectrum Frequencies (LSF).

- If F_N is the quantised LPC vector in the present frame and F_{N-1} is the quantised LPC vector from the past frame, then the interpolated LPC vector SF_k in a subframe k is given by

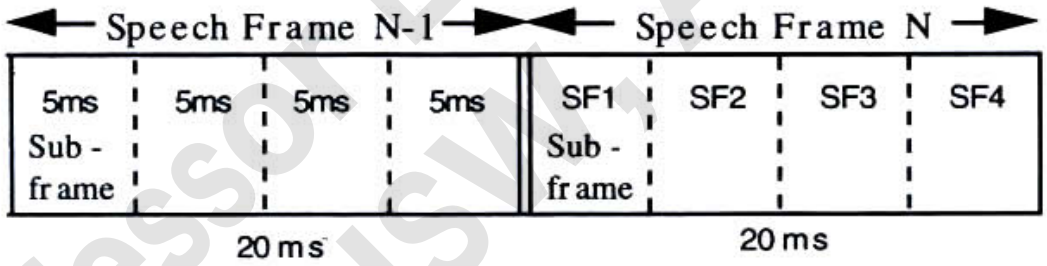
$$SF_k = \delta_k \; F_{N-1} + (1- \delta_k) \; F_N$$

where δ_k is a fraction between 0 and 1.

- The variable δ_k is gradually decreased with the sub frame index.
- Figure given in the next slide shows two frames and corresponding sub frames.

- For specific example, a good choice of values of δ_k is 0.75 ,0.5 ,0.25 and 0 , for k=1,…..,4.
- Using these values, the interpolated LPC coefficients in the four subframes are given by,

$$SF_1 = 0.75\ F_{N-1} + 0.25\ F_N,\ SF_2 = 0.5\ F_{N-1} + 0.5\ F_N$$
$$SF_3 = 0.25\ F_{N-1} + 0.275\ F_N,\ SF_4 = 1.0\ F_N.$$



Non overlapping LPC analysis frames and subframes

Interpolation in this manner, however, increases the overall codec delay.

The Long-Term Predictor

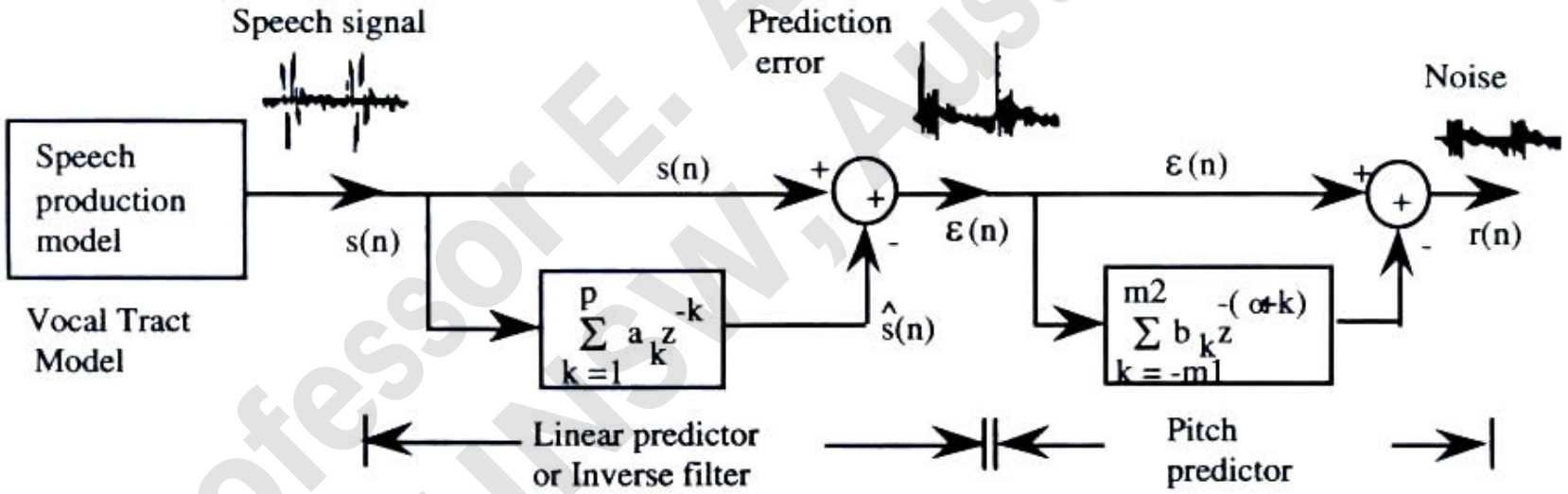
- While the short-term predictor models the spectral envelope of the speech segment being analyzed, the long term predictor, or the pitch, is used to model the fine structure of that envelope.

- After inverse filtering the prediction error, $\varepsilon(n)$, still exhibits some periodicity related to the pitch period of the original speech when it is voiced

- This periodicity is of the order of 20 to 147 samples (55 to 400 HZ) .

The Long-Term Predictor.....

- Adding the pitch predictor to the inverse filter further removes redundancy in the residual signal and turns it into a noise like process (see below).



Linear Predictor and Pitch Predictor

The Long-Term Predictor .....

- It is called a **pitch predictor** since it removes the pitch periodicity or a long-term predictor since the predictor delay is between 20 and 147 samples.
- The output signal r(n) after the pitch predictor is very close to random Gaussian noise.
- The transfer function of the pitch predictor is given by

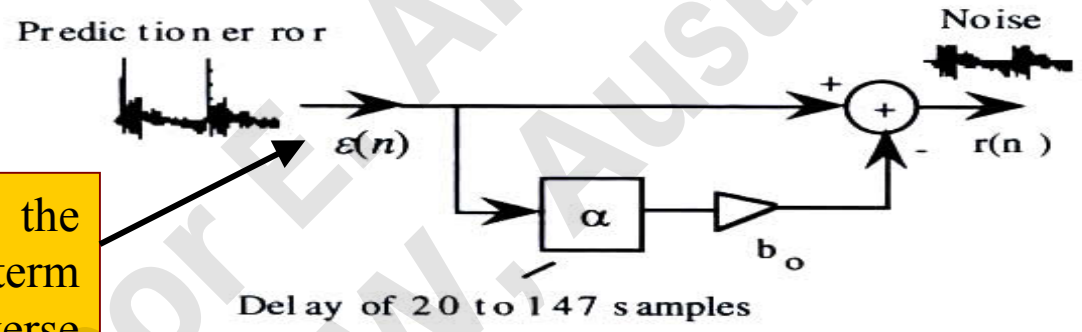
$$P(z) = \frac{R(z)}{\mathcal{E}(z)} = 1 - \sum_{k=-m1}^{m2} b_k z^{-(\alpha+k)}$$

The Long-Term Predictor .....

- For m1=m2=0 , a one-tap predictor is obtained ,for m1=m2=1 , a three tap predictor results .
- The delay α usually represents the pitch period in samples (or a multiple of it and b_k represents the long-term predictor coefficients.
- The pitch predictor is not an essential requirement in medium bit rate LPC coders such as **multi-pulse excited LPC** and **regular pulse excited LPC coders**, however including a pitch predictor improves their performance.

- The pitch predictor is essential in low bit rate speech coders, such as those using the CELP technique.
- A one-tap pitch predictor is shown below and is given by

$$P(z) = 1 - b_o z^{-\alpha}$$



One-Tap Pitch Predictor

Where ε(n) is the residual after short-term prediction or inverse filtering

- For the one-tap predictor,the residual r(n) is given by,

$$r(n) = \varepsilon(n) - b_o\,\varepsilon(n-\alpha)$$

Open-Loop Method

- The parameters α and b_0 are determined by minimizing the mean squared residual error after short-term and long-term prediction over a period of N samples.
- The mean squared residual E is given by

$$E = \sum_{n=0}^{N-1} r(n)^2 = \sum_{n=0}^{N-1} [\varepsilon(n) - b_o\, \varepsilon(n-\alpha)]^2$$

$$\text{Setting } \frac{\partial E}{\partial b_o} = 0 \Rightarrow b_o = \frac{\sum_{n=0}^{N-1} \varepsilon(n).\varepsilon(n-\alpha)}{\sum_{n=0}^{N-1} [\varepsilon(n-\alpha)]^2} \qquad |b_0| \leq 1$$

- Substituting b_0 into mean squared residual

$$E = \sum_{n=0}^{N-1} r(n)^2 - \frac{\left[\sum_{n=0}^{N-1} \varepsilon(n).\varepsilon(n-\alpha)\right]^2}{\sum_{n=0}^{N-1} [\varepsilon(n-\alpha)]^2}$$

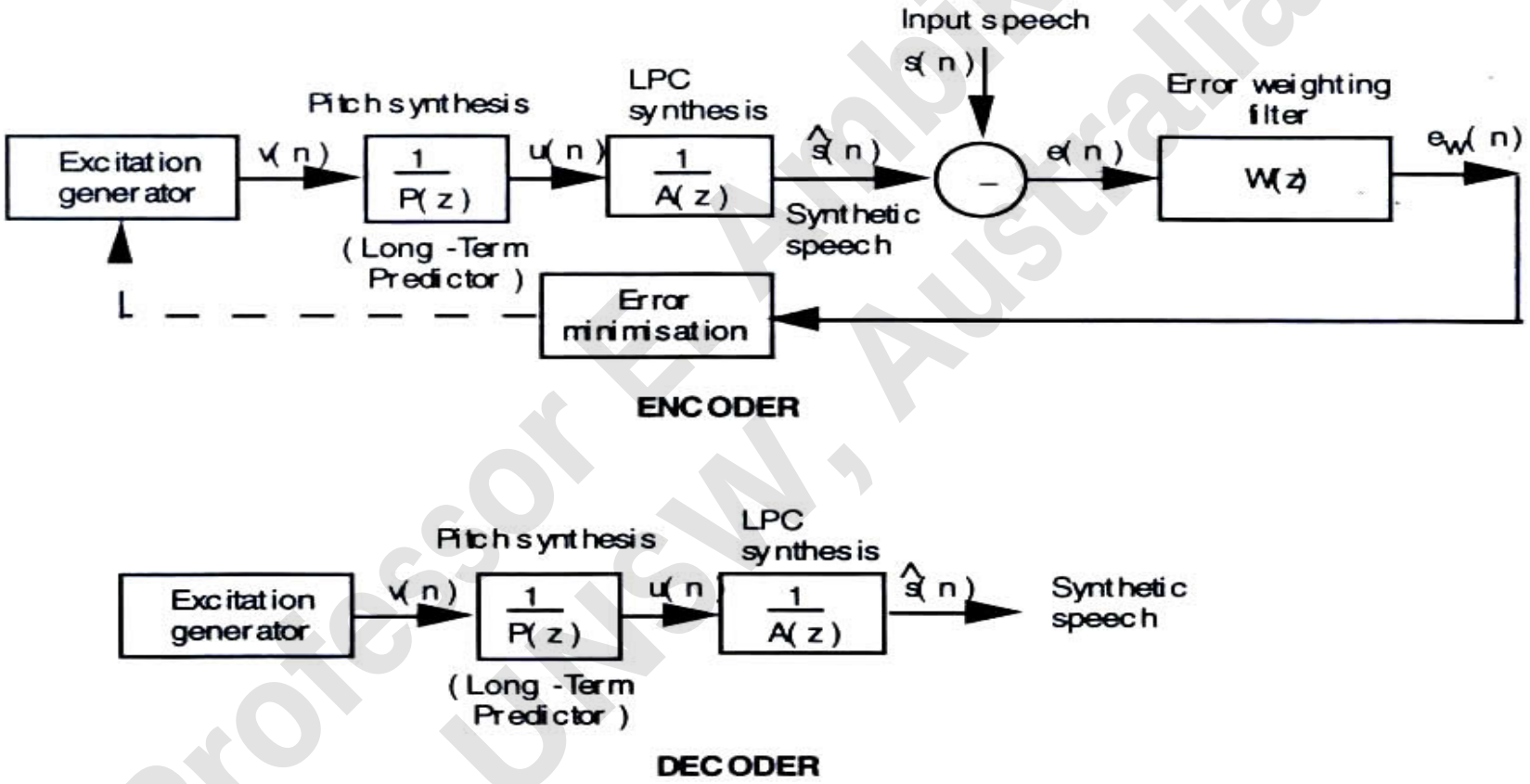
Minimizing E requires maximizing the normalized correlation term in this equation.

↑

Normalised Correlation between the residual $\varepsilon(n)$ and its delayed version.

This term is computed for all possible values α over its specified range, and the value α which maximizes this term (or minimising E) is chosen.

The general model for analysis-by-synthesis linear predictive coding as given previously may be modified in order to include a long-term predictor or pitch predictor.



- It was shown above that the pitch predictor parameters b_0 and α are calculated directly from the LPC residual signal $\varepsilon(n)$.
- This is known as the **open-loop method** for calculating the pitch predictor parameters.
- However, a significant improvement is achieved when pitch predictor parameters b_0 and α are optimized inside the analysis-by-synthesis loop. In this case, the computation of the parameters contributes directly to the weighted error minimization procedure.

Closed-Loop Method

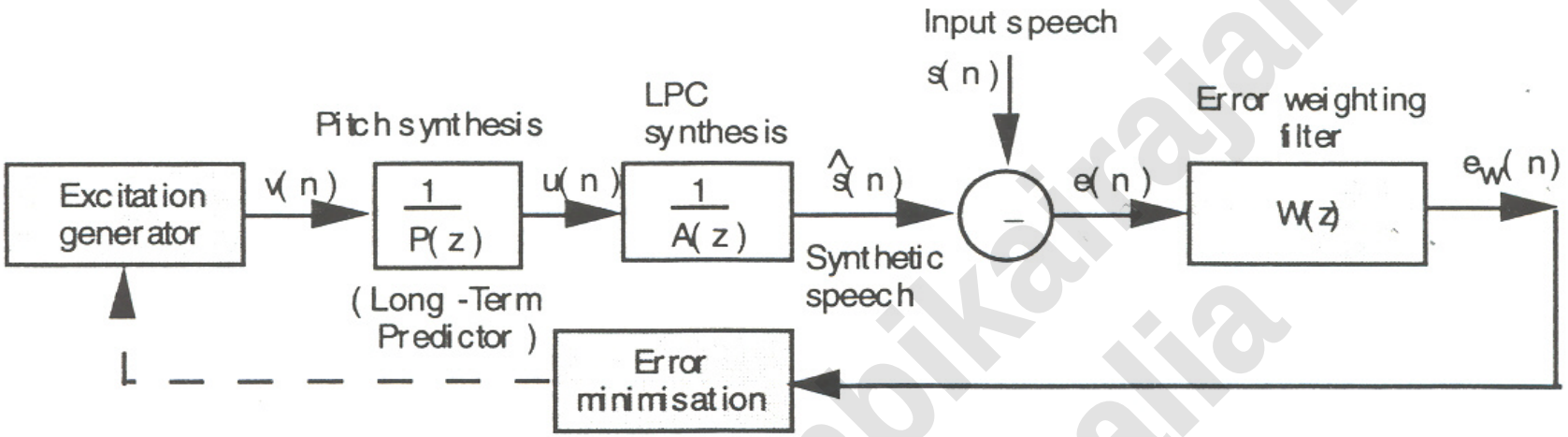
- A different configuration (see next slide) can be derived from the general model for analysis by synthesis LPC coding with Long-term predictor that may be useful for closed-loop analysis.
- The weighting filter W(z) is given by

$$W(z) = \frac{A(z/\gamma_N)}{A(z/\gamma_D)}$$

$$e(n) = s(n) - \hat{s}(n) \Rightarrow E(z) = S(z) - \hat{S}(z) \quad \textit{and}$$

$$E_w(z) = W(z).E(z)$$

$$\therefore \quad E_W(z) = W(z).S(z) - W(z).\hat{S}(z)$$

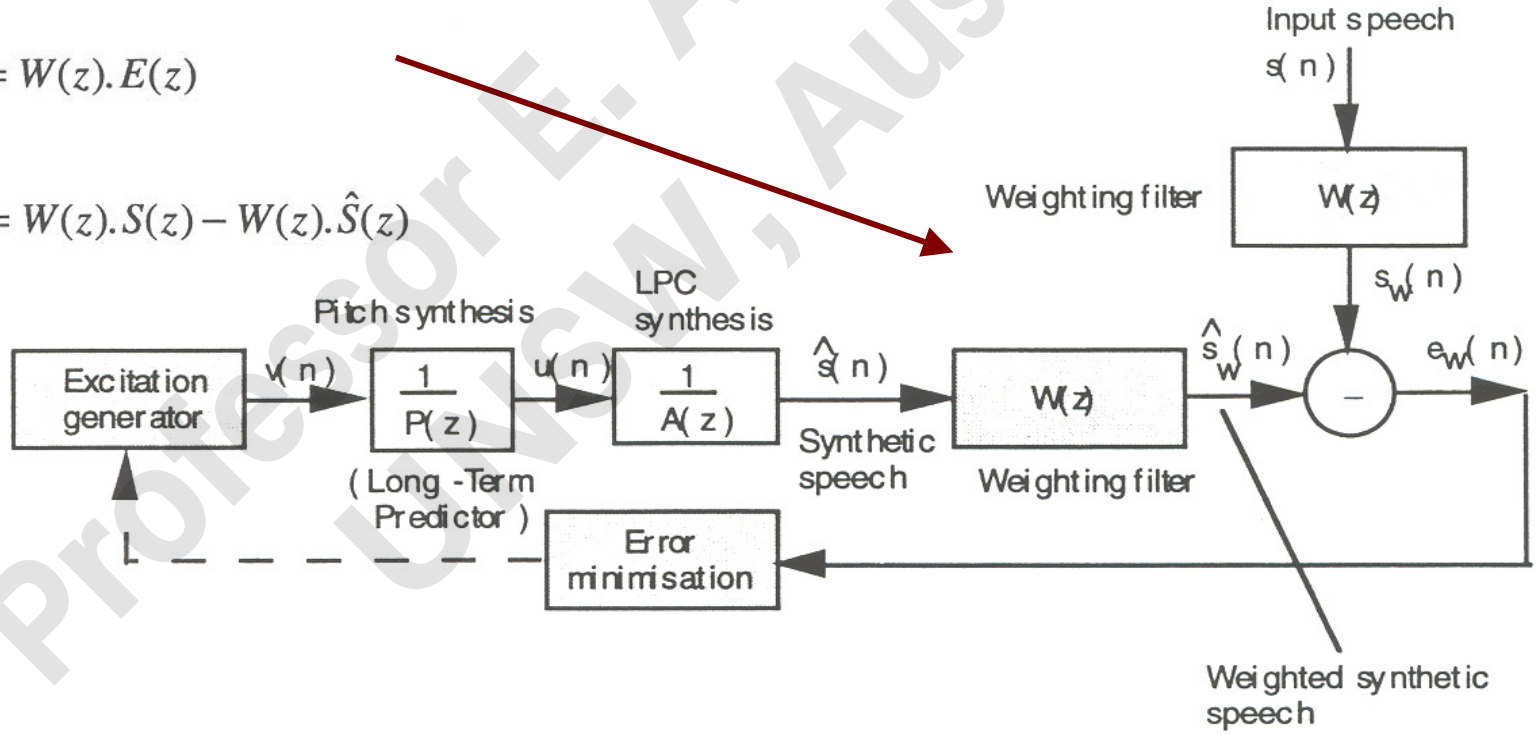


ENCODER

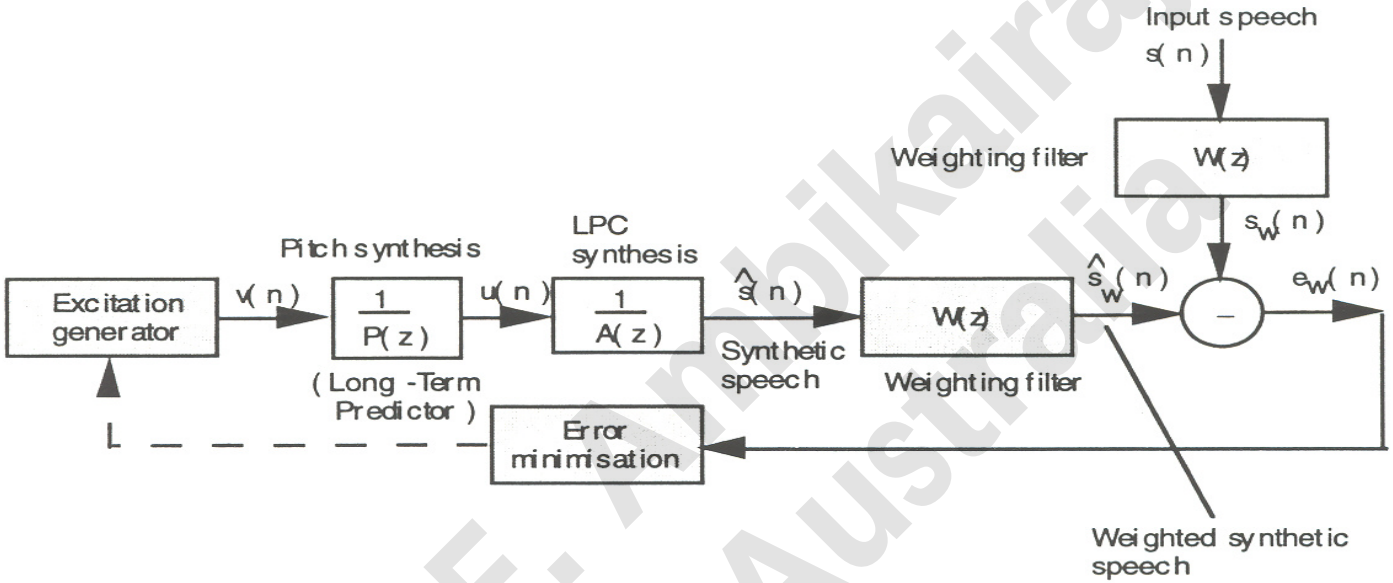
$$e(n) = s(n) - \hat{s}(n) \Rightarrow E(z) = S(z) - \hat{S}(z)$$

$$E_w(z) = W(z).E(z)$$

$$E_W(z) = W(z).S(z) - W(z).\hat{S}(z)$$



- In this new configuration , the original and the synthesized speech are weighted separately before subtraction.



- Assuming one-tap long-term prediction , the output of the pitch synthesis filter given by

$$u(n) = v(n) + b_o\, u(n-\alpha)$$

- It is first assumed that no excitation has been determined, so the above equation reduces to

$$u(n) = 0 + b_o\, u(n-\alpha)$$

- The weighted synthesis speech , $\hat{s}_W(n)$ is given by

$$\hat{s}_W(n) = \sum_{i=0}^{n} u(i)h(n-i) + \hat{s}_O(n)$$

Where h(n) is the impulse response of the combined filter W(z)/A(z) and $\hat{s}_0(n)$ is the zero input response of the combined filter W(z)/A(z), i.e. the output of the combined filter due to its initial states.

- The weighted error between the original and synthesized speech is given by

$$e_w(n) = s_w(n) - \hat{s}_w(n)$$

By combining $\hat{s}_w(n) = \sum_{i=0}^{n} u(i)h(n-i) + \hat{s}_o(n)$ 1

$e_w(n) = s_w(n) - \hat{s}_w(n)$ we obtain

$$e_w(n) = s_w(n) - \hat{s}_o(n) - b_o \sum_{i=0}^{n} u(i-\alpha)h(n-i)$$

$$\Downarrow$$

$$y_\alpha(n) = u(n-\alpha) * h(n)$$

- The mean squared weighted error is given by

$$E_w = \sum_{n=0}^{N-1} [s_w(n) - \hat{s}_o(n) - b_o y_\alpha(n)]^2$$

Setting $\frac{\partial E_w}{\partial b_o} = 0 \quad \Rightarrow \quad b_o = \frac{\sum_{n=0}^{N-1} \{s_w(n) - \hat{s}_o(n)\} y_\alpha(n)}{\sum_{n=0}^{N-1} [y_\alpha(n)]^2}$

- Substituting b_0 we obtain,

$$E_w = \sum_{n=0}^{N-1} [s_w(n) - \hat{s}_o(n)]^2 - \frac{\left[\sum_{n=0}^{N-1} \{s_w(n) - \hat{s}_o(n)\} y_\alpha(n)\right]^2}{\sum_{n=0}^{N-1} [y_\alpha(n)]^2}$$

$$b_o = \frac{\sum_{n=0}^{N-1} \{s_w(n) - \hat{s}_o(n)\} y_\alpha(n)}{\sum_{n=0}^{N-1} [y_\alpha(n)]^2}$$

↑

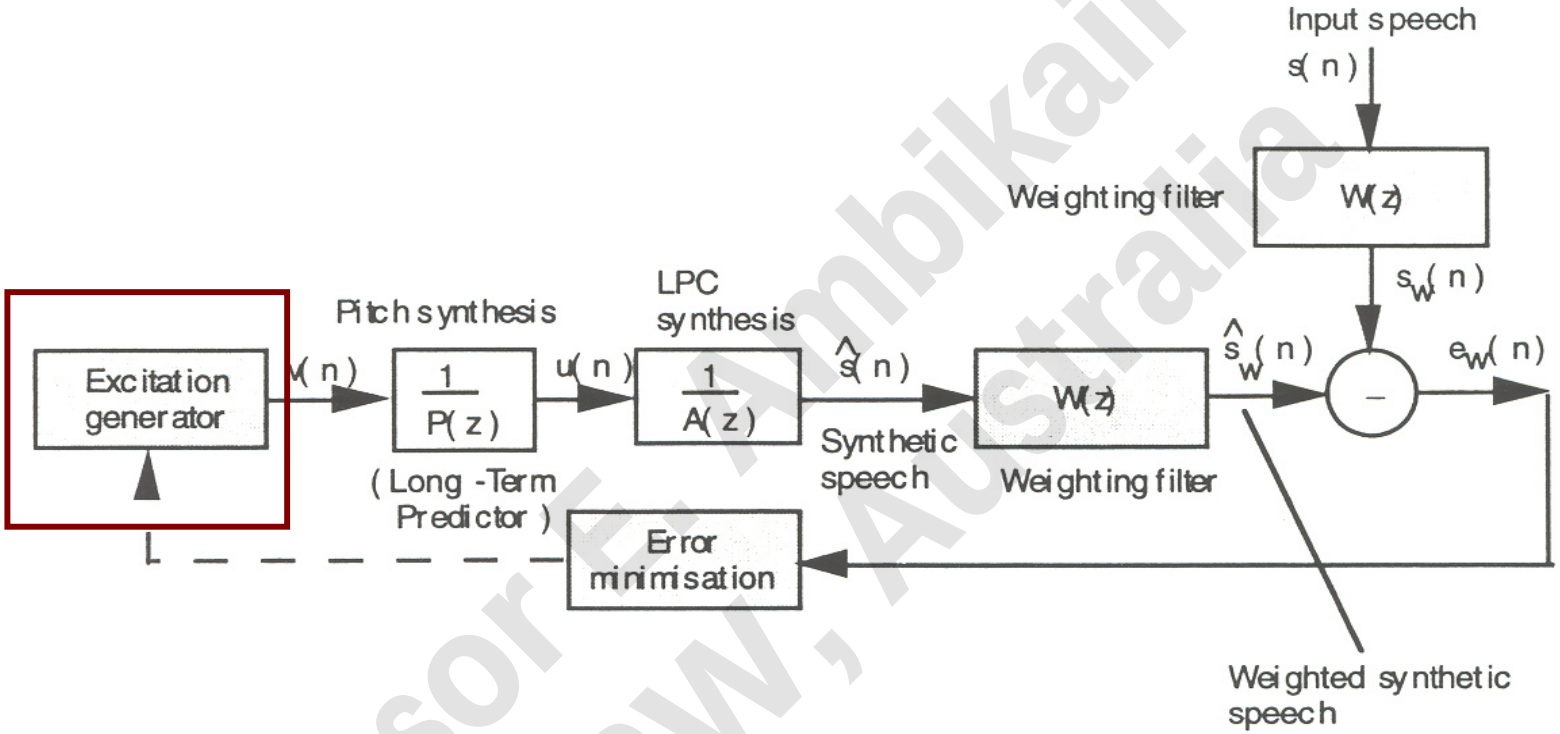
The pitch delay α is selected as the delay which maximises this term.

- Significant speech quality improvement over the open loop solution is achieved when the long-term predictor parameters b_0 and α are computed inside the optimization loop using the above equations.

- The disadvantage of the closed-loop solution is the extra computation needed to compute the convolution equation in the previous slide over the range of α.

- However, a fast procedure to compute this convolution $y_\alpha(n)$ for all the possible delays is available.

- Normally α is calculated using open-loop method and b_0 is calculated using closed-loop method.



A basic Structure for high quality analysis-by-synthesis LPC coding

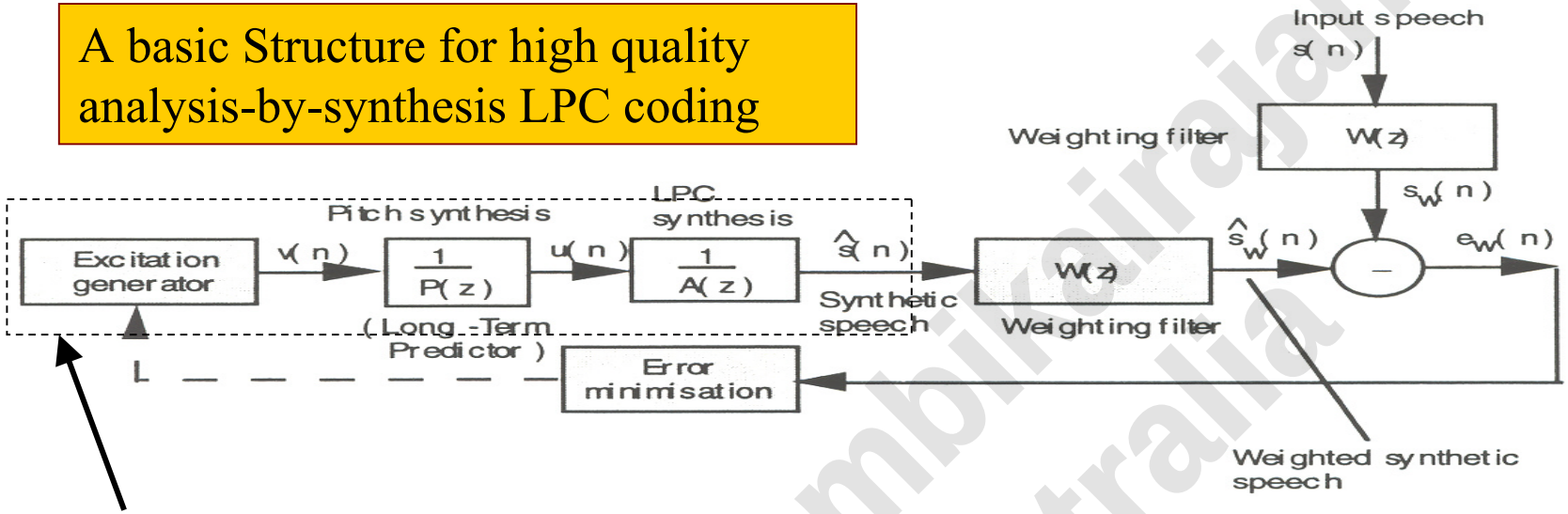
Code-Excited Linear Prediction

- The residual signal after short-term and long-term prediction becomes noise like and its is assumed that this residual can be modelled by a zero-mean Gaussian process with a slowly varying power spectrum.
- As a result the excitation frame may be vector quantized using a large stochastic codebook.
- In the CELP approach , a 5ms (40 samples) excitation frame is modelled by a Gaussian vector chosen from a large Gaussian codebook by minimizing the perceptually weighted error between the original and synthesized speech.

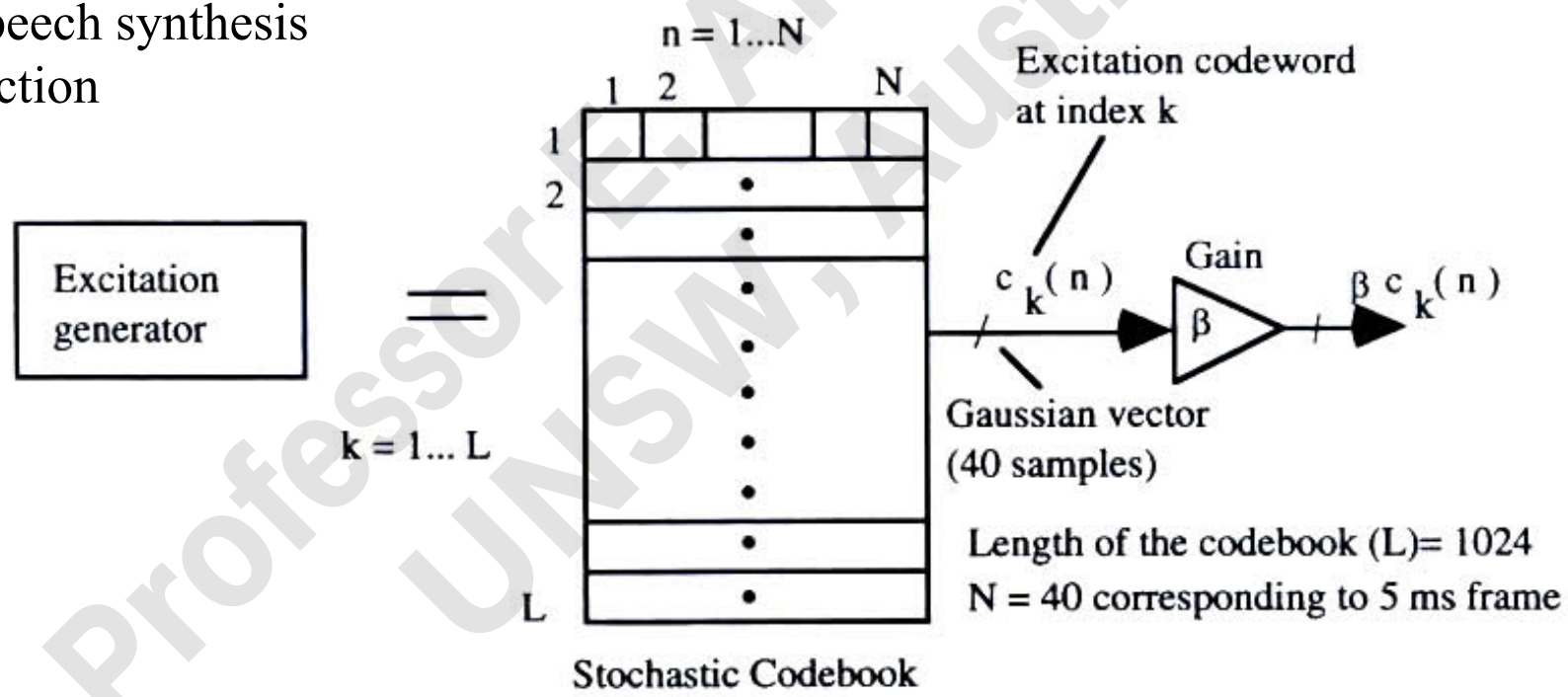
Code-Excited Linear Prediction .....

- Usually a codebook of 1024 entries is needed, and the optimum innovation sequence is chosen by an exhaustive search.
- This exhaustive search of the excitation codebook greatly complicates the CELP implementation.
- The excitation generation in the previous slide(s) can now be replaced by a stochastic codebook with a gain β
- This is show in the next slide.

A basic Structure for high quality analysis-by-synthesis LPC coding



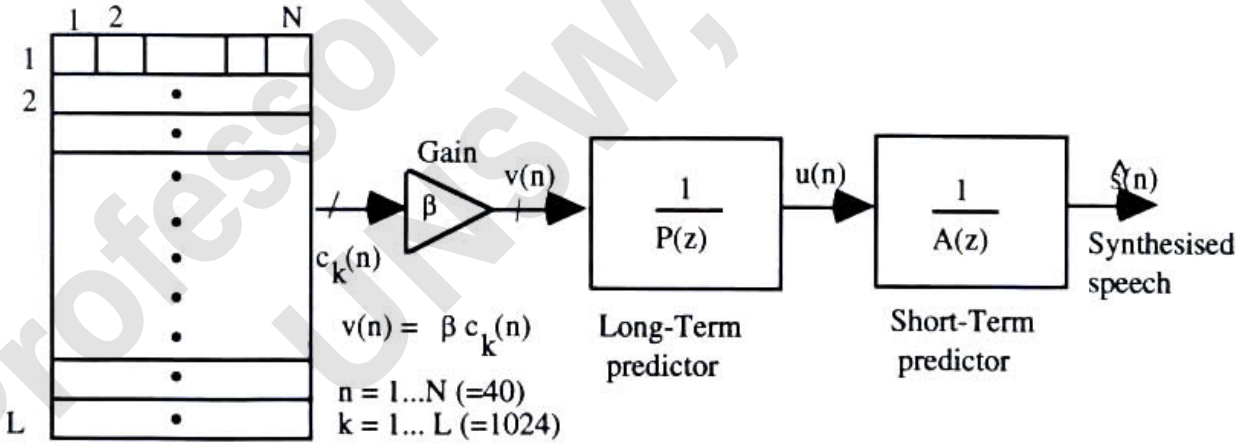
Speech synthesis section



Excitation generation using stochastic codebook

Code-Excited Linear Prediction ..

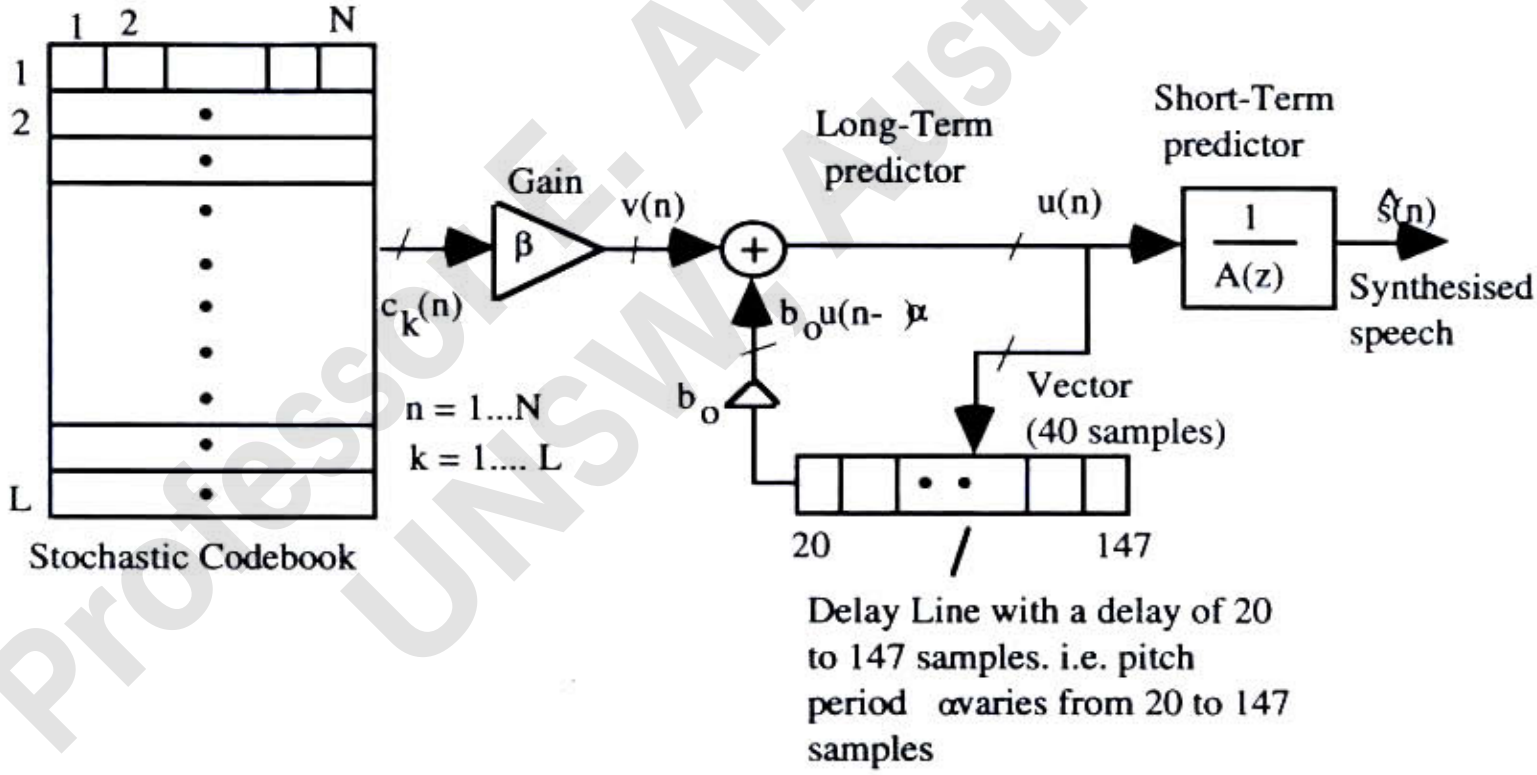
- The stochastic codebook is sometimes referred to as a fixed codebook and the gain β is referred to as a fixed gain.
- The speech synthesis section (see previous slide) is redrawn in the diagram below to include the stochastic codebook.



Speech synthesis section using stochastic codebook

Code-Excited Linear Prediction ..

- Using $P(z)=1-b_0 z^{-\alpha}$, the diagram in the previous slide can be modified as shown below.
- Here α is the pitch period.



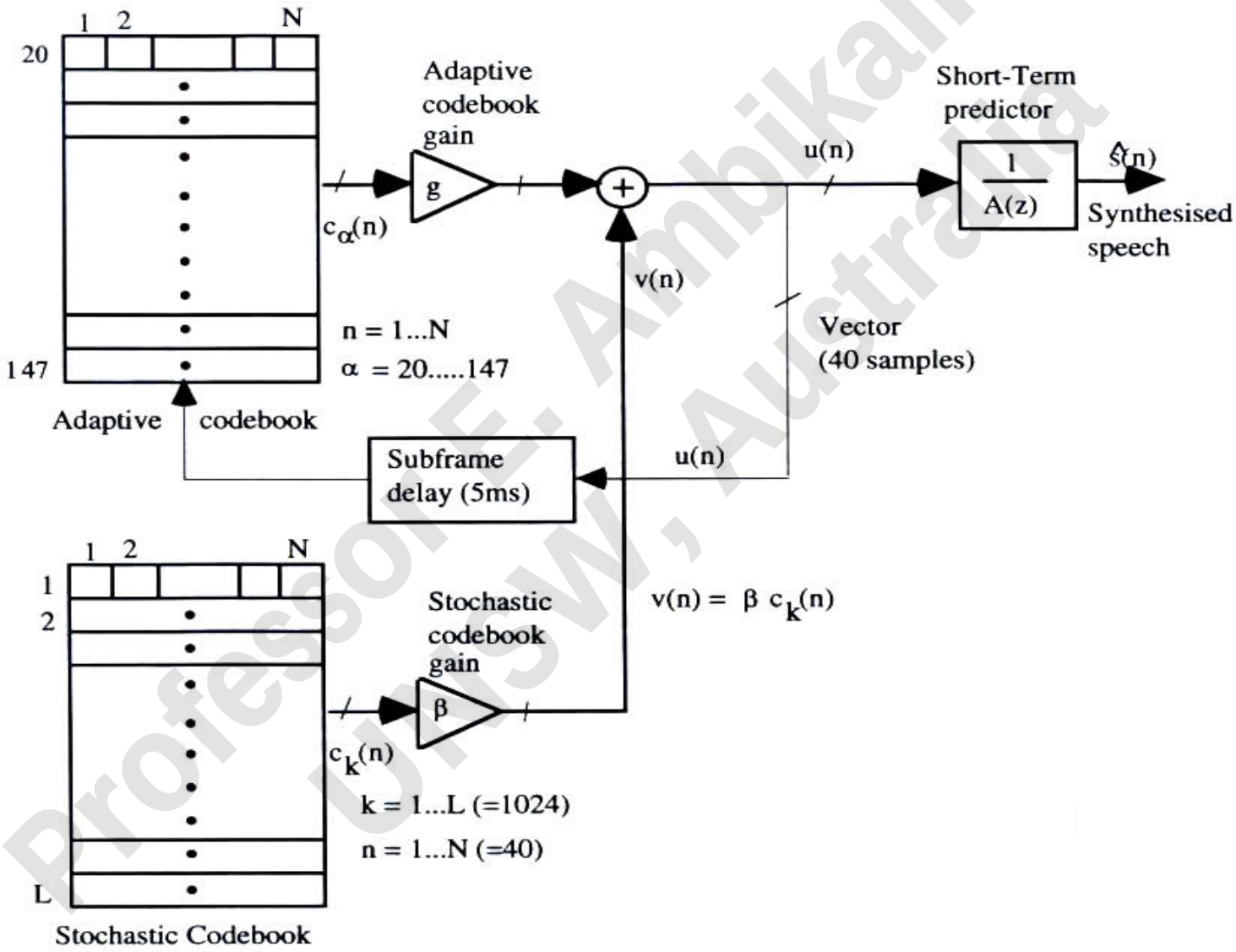
Code-Excited Linear Prediction ..

- It can be seen that each delay cell contains 40 samples and these values are updated every 5ms (sub frame delay).
- The long-term predictor can now be replaced by an adaptive codebook and the appropriate address selected from this codebook will respond to the pitch delay α.
- The gain b_0 is now denoted by the adaptive gain g. Figure.in the previous slide is now redrawn to include the adaptive codebook (see next slide)

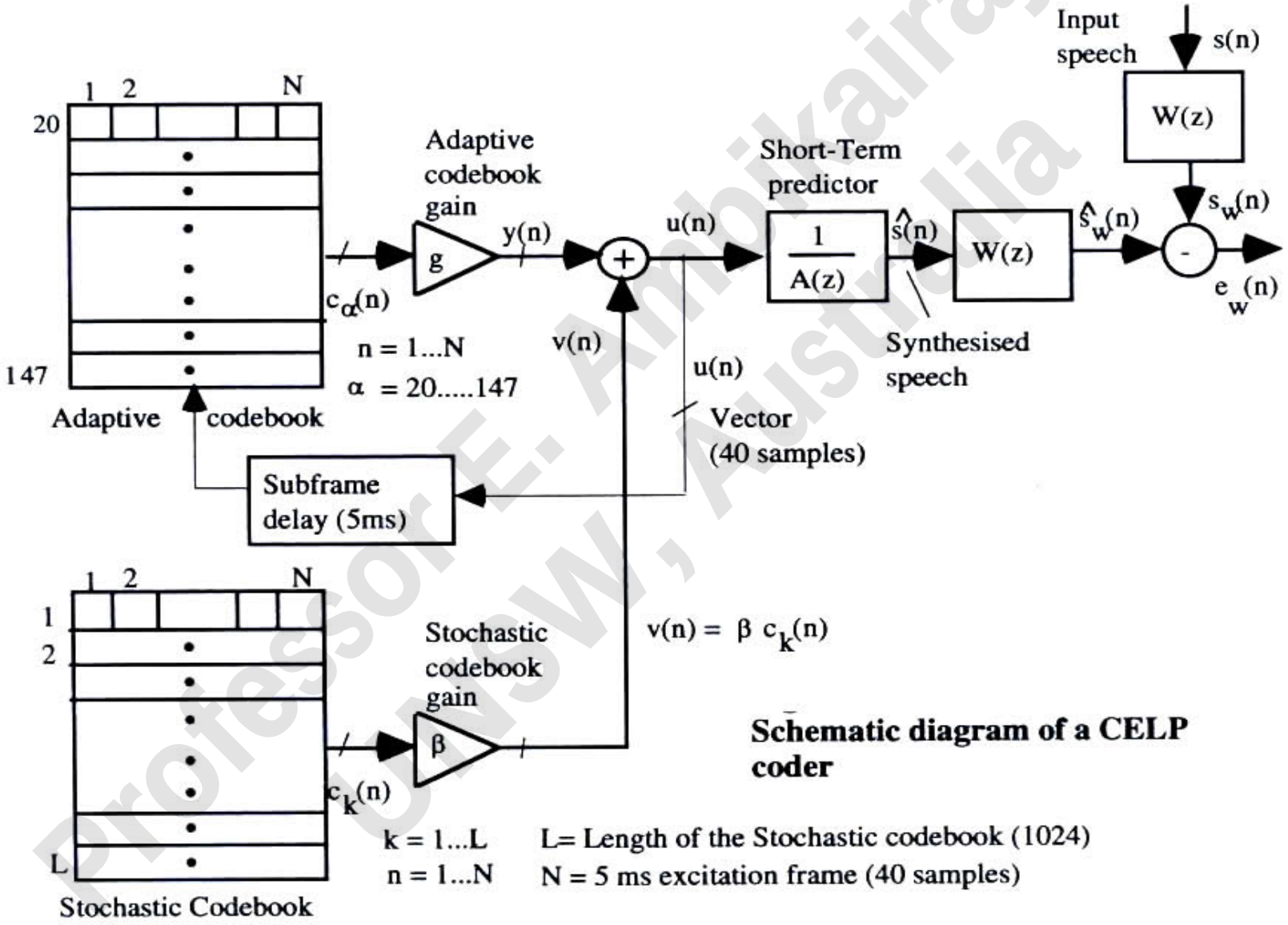
Code-Excited Linear Prediction (continued)

- Figure 11 shows a schematic diagram of the CELP coder including the adaptive codebook.
- The address selected from the adaptive codebook (α), the corresponding gain (g), the address (k) selected from the stochastic codebook and the corresponding gain(β) are sent to the decoder.
- This uses identical codebooks to determine the excitation signal at the input to the Short-term predictor in order to produce the reconstructed speech.

Schematic Diagram of the CELP synthesis model



Schematic Diagram of a CELP Coder



Schematic diagram of a CELP coder

Code-Excited Linear Prediction ...

- The first stage in determining the excitation parameters is to calculate the pitch delay α and the adaptive gain g .
- Then the equations for stochastic codebook gain β and the best excitation codeword (k) are derived.
- Using the Figure in the previous slide, the weighted synthesized speech can be written as

$$\hat{s}_w(n) = \beta\ c_k(n) * h(n) + g\ c_\alpha(n) * h(n) + \hat{s}_o(n)$$

$$\hat{s}_w(n) = \beta\ c_k(n) * h(n) + g\ c_\alpha(n) * h(n) + \hat{s}_o(n)$$

where

β - the Stochastic codebook gain,
$c_k(n)$ - the excitation codeword at index k,
h(n) - the impulse response of the combination filter W(z)/A(z),
g - the adaptive codebook gain,
$c_\alpha(n)$- the codeword selected from the adaptive codebook,
$\hat{s}_o(n)$ - zero input response of the combination filter W(z)/A(z).

- The weighted error between error the original and synthesized speech is given by:

$$e_w(n) = s_w(n) - \hat{s}_w(n)$$

- By Combining the above two equations we obtain:

$$e_w(n) = \left[\{s_w(n) - \hat{s}_o(n)\} - g\{c_\alpha(n) * h(n)\}\right] - \beta\{c_k(n) * h(n)\}$$

- Assume that no excitation has been determined, i.e. $c_k(n)=0$

- Therefore the above equation is rewritten as

$$e_w(n) = \left[\{s_w(n) - \hat{s}_o(n)\} - g\{c_\alpha(n) * h(n)\}\right]$$

- The mean squared weighted error is given by

$$E_w = \sum_{n=0}^{N-1} [s_w(n) - \hat{s}_o(n) - g(c_\alpha(n) * h(n)]^2$$

- Setting $\frac{\partial E_w}{\partial g} = 0$ gives

$$g = \frac{\sum_{n=0}^{N-1} \{s_w(n) - \hat{s}_o(n)\}\{c_\alpha(n) * h(n)\}}{\sum_{n=0}^{N-1} [c_\alpha(n) * h(n)]^2}$$

- Substituting g in equation given in the previous slide results in

$$E_w = \sum_{n=0}^{N-1} [s_w(n) - \hat{s}_o(n)]^2 - \frac{\left[\sum_{n=0}^{N-1} \{s_w(n) - \hat{s}_o(n)\}\{c_\alpha(n) * h(n)\}\right]^2}{\sum_{n=0}^{N-1} [c_\alpha(n) * h(n)]^2}$$

↑

This term is computed for all possible values of α over its specified range and the value of α which maximises this term.

- The stochastic codebook parameters β and k are derived in the following sequence:

$$e_w(n) = \left[\left\{s_w(n) - \hat{s}_o(n)\right\} - g\left\{c_\alpha(n) * h(n)\right\}\right] - \beta\left\{c_k(n) * h(n)\right\}$$

$$\Big\| \quad p(n)$$

$$e_w(n) = p(n) - \beta\left\{c_k(n) * h(n)\right\}$$

- The mean squared weighted error is given by

$$E_w = \sum_{n=0}^{N-1} [p(n) - \beta(c_k(n) * h(n)]^2$$

Setting $\frac{\partial E_w}{\partial \beta} = 0$ gives

$$\beta = \frac{\sum_{n=0}^{N-1} \left\{ [s_w(n) - \hat{s}_o(n)] - g[c_\alpha(n) * h(n)] \right\} \{ c_k(n) * h(n) \}}{\sum_{n=0}^{N-1} [c_k(n) * h(n)]^2}$$

E_w now becomes

$$E_w = \sum_{n=0}^{N-1} [p(n)]^2 - \frac{\left[\sum_{n=0}^{N-1} \left\{ [s_w(n) - \hat{s}_o(n)] - g[c_\alpha(n) * h(n)] \right\} \{ c_k(n) * h(n) \} \right]^2}{\sum_{n=0}^{N-1} [c_\alpha(n) * h(n)]^2}$$

↑

The codeword with index k which maximises this term is chosen, and the scalar gain β is then computed from equation (31).

- Letting $y_\alpha(n) = c_\alpha(n) * h(n) = \sum_{i=0}^{n} h(n) c_\alpha(n-i) \qquad n = 0,1,\ldots N-1$ leads to the following matrix form:

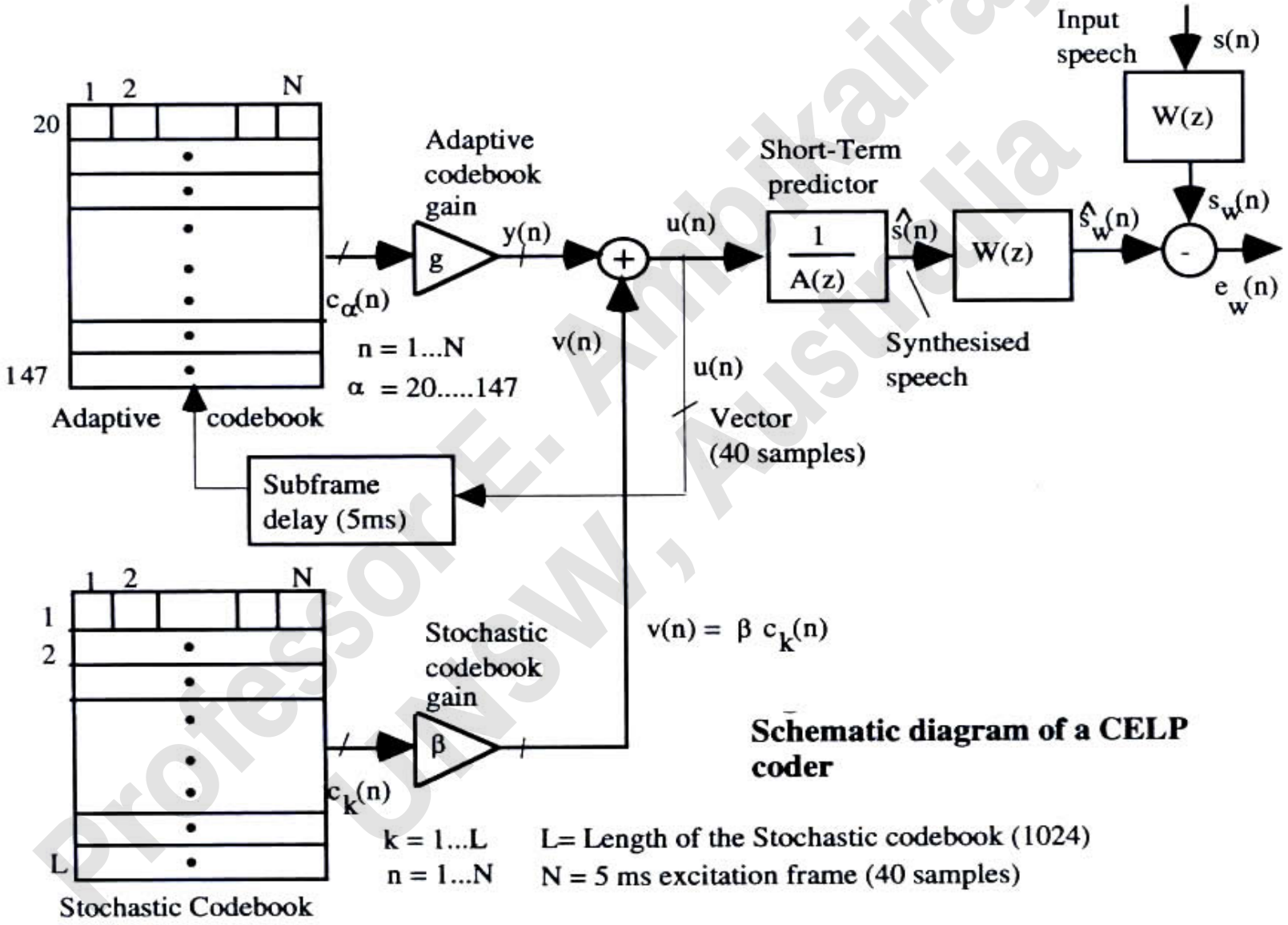
$$\begin{bmatrix} y_\alpha(o) \\ y_\alpha(1) \\ y_\alpha(2) \\ . \\ . \\ . \\ y_\alpha(N-1) \end{bmatrix} = \begin{bmatrix} h(0) & 0 & 0 & 0 & . & . & 0 \\ h(1) & h(0) & 0 & 0 & . & . & 0 \\ h(2) & h(1) & h(0) & 0 & . & . & 0 \\ . & . & . & . & . & . & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ h(N-1) & h(N-2) & h(N-3) & . & . & . & h(0) \end{bmatrix} \begin{bmatrix} c_\alpha(0) \\ c_\alpha(1) \\ c_\alpha(2) \\ . \\ . \\ . \\ c_\alpha(N-1) \end{bmatrix}$$

↑

A convolution matrix of the impulse response h(n)

- Similarly $y_k(n) = c_k(n) * h(n) = \sum_{i=0}^{n} h(n) c_k(n-i) \qquad n = 0,1,\ldots N-1$

Schematic Diagram of a CELP Coder



Schematic diagram of a CELP coder

Code-Excited Linear Prediction ..

- The excitation codebook contains (see figure in the previous slide) L code words (stochastic vectors) of length N samples (typically L=1024 and N=40 samples corresponding to a 5 ms excitation frame).
- The excitation signal for a speech frame of length N is chosen by an exhaustive search of the codebook .
- This is a computationally demanding procedure which is difficult to implement in real time.

- There are alternative methods to simplify the codebook search procedure without affecting the quality of the output speech available such as:
  - CELP search procedure using autocorrelation approach
  - Structured codebooks
  - Sparse exciting codebooks
  - Ternary codebooks
  - Algebraic codebooks
  - Overlapping codebooks

Code-Excited Linear Prediction

- It can be seen from the diagram that CELP coders do not distinguished between voiced and unvoiced frames.
- Input speech is divided into non-overlapping frames typically 20ms (160 samples) in duration and ten LPC coefficients or Line Spectrum Paris (LSP) are calculated for a single frame.
- A speech frame is divided into sub frames (normally 4 sub frames = 4x40 samples) and the parameters g, α, β and k given by the previous equations are calculated for each sub frame.

Code-Excited Linear Prediction ...

- These parameters along with LSP's quantized before transmitting them to the decoder.
- At the decoder 40 samples are synthesized at any one time.
- The fundamental frequency in speech can change more quickly than its spectral content and hence the parameters g ,β,α and k are calculated more frequently than LPC analysis.

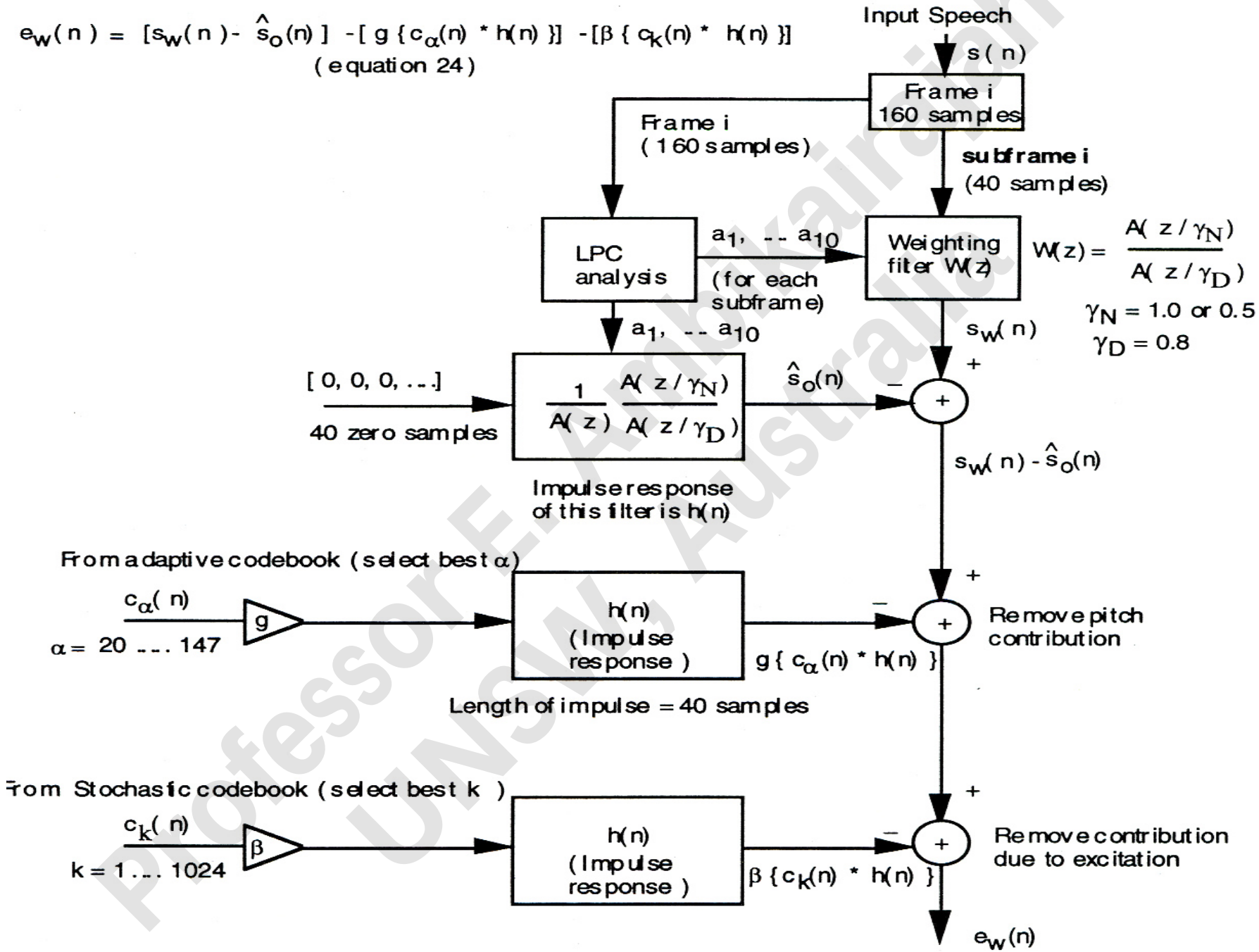
The bit allocation for a 6.8 kbits/s CELP coding is shown in the following table.

Parameter	Number of bits
LSP's	36 bits (3, 3, 4, 4, 4, 4, 4, 4, 3, 3)/20 ms
Stochastic codebook indices (k)	10 bits/5 ms
Stochastic codebook gain (β)	5 bits/5 ms
Adaptive codebook indices (α)	7 bits/5 ms
Adaptive codebook gain (g)	3 bits/5 ms
Total	{36 + (4*25)} bits/20 ms = 6800 bits/s

Implementation of a CELP coder

- Table below shows the default analysis conditions that can be used in a CELP coder simulation.
- The implementation of the CELP coder is based the calculations of the parameters and a block diagram of the excitation sequence is shown in the next slide. in This excitation sequence is repeated for every sub frame.

Sampling frequency	8000 Hz
LPC analysis frame (non-overlapping)	160 samples (20 ms)
Predictor order	10
Analysis Method	Autocorrelation
No. of subframes	4 (4 x 5 ms = 20 ms)
Excitation frame	40 samples (5 ms)
Long-term predictor taps	1
Long-term Predictor update	40 samples (5 ms)
Long-term predictor analysis	Adaptive codebook, integer delays (20 to 147)
Stochastic codebook length	1024



Block diagram of the excitation sequence

Speech Quality Measurement

- The original speech and the decoded speech may be compared in two ways:

(a) By calculating the overall signal to noise ration as follows:

$$SNR(dB) = 10 \log \frac{\sum_{n=1}^{M} [s(n)]^2}{\sum_{n=1}^{M} [s(n) - \hat{s}(n)]^2}$$

s(n) - original speech, $\hat{s}(n)$ - synthesised speech

M- total number of samples (500 frames x 160 samples)

(b) by calculating the segmental signal to noise ratio as follows:

$$SEGSNR(dB) = \frac{1}{K}\sum_{k=1}^{K}\left(SNR_k\right)$$

K-number of frames ;SNR_k is the S/N of the k^{th} frame calculated using equation given in the previous slide