ELEC9344:Speech & Audio Processing

Chapter 13 (Week 13)

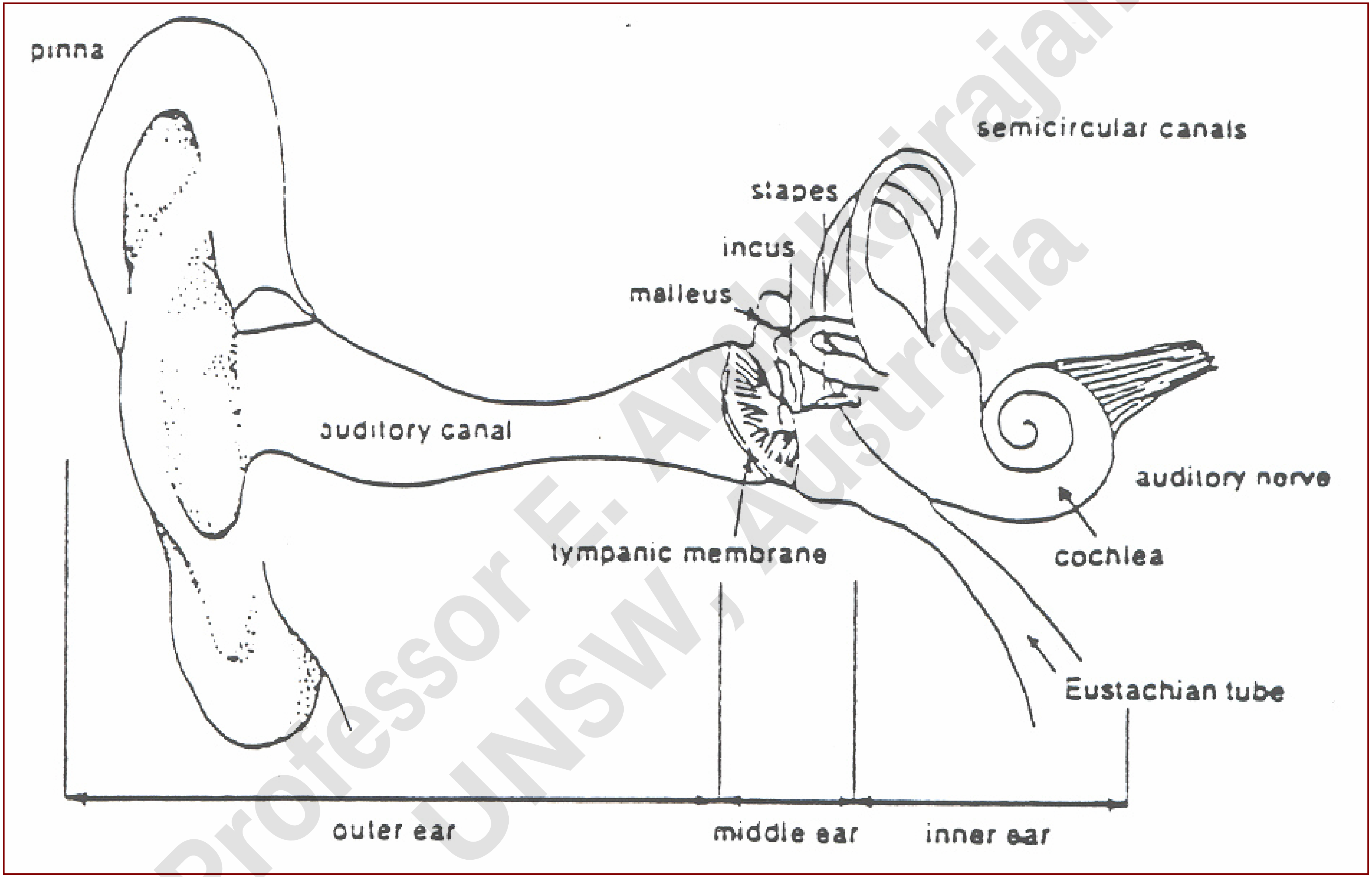
Auditory Masking

Anatomy of the ear

- The ear divided into three sections:
  - The outer
  - Middle
  - Inner ear (see next slide)

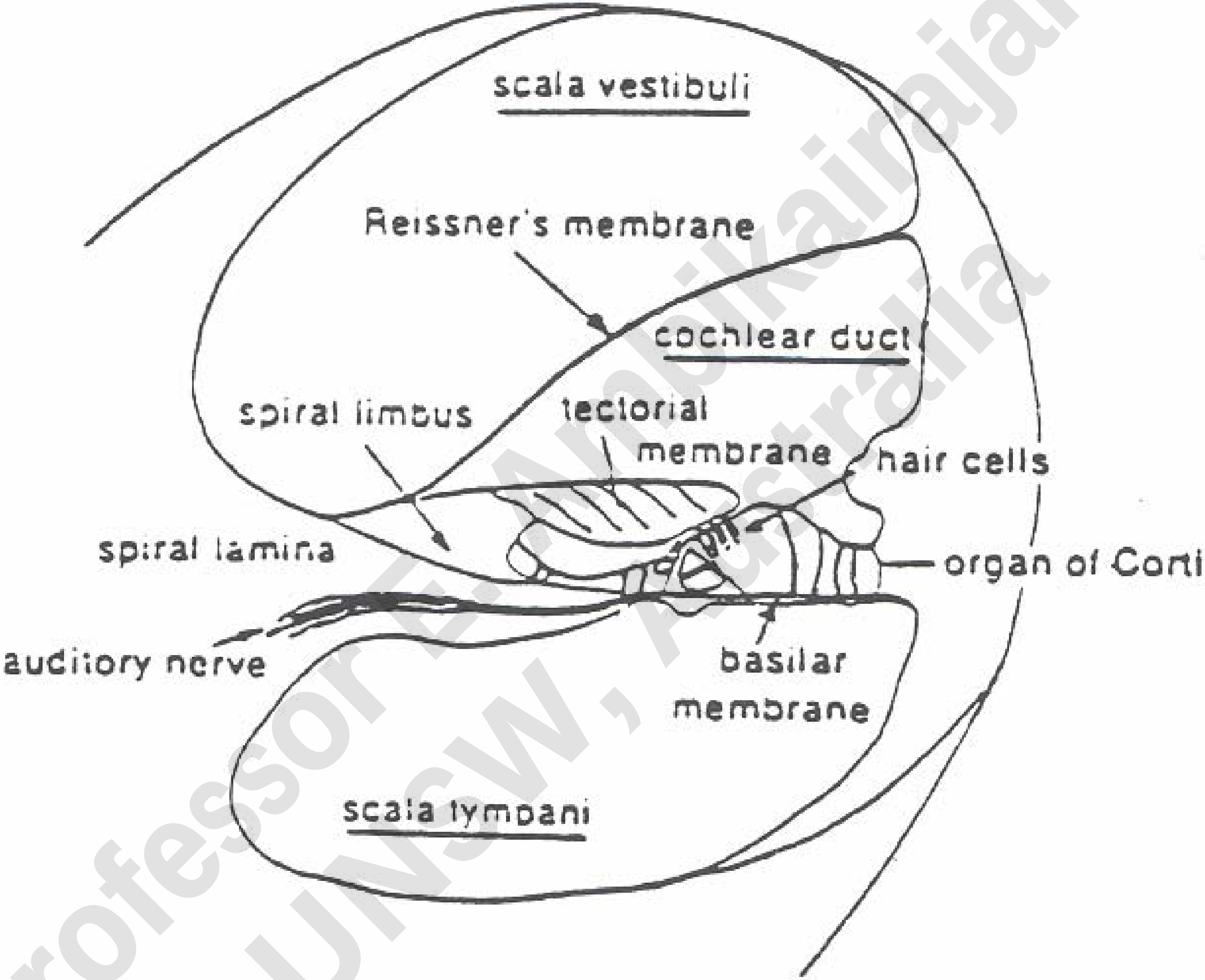
- The outer ear is terminated by the eardrum (tympanic membrane).

- Sound waves entering the auditory canal of the outer ear are directed into the ear drum and cause it vibrate



Schematic diagram of the parts of the ear

- The vibrations are transmitted by the middle ear, an air filled section comprising a system of three tiny bones, the malleus ,incus and stapes , to the cochlea (the inner ear).
- The cochlea is a spiral if about 2 ¾ turns which unrolled would be about 3.5cm long.
- The cochlea consists of three fluid-filled sections (see fig below).
- One, the cochlear duct , is relatively small in cross-sectional area, and other two, the scala vestibuli and the scala tympani are larger and roughly equal in area.



Cross section of the cochlea

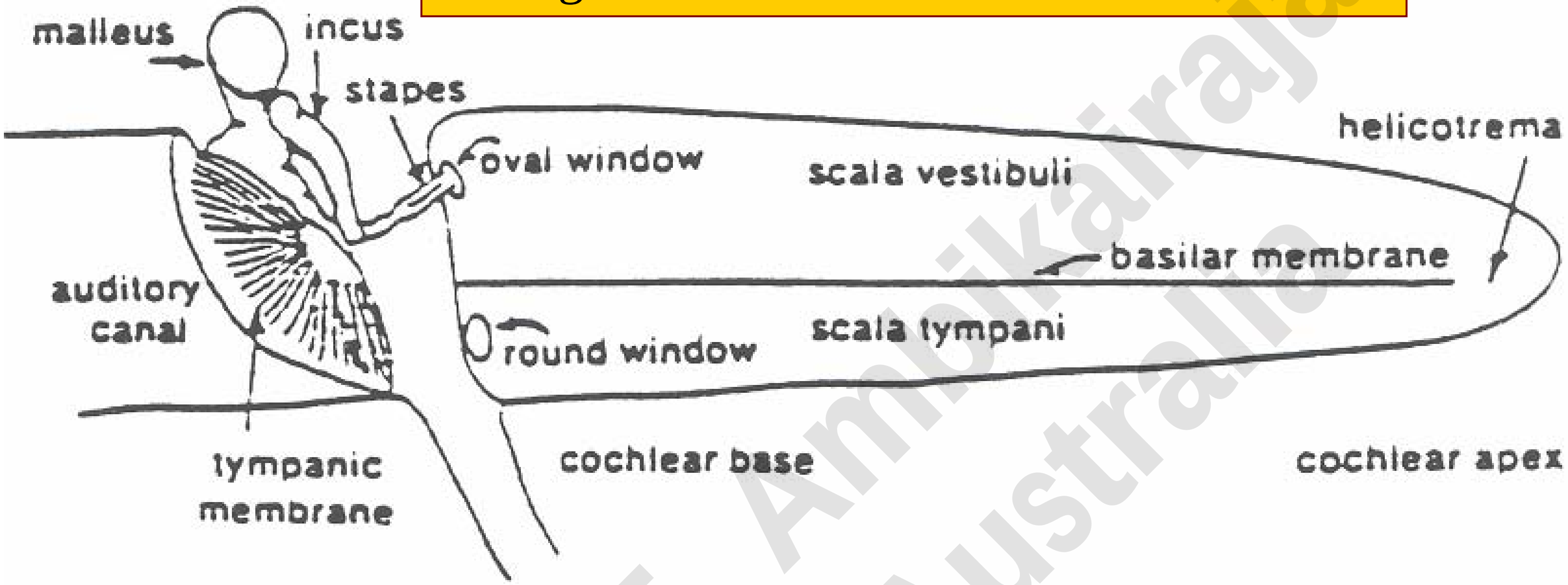
- The scala vestibuli is connected to the stapes via the oval window (see next slide) .

- The scala tympani terminates in the round window which is a thin membranous cover allowing the free movement of the cochlear fluid.

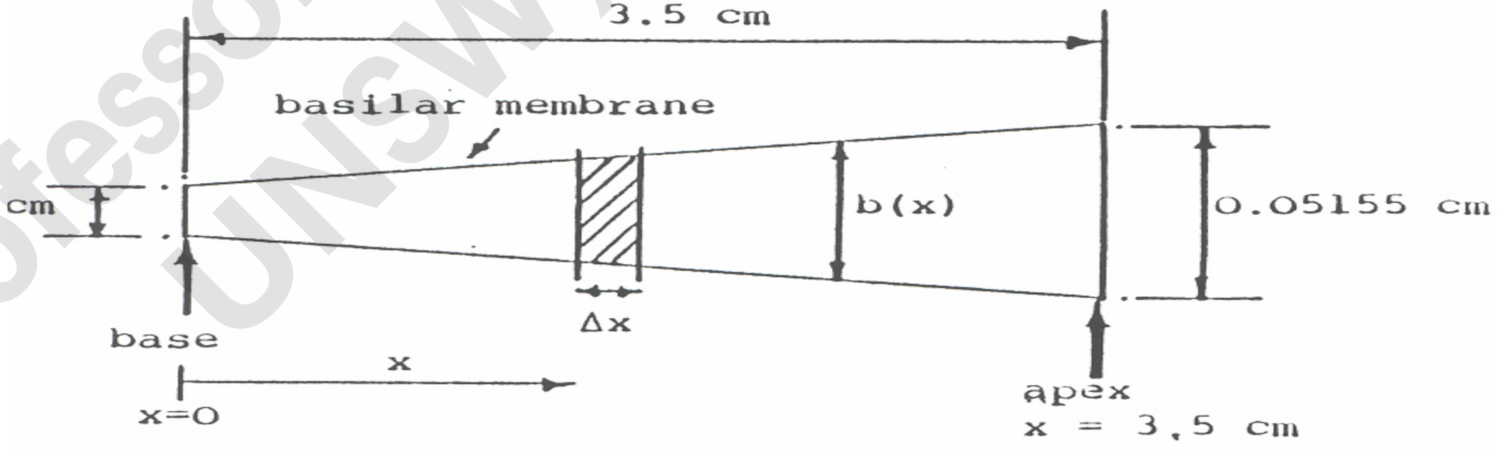
- Running the full length of the cochlea is the Basilar Membrane (BM) which separates the cochlear duct from the scala vestibuli.

- The Reissner membrane is very thin compared to the basilar membrane.

A longitudinal section of an uncoiled cochlea



width, $b(x) = 0.019 + 0.0093\,x$ cm

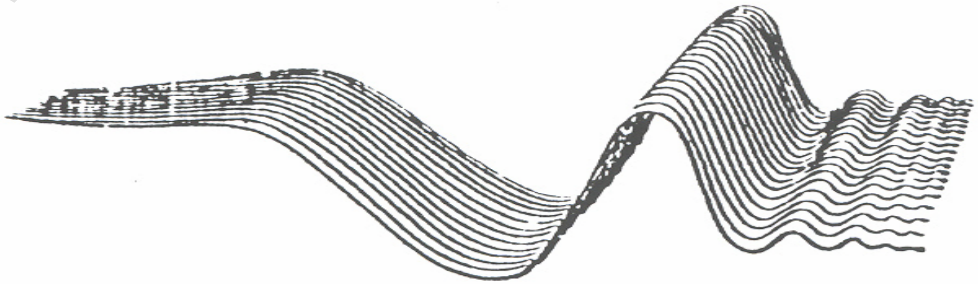


unrolled basilar membrane

- It has been shown by Bekesey (1960) that when the vibrations of the eardrum are transmitted by the middle ear into movement of the stapes, the resulting pressure within the cochlea fluid generates a traveling wave of displacement on the basilar membrane.

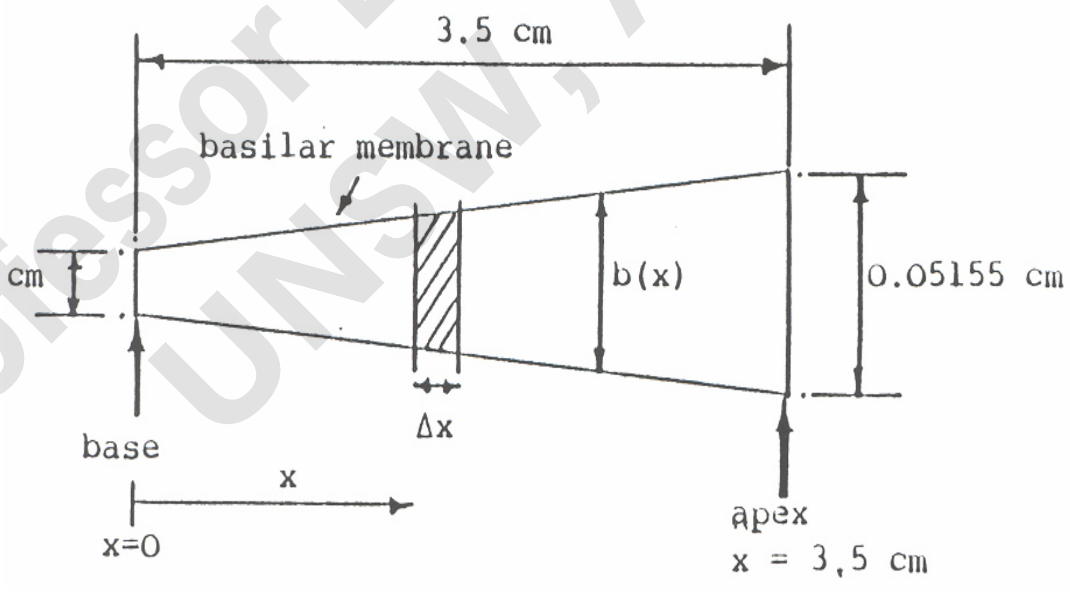
- The location of the maximum amplitude of this traveling wave varies with frequency of the eardrum vibrations

- The response of the BM at an instant of time to a pure tone at the stapes is schematically shown below

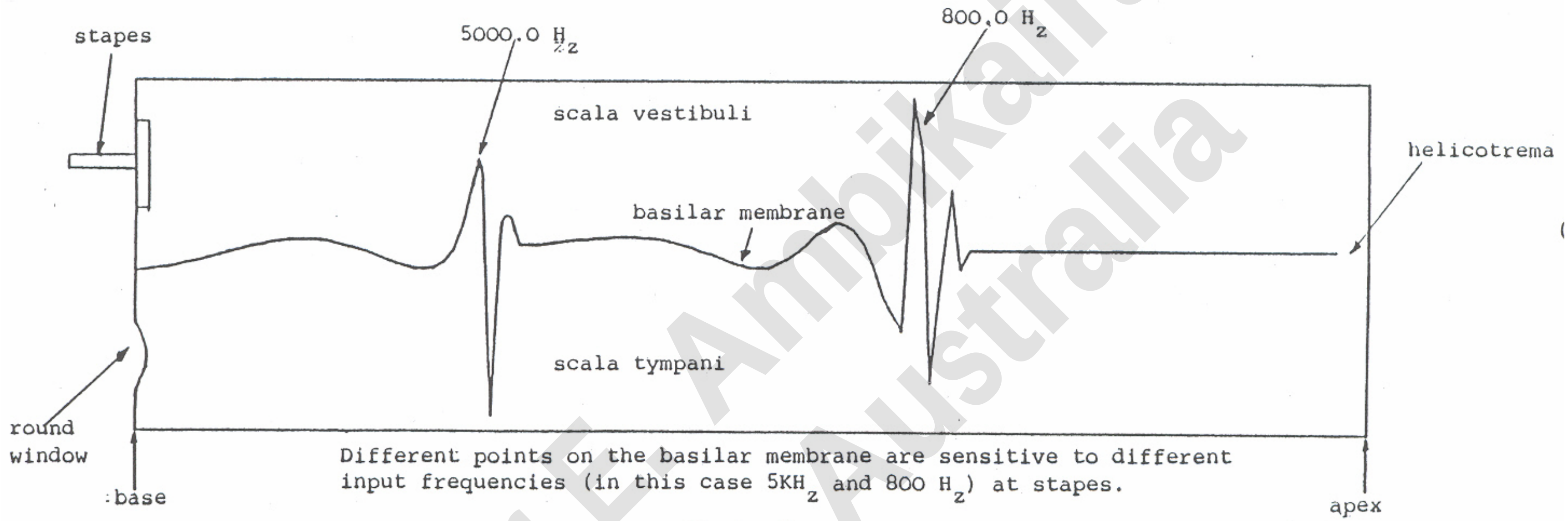


Schematic view of an instant travelling wave along the basilar membrane

- The basilar membrane varies in width and stiffness along its length
- At the basal end it is narrow and stiff whereas towards the apex it is wider and more flexible.
- The maximum membrane displacement will occur at the stapes end for high frequencins and at the far end (apex) for low frequencies.



- The wave motion along the BM is governed by the mechanical properties of the membrane and hydrodynamic properties of the surrounding fluid (scalas)

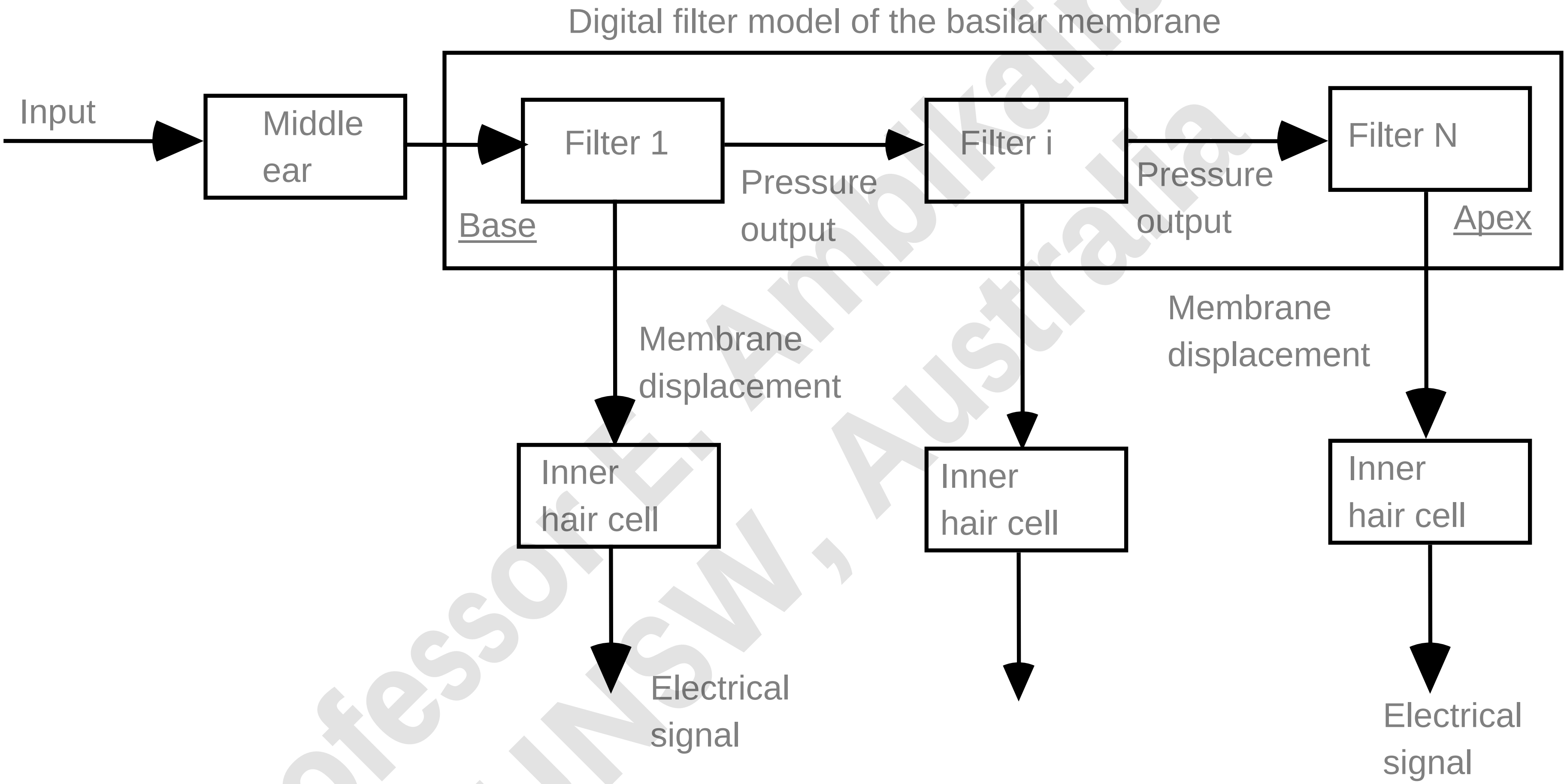


Different points on the basilar membrane are sensitive to different input frequencies (in this case 5KHz and 800 Hz) at stapes.

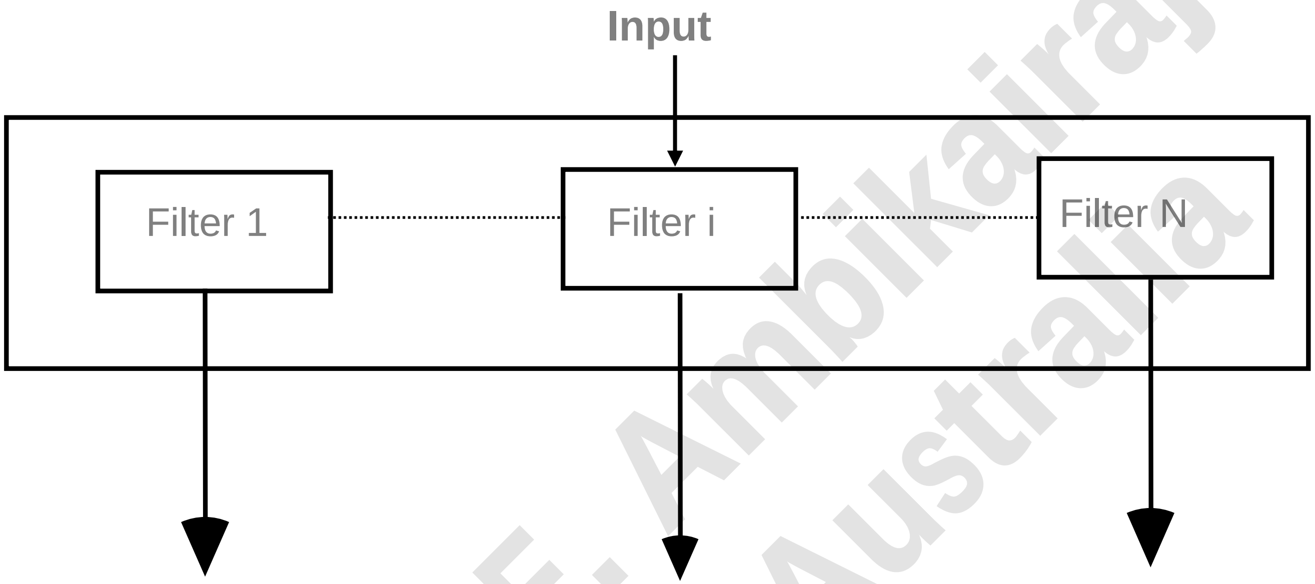
- It appears that each point of the BM moves independently (i.e. a point on the basilar membrane is assumed to have no direct mechanical coupling to neighboring points).

- However, the neighboring points are coupled through the surrounding fluid.

Transmission Line Model



Parallel Filter Bank Model



Sound Pressure Level

- Atmospheric pressure is approximately 15 lb/in^2 or 1 bar. A variation of one millionth of the atmospheric pressure (or 1 µbar) is an appropriate stimulus for hearing. Such a pressure variation is generated in normal conversation by the human voice.
- The minimum level of pressure changes to which man is sensitive is well over 0.0002 µbars.
- A figure commonly used as the upper limit of hearing is 2000 µbars.

- At this upper limit, acoustic stimulus is accompanied by pain. We know,

$$\text{dB (power)} = 10 \log[P_0/P_i]$$

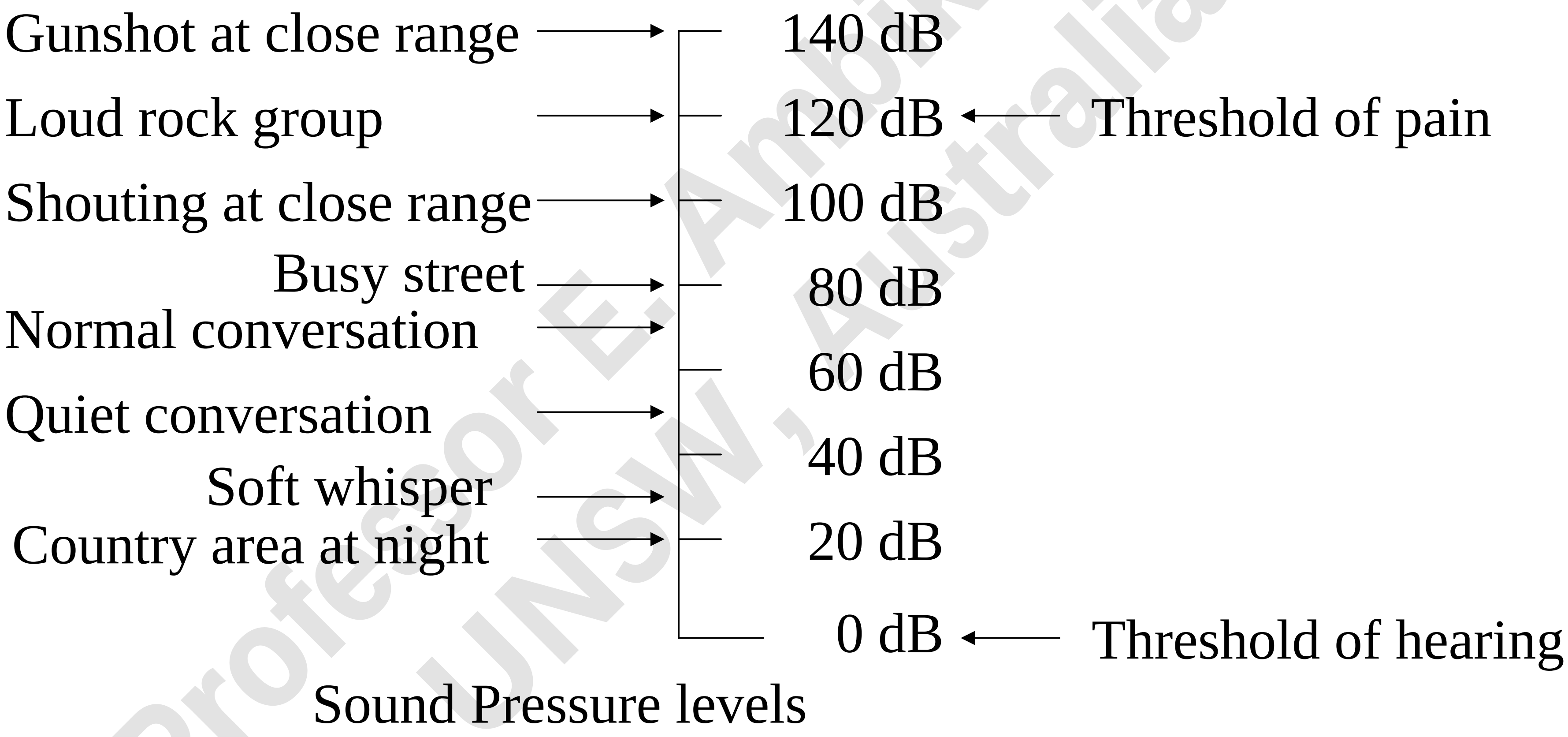
- Since acoustic power is directly related to the square of acoustic pressure,

$$\text{dB (pressure)} = 10 \log[(P_0)^2/(P_i)^2 = 20 \log[P_0/P_i]$$

- P_i is commonly taken as 0.0002 μbars (at or below the threshold for hearing).
- Given an upper limit of p_0 as 2000 μbars, the Sound Pressure Level (SPL) of an acoustic stimulus is:

$$\text{SPL} = 20 \log(2000\ \mu\text{bars}/0.0002\ \mu\text{bars}) = 20 \log(10^7)$$
$$= 140\ \text{dB.}$$

Figure below shows typical sound levels in dB SPL for various common sounds.



Gunshot at close range

Loud rock group

Shouting at close range

Busy street

Normal conversation

Quiet conversation

Soft whisper

Country area at night

140 dB

120 dB

100 dB

80 dB

60 dB

40 dB

20 dB

0 dB

Threshold of pain

Threshold of hearing

Sound Pressure levels

Auditory Masking

- The human auditory system is often modelled as a filter bank which is based on a particular perceptual frequency scale.
- These filters are called ‘critical-band’ filters
- From the point of view of perception, critical bands can be treated as single entities within the spectrum.
- Signal components within a given critical band can be masked by other components within the same critical band.
- This is called **intra-band** masking.

- In addition, sounds on one critical band can mask sounds in different critical bands.
- This is called **inter-band** masking.
- While the masking process is very complex and only partially understood, the basic concepts can be successfully used in audio compression systems, so that better compression is achieved.
- Many people have examined the human auditory system and have concluded that the ear is primarily a frequency analysis device and can be approximated by a bandpass filter bank, consisting of strongly overlapping bandpass filters (known as the **critical-band filters**).
- Twenty five critical bands are required to cover frequencies of up to 20 kHz

- These filters may be spaced on a perceptual frequency scale known as **‘Bark scale’**.

- Experiments on the response of the basilar membrane in the ear have shown a relationship between acoustical frequency and perceptual frequency resolution.

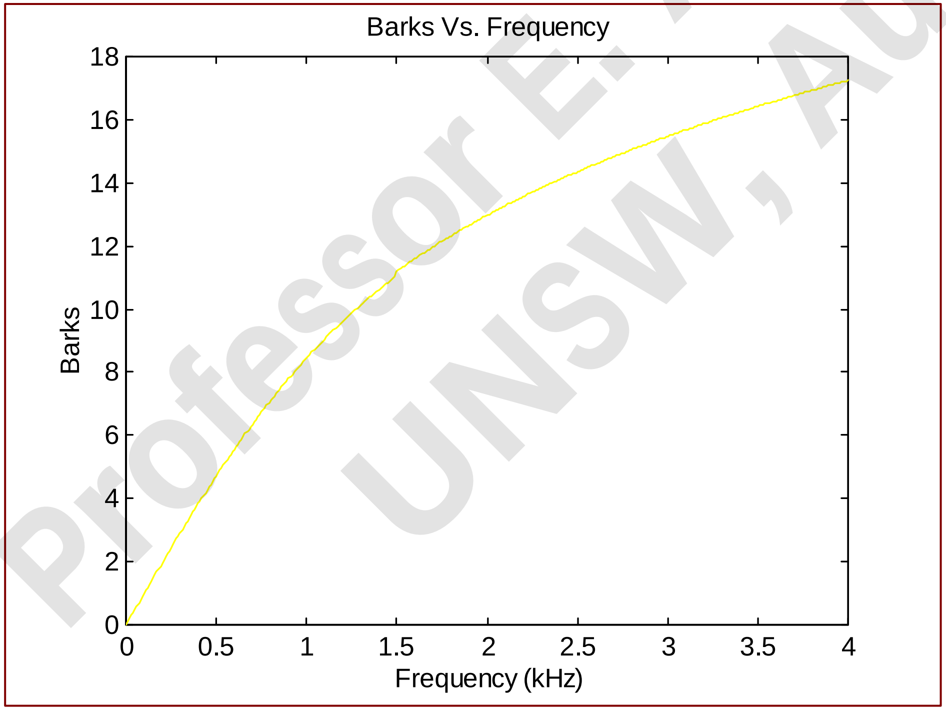
- A perceptual measure, called the Bark scale, provides the relationship between the two.

- The relationship between the frequency in Hz and the ‘critical band rate’ (with the unit of Bark) can be approximated by the following equations:

$$z_v(Bark) = 13.0\tan^{-1}\left(0.76f\right) \quad f < 1.5kHz$$

$$z_v(Bark) = 8.7 + 14.2\log_{10}\left(f\right) \quad f > 1.5kHz$$

Where f is the frequency in kHz and z_v is the frequency in Barks. Figure below shows a plot of Barks vs. frequency (in kHz) up to 4 kHz



The non-linear nature of the Bark scale can be clearly seen.

- Critical bandwidth is roughly constant at about 100 Hz for low centre frequency (< 500 Hz) (see next slide)

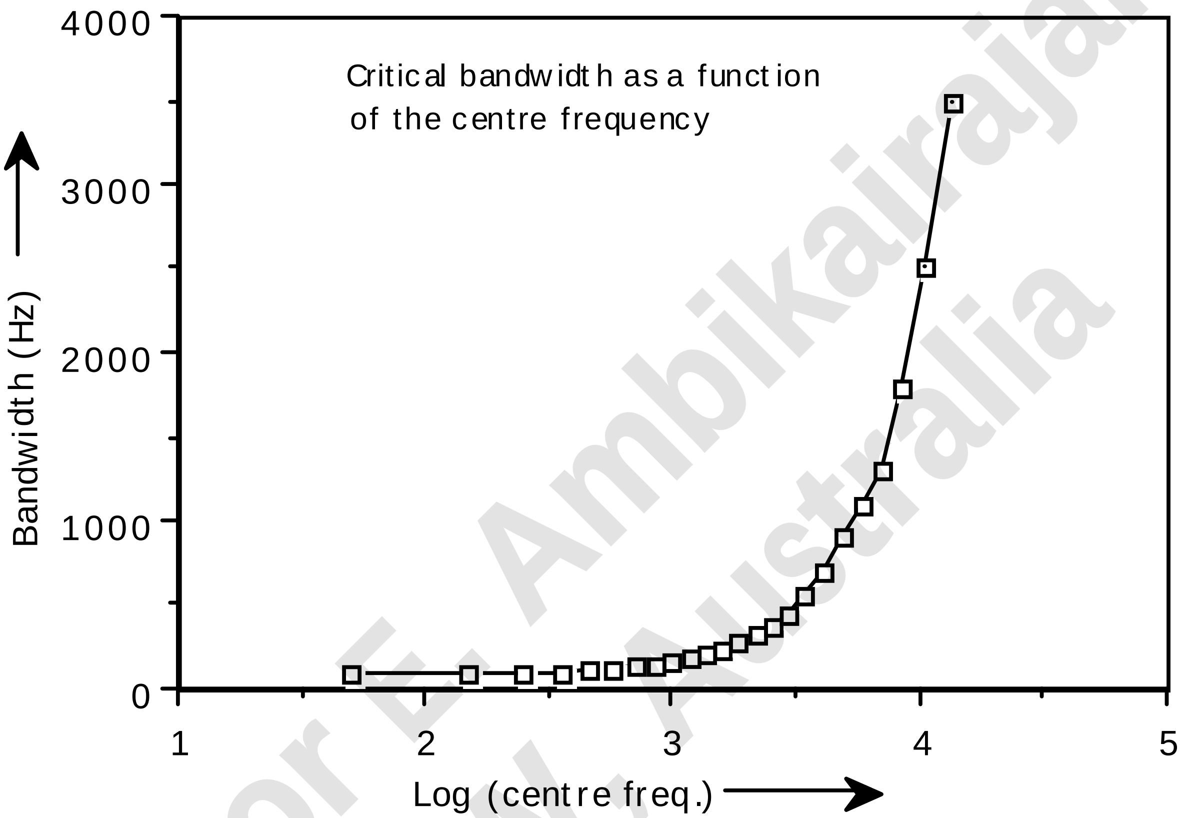
- For high frequencies, the critical bandwidth increases, reaching approximately 700 Hz at centre frequencies around 4 kHz.

- The filters are approximately constant Q at frequencies above 1000 Hz, with a Q value of 5 or 6.

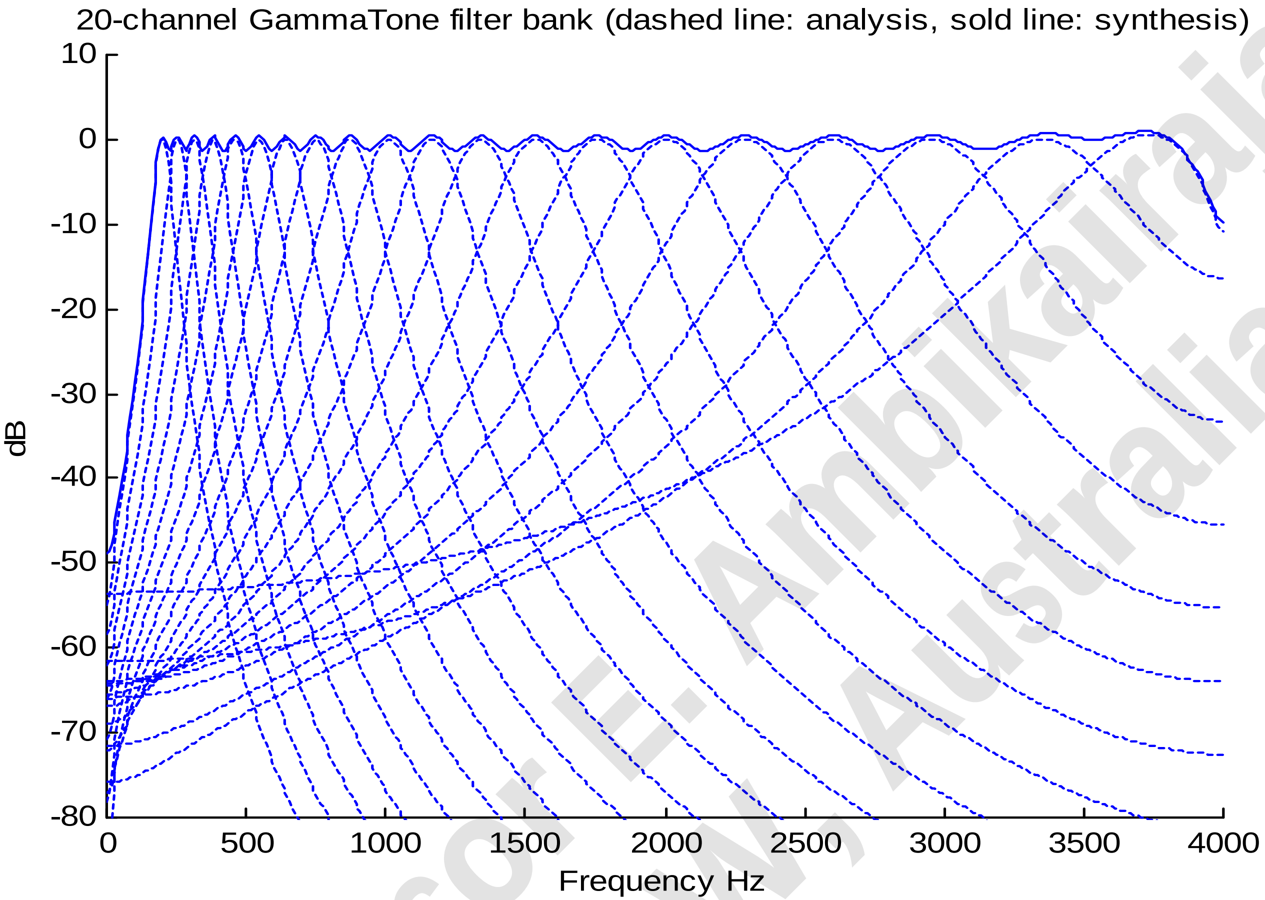
- Twenty five critical bands are required to cover

Critical Band-Rate (Bark)	Lower Edge (Hz)	Centre Freq. (Hz)	Upper Edge (Hz)	BW(Hz)	Q-factor
1	0	50	100	100	0.5
2	100	150	200	100	1.5
3	200	250	300	100	2.5
4	300	350	400	100	3.5
5	400	450	510	110	4.5
6	510	570	630	120	4.75
7	630	700	770	140	5
8	770	840	920	150	5.6
9	920	1000	1080	160	6.25
10	1080	1170	1270	190	6.15
11	1270	1370	1480	210	6.52
12	1480	1600	1720	240	6.66
13	1720	1850	2000	280	6.6
14	2000	2150	2320	320	6.72
15	2320	2500	2700	380	6.58
16	2700	2900	3150	450	6.44
17	3150	3400	3700	550	6.18
18	3700	4000	4400	700	5.71
19	4400	4800	5300	900	5.33
20	5300	5800	6400	1100	5.27
21	6400	7000	7700	1300	5.38
22	7700	8500	9500	1800	4.72
23	9500	10500	12000	2500	4.20
24	12000	13500	15500	3500	3.86
25	15500	19500	-	-	-

Critical bands of the auditory system



Variation in critical bandwidth as a function of centre frequency.



Auditory Filtering may be carried out using Gammatone filters

$$g(n) = a(nT)^{N-1} e^{-2\pi b ERB(f_c) nT} \cos(2\pi f_c nT)$$ ← Impulse response

f_c centre frequency, *T* is the sampling period, *n* is the discrete time sample index, *a*, *b* constants, and ERB(f_c) is the equivalent rectangular bandwidth of an auditory filter. At a moderate power level, $ERB(f_c) = 24.7 + 0.108 f_c$

Human Auditory Perception

- For the human auditory system, the *perception* of the sound is important.
- We do not perceive frequency but instead perceive *pitch*.
- We do not perceive level, but *loudness*.
- We do not perceive spectral shape, modulation depth, or frequency of modulation, instead we perceive *sharpness*, fluctuation strength or *roughness*.
- Also we do not perceive time directly, but perceive the subjective *duration*.

Human Auditory Perception……

- In all the hearing sensations, *masking* plays an important role in the frequency domain, as well as in the time domain.
- The information received by our auditory system can be described most effectively in the three dimensions of *loudness*, *critical-band rate* and *time*.
- The resulting three-dimensional pattern is the measure from which the assessment of sound quality can be achieved.

Masking

- The effect of masking plays a very important role in hearing, and is differentiated into two forms:
  - Simultaneous masking;
  - Nonsimultaneous masking.

Simultaneous Masking

- An example of simultaneous masking would be the case where a person is having a conversation with another person while a loud truck passes by.
- In this case, the conversation is severely disturbed and to continue the conversation successfully, the speaker has to raise his voice to produce more speech power and greater loudness.
- In music, similar effects take place where different instruments can mask each other and softer instruments become only audible when the loud instrument pauses

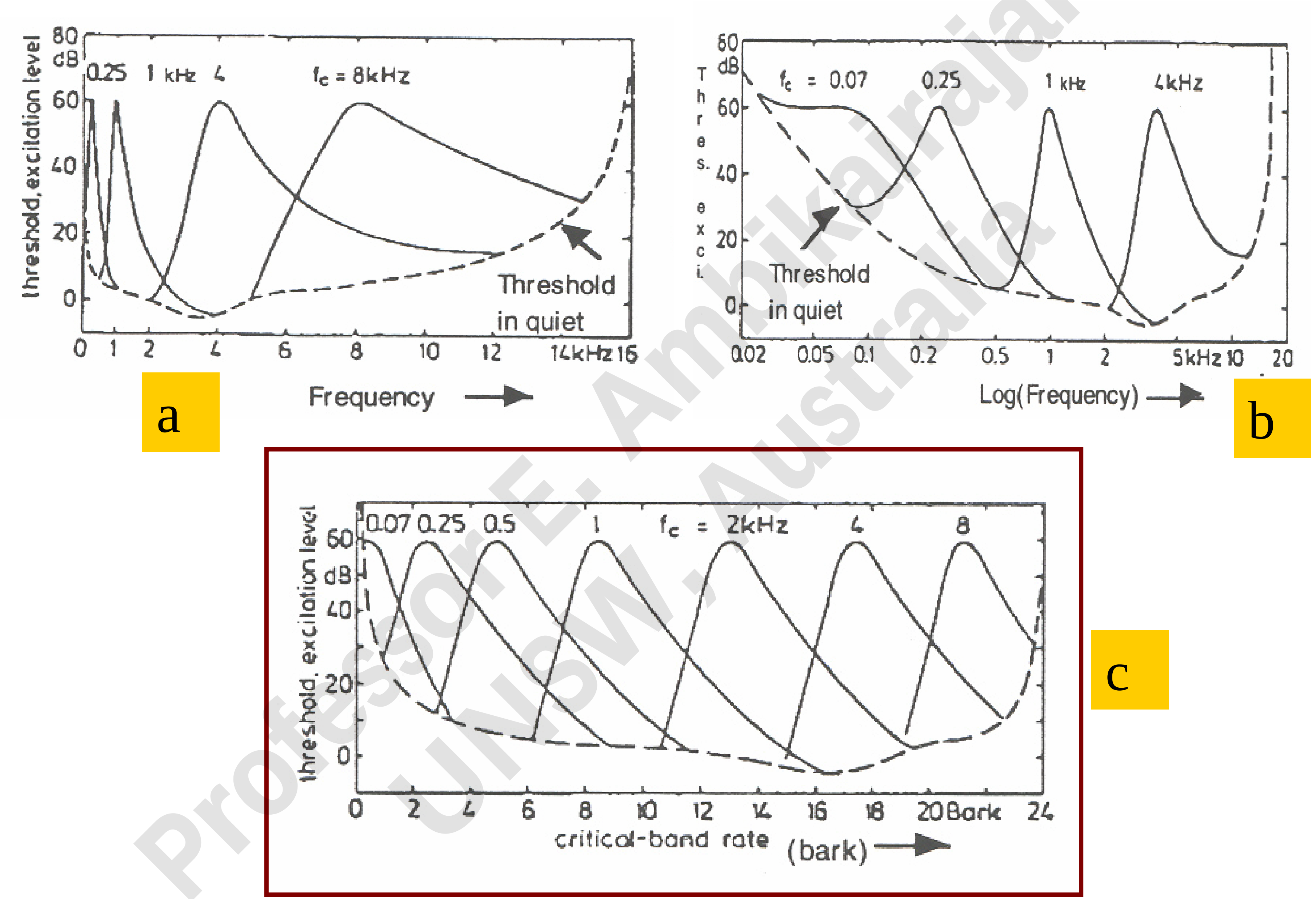
- Masking is usually described in terms of the minimum sound-pressure level of a test sound (a pure tone in most cases) that is audible in the presence of a masker.

- Figure below contains examples of maskers at different frequencies and their masking patterns.

- Most often, narrow-band noise of a given centre frequency and bandwidth is used as a masker.

- The excitation level of each masker is 60 dB.

- Comparing the results produced for different centre frequencies of the masker, we find the shapes of the masking curves are rather dissimilar irrespective of the frequency scaling (linear/ log) used.



Example of masking Curves

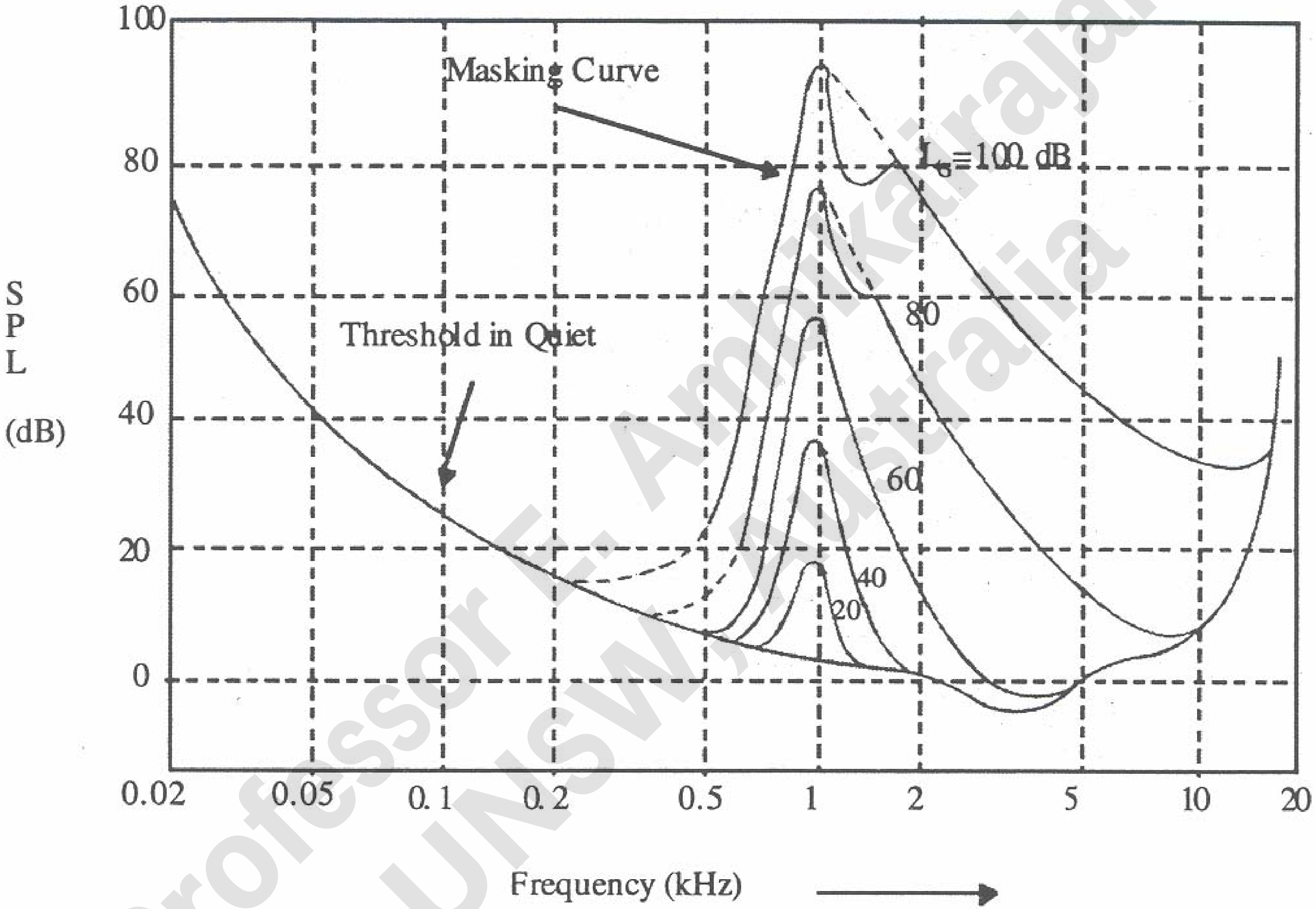
- However, one can observe that the shapes of the masking curves are similar up to about 500 Hz on linear frequency scale (Fig.(a)) while for centre frequencies above 500 Hz there is some similarity on the logarithmic frequency scale (Fig. (b)).

- These results match the critical band scale quite well, since the critical band-rate scale (as explained before) follows a linear frequency scale up to about 500 Hz and a logarithmic frequency scale above 500 Hz, and supports the notion that signals within a given critical band can be treated as a single ‘perceptual entity’.

- When frequency is converted to critical-band rate the masking pattern shown in Figs. (a) and (b) changes to those shown in Fig (c) (see previous diagram)
- The advantage of using the critical band-rate scale is obvious, namely that the shape of the masking curves for different centre frequencies are very similar (Fig. c.)
- Many other effects such as pitch, loudness etc. can be described more simply using the critical-band rate scale than using the normal linear frequency scale.

Threshold in Quiet

- The effect of masking produced by narrow-band maskers is level dependent and therefore has a nonlinear effect.
- Figure below shows the masking thresholds of narrow-band noise signals with a bandwidth of 90 Hz, centred at 1 kHz, at various sound pressure levels L_G.
- The masking thresholds for narrow-band noise signals show an asymmetry around the frequency of the masker.
- The low frequency slopes (see next slide) appear to be unaffected by the level of the masker



Threshold in quiet and masking curve of narrowband noise signals centred at 1.0 kHz at various SPLs (L_G)

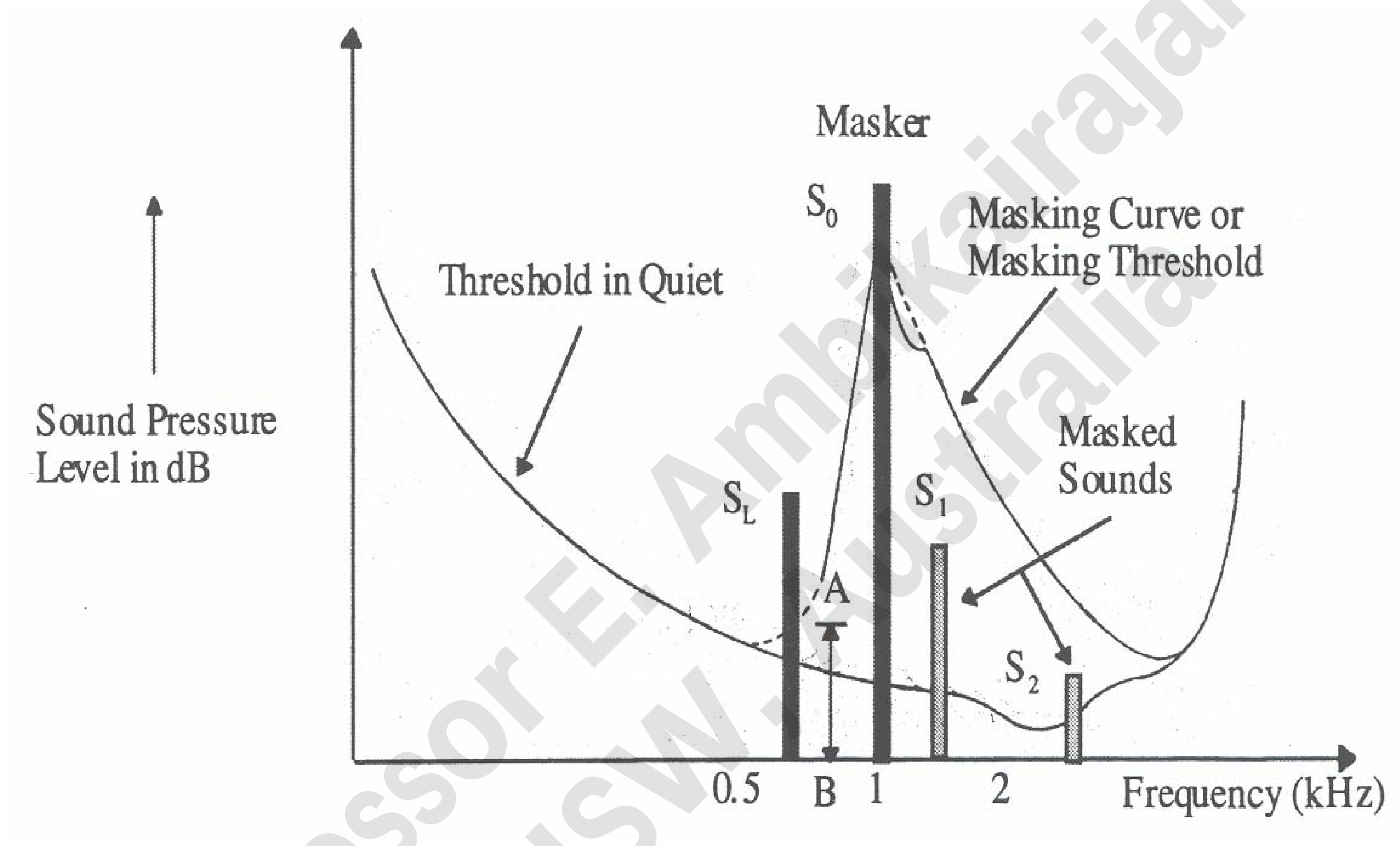
- In the figure (previous slide) threshold in quiet or absolute threshold of hearing is given as a baseline.

- All of the masking thresholds show a steep rise from low to higher frequencies up to the frequency of maximum threshold. Beyond this frequency, the masking threshold decreases quite rapidly toward higher frequencies for low and medium masker levels(L_G = 20, 40 and 60 dB).

- At higher masker levels (L_G = 80 and 100 dB) the slopes towards the higher frequencies becomes increasingly shallow. That is, signals with frequencies higher than the masker frequency are masked more effectively than signals with frequencies lower than the masker frequency

Simultaneous masking

- Simultaneous masking is a frequency domain phenomenon where a low-level signal (s_u) can be made inaudible by a simultaneously occurring stronger signal (s_o), if both signals are close enough to each other in frequency (See Figure).
- The masker is the signal S_o, which produces a masking threshold similar in shape to a Gaussian distribution.
- Any signal within the skirt of this masking threshold will be masked by the presence of S_o.
- The weaker signals S_1 and S_2 are completely inaudible. This is because their individual sound pressure levels are below the masking threshold.



Without a masker, a signal is inaudible if its sound pressure level is below the threshold in quiet

- The signal S_L is only partially masked and the perceivable portion of the signal lies above the masking curve.
- Thus, in the context of signal coding, it is possible to increase the quantisation noise in the subband containing the signal S_L up to the level AB, which means that fewer bits are needed to represent the signal in this subband.

- We have just described masking by only one masker. If the source signal consists of many simultaneous maskers, a global masking threshold can be computed as a function of frequency for the signal as a whole.

Terhardt's Auditory Masking Model

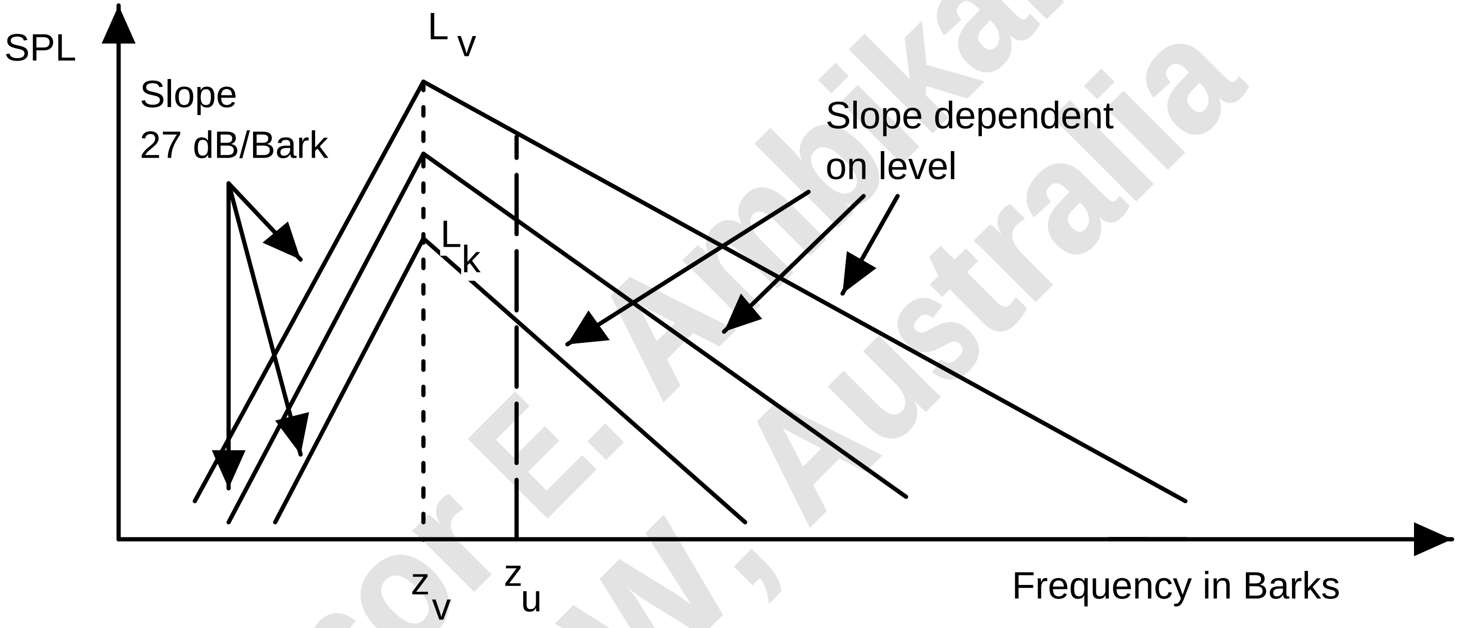
- This model is based on Tehardt's psychoacoustic model where the auditory system is represented using the critical-band rate scale.
- Spectral components within a given critical band can be masked by other components within the same critical band; this is called *intra-band* masking.
- In addition, sounds within one critical band can also mask other sounds in different critical bands. This is called *inter-band* masking.

Auditory Masking Model

- Experiments on pitch perception carried out by Terhardt have shown that there is a direct relationship between the level of a masker and the amount of masking it induces on another frequency component.

- Tehardt approximated the masking curves shown in the next slide using straight lines and used the characteristic to represent the masking effect produced by a spectral component of frequency z_v (Barks) on another spectral component of frequency z_u (Barks).

Masking Threshold produced by a spectral component at frequency z_v (Barks) for various SPLs



The high frequency slope (s_{vh}) for the masking threshold curve is given by

$$s_{vh} = -24 - \frac{230}{f_v} + 0.2\ L_v \qquad dB/Bark$$

- where L_v is the level of the masker (in dB SPL), f_v is the masker component frequency in Hz and s_{vh} is the slope. Tehardt’s experiments showed that the sound pressure level of the masker is not so important when computing the masking effect on lower frequencies.

- Thus, the low-frequency slope(s_{vl}) of the masking curve is independent of L_v and is set to 27 dB/Bark.

- If the spectrum contains N frequency components, the overall masking threshold of a component at z_u (Barks) due to all other components in the spectrum is given by

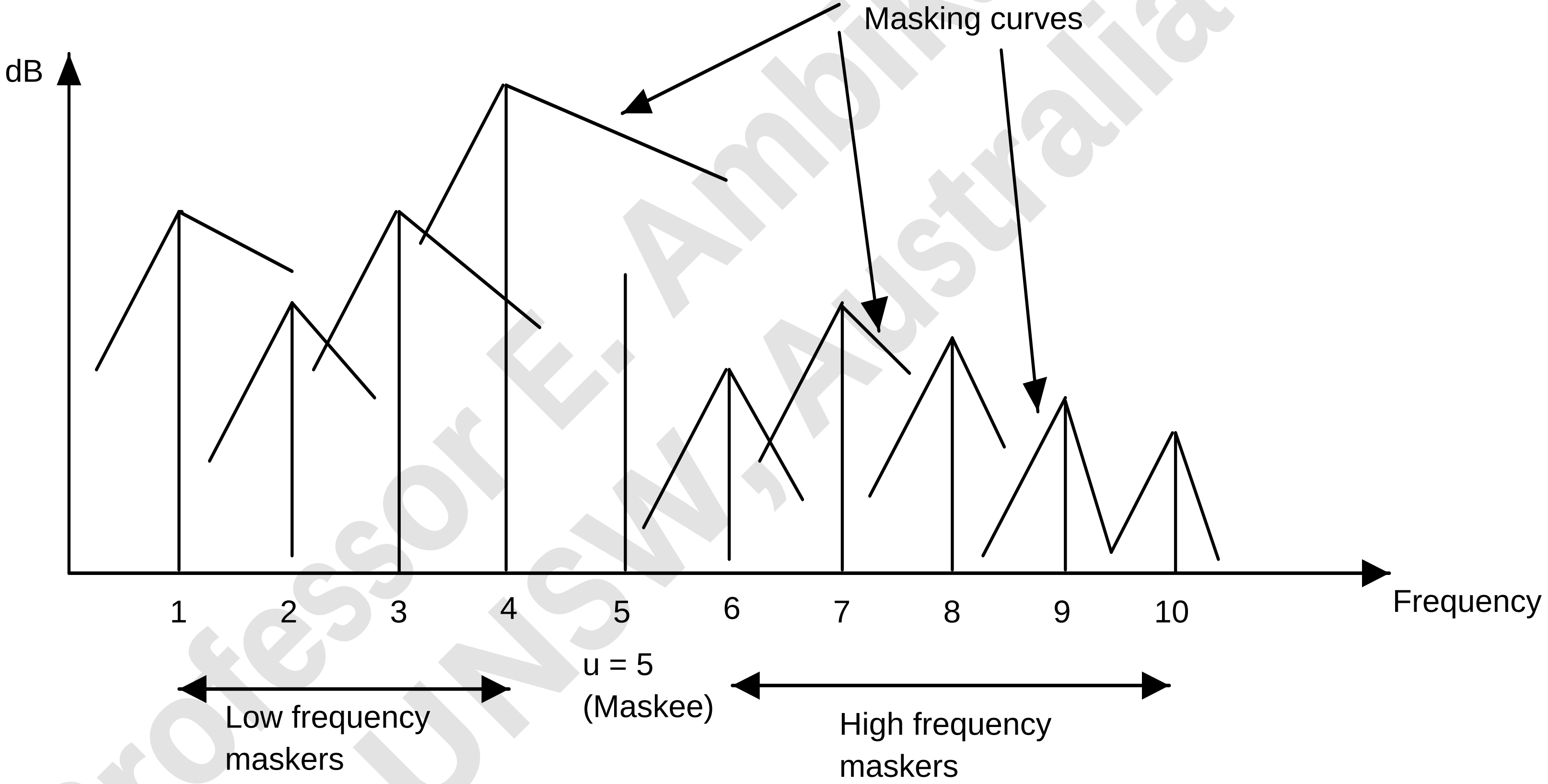
$$Th\,(z_u) = 20\,\log_{10}\left\{ \sum_{v=1}^{u-1} \underset{\uparrow}{10^{\frac{1}{20}\left[L_v - S_{vh}(z_v - z_u)\right]}} + \sum_{v=u+1}^{N} \underset{\uparrow}{10^{\frac{1}{20}\left[L_v - S_{vl}(z_v - z_u)\right]}} \right\} v \neq u$$

A maskee u being masked by a lower frequency masker v

A maskee u being masked by a higher frequency masker v

- Note that the above equation is not evaluated for u=v. i.e.it is assumed that the maskee does not mask itself. The resultant **inter-band** masking threshold value can be estimated using the above equation (previous slide)
- Example: There are $N = 10$ spectral components, with the component at $u = 5$ being the maskee.
- All other frequency components will mask this component. The resultant masking threshold value can be estimated using the equation given in the previous slide

Masking calculation



Intra-band masking

- The next step is to take the effect of intra-band masking into account.
- There are two types of masking that have been experimentally observed, which can occur within a critical band.
- The first one is usually referred to as **‘tone-masking-noise’** and
- The second one is **‘noise-masking-tone’**.

- **Tone Masking Noise : $E_N = -(14.5 + i)$ dB**

- **Noise masking Noise: $E_T = -5.5$ dB**

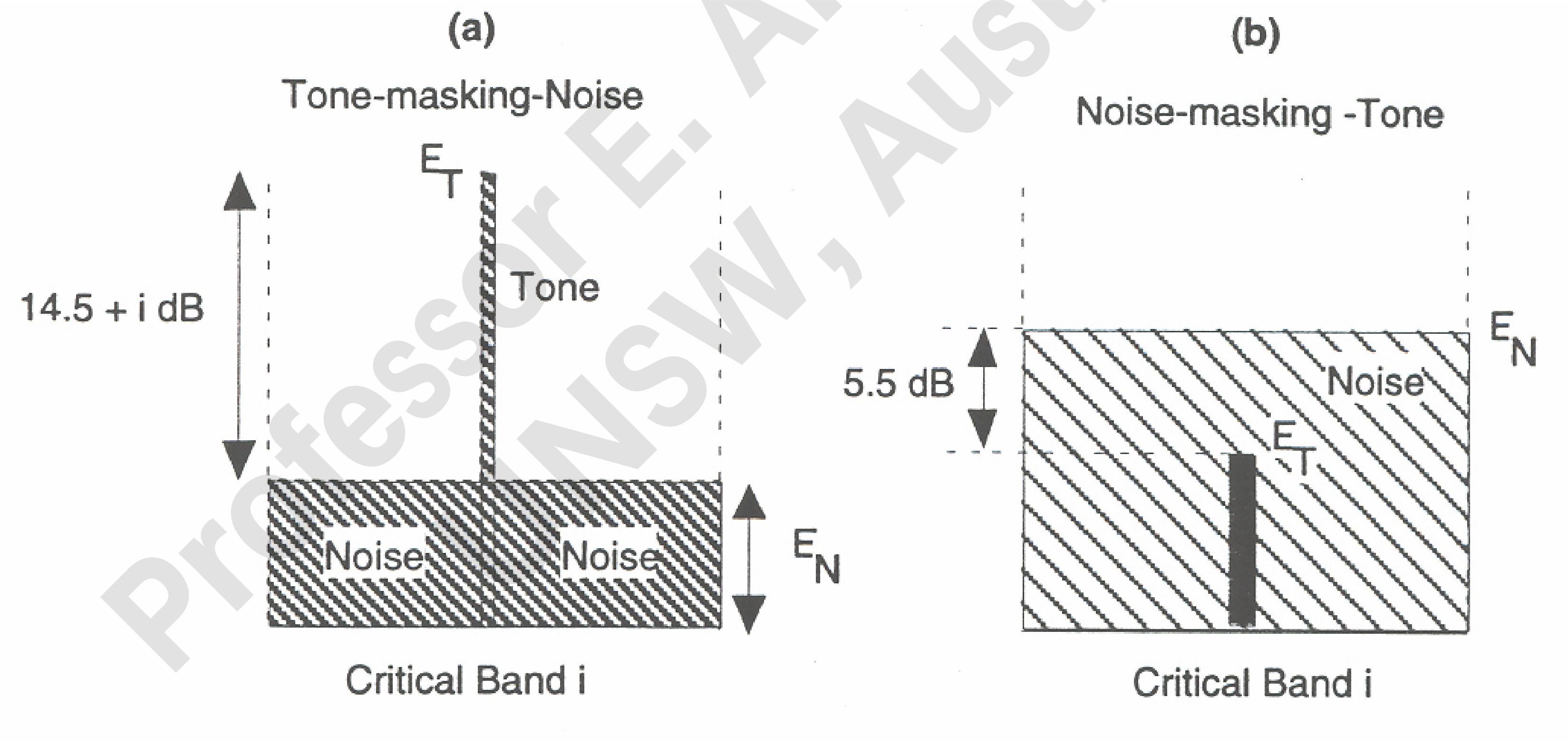
- **where E_T and E_N are tone and noise energies, i is the critical band number.**

- **From the first equation (see above) states that a tone will mask the noise in a critical band if the power of the tone is at least 14.5 + *i* dB higher than the noise power (see next slide (a)).**

- **It is evident from the above equation that in higher critical bands the power of the tone must be higher in order to mask the same noise power as in the lower critical band. This is because the critical bandwidth increases with**

❑Noise masking Noise: E_T = -5.5 dB

❑Similarly using the above Equation, one can see that a tone will be masked within a critical band if the tone is 5.5 dB lower than the noise energy in the same band (see slide b below)



- **There are many ways of calculating the tone-like or noise-like nature of the signal .**

- **For simplicity it is assumed here that a signal in a lower critical band (up to 2.5 kHz) is more tone-like in nature while a signal in a higher critical band is more noise-like, as the higher critical bands have wider bandwidths.**

- **Previous equations can now be rewritten as**

$$E_N = -K.(14.5 + i)\ \text{dB} \qquad 2.5\,kHz < f \leq 4\ kHz \quad 15 \leq i \leq 17$$

$$E_T = -K.(42.5 - i)\ \text{dB} \qquad 0 \leq f \leq 2.5\,kHz \quad 0 \leq i \leq 14$$

where *K* is a scaling factor that takes a value between 0.5 and 1.0.

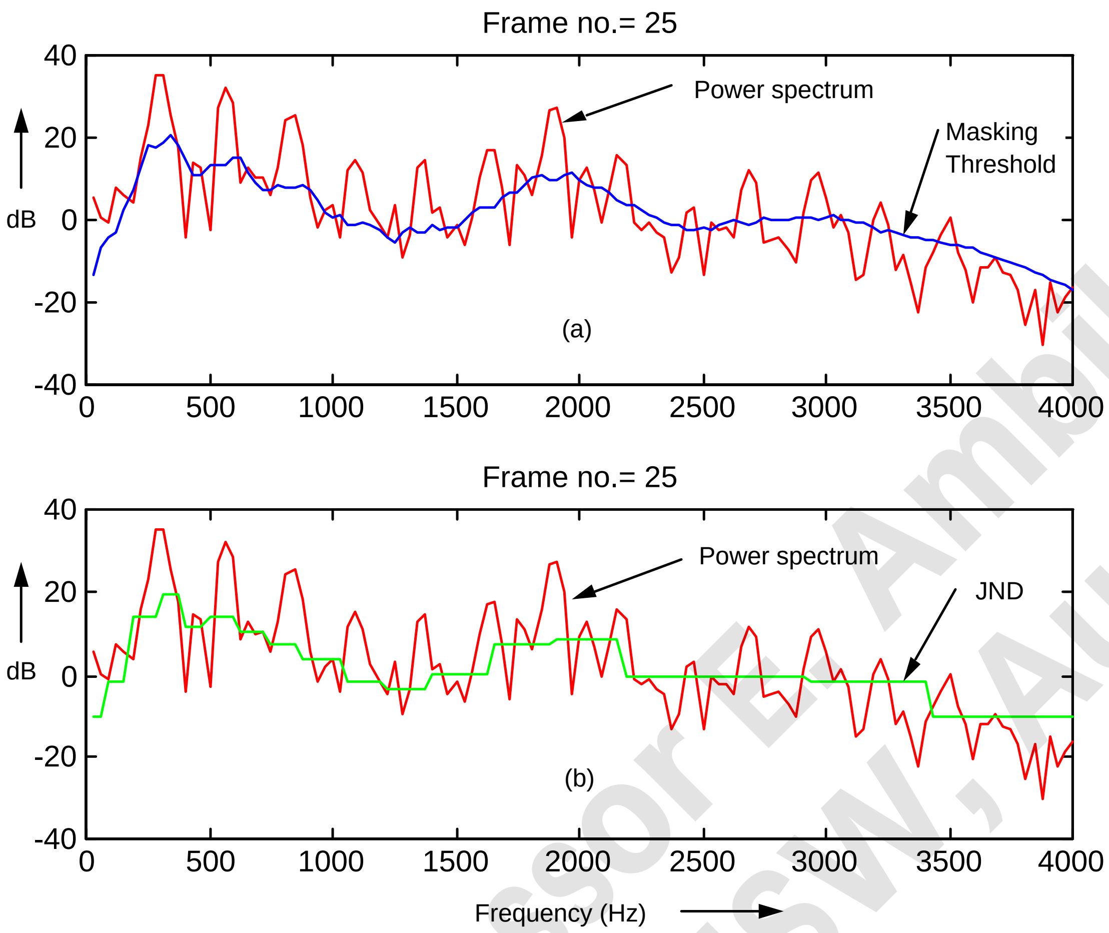
- The overall masking threshold is now given by

$$Nth(z_u) \quad = \quad Th(z_u) + E_N \text{ (or } E_T)$$

- Above Equation is evaluated for every frequency component in the spectrum thus obtaining a global masking threshold as a function of frequency.

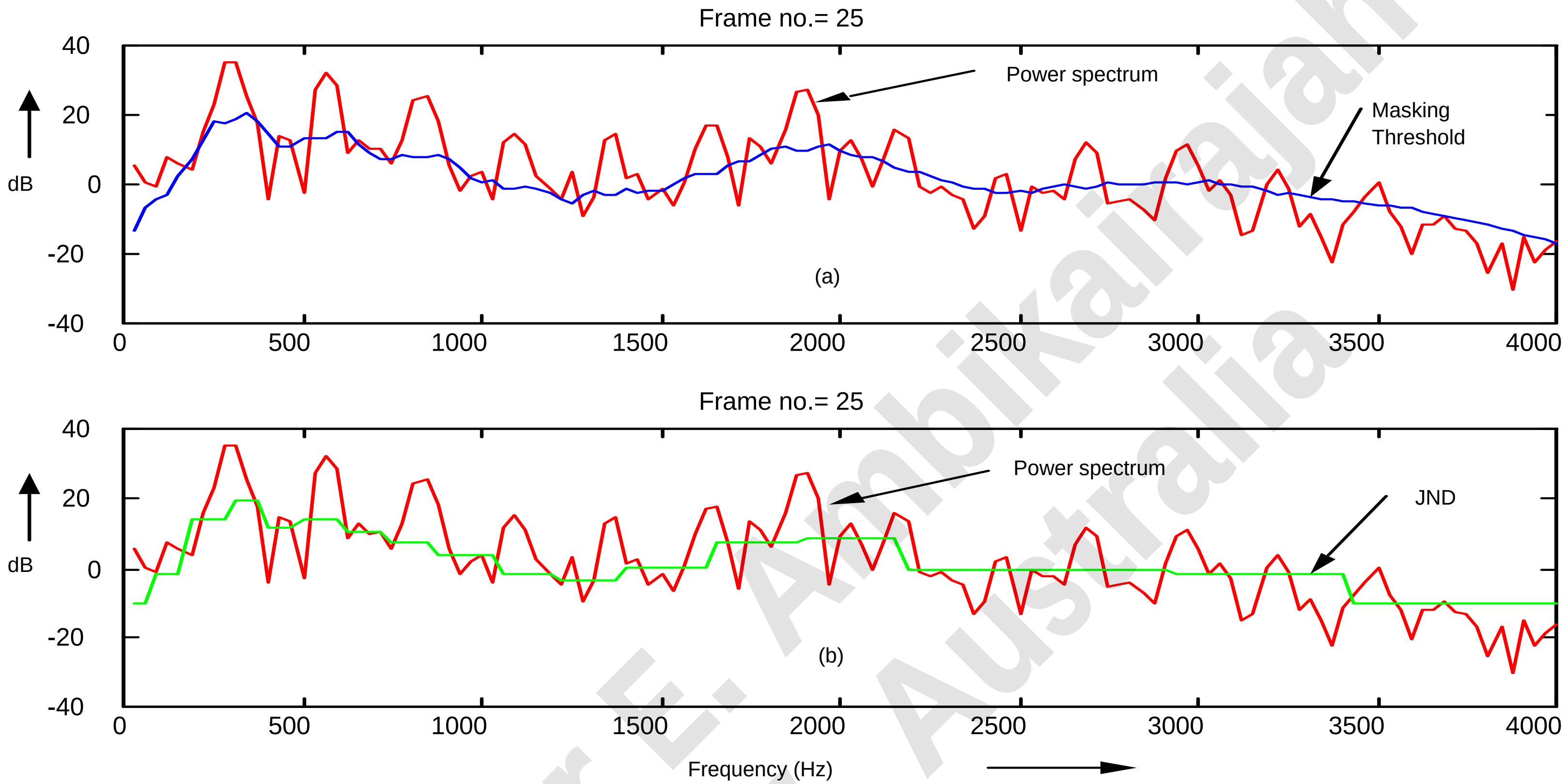
- From the overall masking threshold values, the Just Noticeable Distortion (JND) vale in each critical band can be calculated, by selecting the minimum value of Nth(z_u) in that band.

- Any signal component above the JND value in each critical band conveys signal information, while signal components below this threshold



- **Figure (a) below shows a plot of the power spectrum of one frame (256-point FFT used) of a voiced speech signal, at 8 kHz, along with the calculated global masking threshold values**

- **Figure (b) plots the same power spectrum along with a plot of the minimum threshold value (JND) in each critical band.**



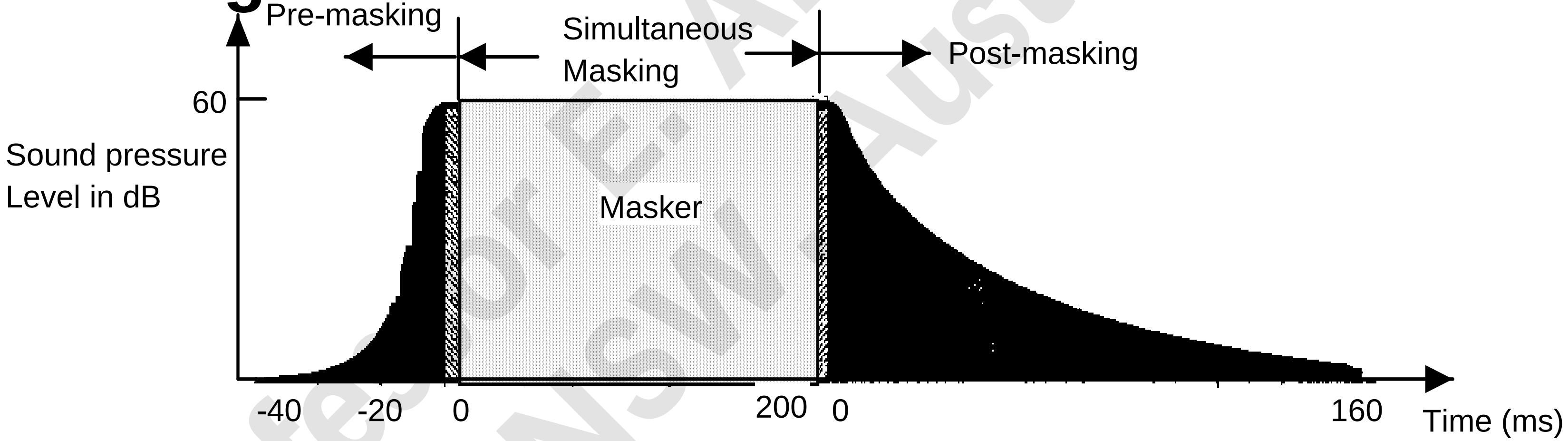
- **As can be seen , the JND value for each band is simply minimum value of the masking threshold in that band.**

- **The distribution of the critical bands can be seen with the JND values changing sharply from band to band.**

Nonsimultaneous masking

- Nonsimultaneous masking is also referred to as *temporal masking*. Temporal masking may occur when two sounds appear within a small interval of time.
- Two time domain phenomena play an important role in human auditory perception,:
  - pre-masking
  - post-masking.

- **Temporal masking is illustrated in the diagram shown below. When the signal precedes the masker in time, the condition is called post-masking; when the signal follows the masker in time, the condition is pre-masking.**



Temporal Masking. Acoustic events in the dark areas will be masked.

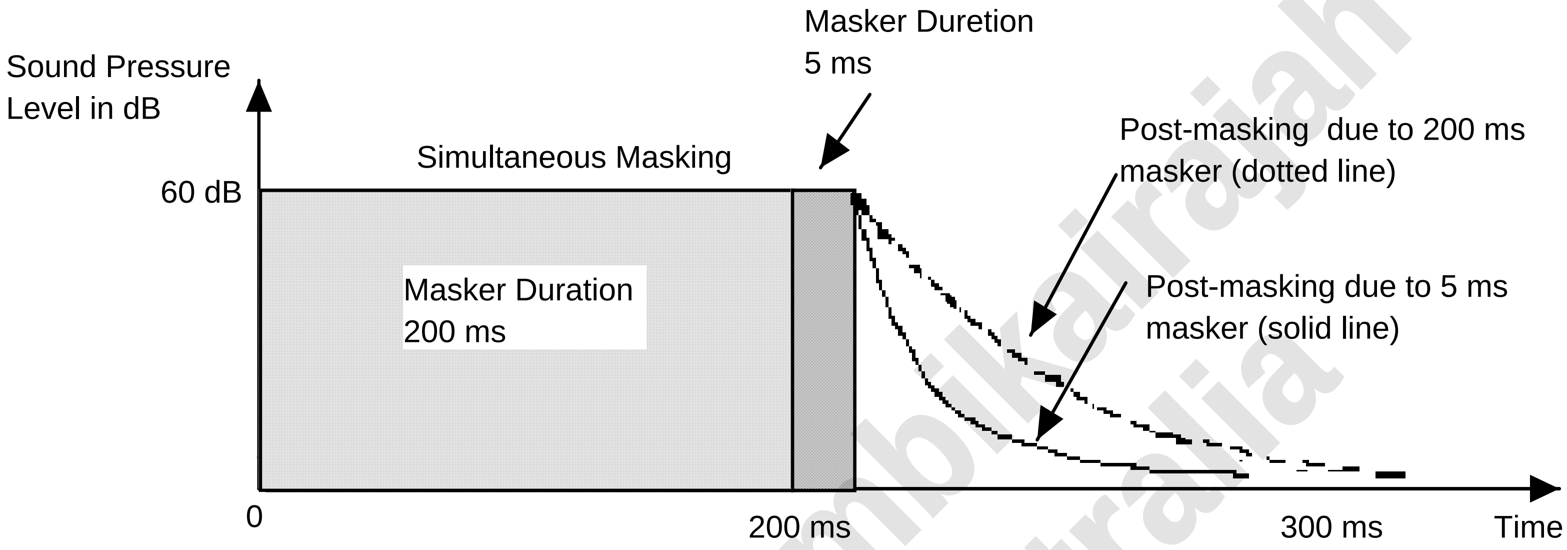
- **Post-masking is the more important phenomenon from the point of view of efficient coding.**

- **It results from the gradual release of the effect of the masker, i.e. masking does not immediately stop when the masker is removed, but rather continues for a period of time following this removal.**

- **The duration of post-masking depends on the duration of the masker.**

- **In the diagram (see next slide), the dotted line indicates post-masking for a long masker duration of at least 200ms.**

- **The degree of post masking decreases from the**



- **Postmasking produced by very short masker burst , such as 5 ms (See above) behaves quite differently.**
- **Post-masking in this case decays much faster so that after only 50 ms the threshold in quiet is reached. This implies that post-masking strongly depends on the duration of the masker and therefore is another highly nonlinear effect.**

Temporal masking Model I

- This model is based on the fact that temporal masking decays approximately exponentially following each stimulus. The masking level calculation for the *m*th critical band signal $M_f(t,m)$ is

$$M_f(t,m) = \begin{cases} L(t,m), & L(t,m) > c_0 L(t-\Delta t, m) \\ c_0 L(t-\Delta t, m), & otherwise \end{cases}$$

where $c_0 = \exp(-\tau_m)$. The amount of temporal masking *TM1* is then chosen as the average of $M_f(t,m)$ for each frame calculation.

- **Normally first order IIR low-pass filters are used to model the forward masking. The time constant,τ_m, of these filters are as follows, in order to model the duration of forward masking more accurately.**

$$\tau_m = \tau_{\min} + \frac{100\,Hz}{fc_m} \cdot \left(\tau_{100} - \tau_{\min}\right)$$

- **The time constants $\tau_{\min}$ and τ_{100} used were 8 ms and 30 ms, respectively. The time constants were verified empirically by listening tests and were found to be much shorter than the 200 ms post-masking effect commonly seen in literature.**

Temporal masking Model II

- Jesteadt et al describe temporal masking as a function of frequency, masker level, and signal delay.
- Based on the forward masking experiments carried out by Jesteadt, the amount of temporal masking can be well-fitted to psychoacoustic data using the following equation:

$$M_f(t,m) = a(b - \log_{10} \Delta t)(L(t,m) - c)$$

$$M_f(t,m) = a(b - \log_{10}\Delta t)(L(t,m) - c)$$

- **where $M_f(t,m)$ is the amount of forward masking (dB) in the *m*th band, Δt is the time difference between the masker and the maskee in milliseconds, is the masker level (dB), and *a*, *b*, and *c*, are parameters that can be derived from psychoacoustic data.**

- **The parameter *a* is based upon the slope of the time course of masking, for a given masker level.**

- **Assuming that forward temporal masking has duration of 200 milliseconds, and thus *b* may be chosen as $\log_{10}(200)$**

- **Similarly *a*, *c* are chosen by fitting a curve to the masker level data provided by Jestead**

Combined Masking Threshold

- A combined masking threshold may be calculated by considering the effect of both temporal and simultaneous masking.

$$MT = \left(TM^{p} + SM^{p}\right)^{1/p}, \ 1 \leq p \leq \infty$$

- **where *MT* is the total masking threshold, *TM* is temporal masking threshold, and *SM* is the simultaneous masking threshold. The parameter *p* defines the way the masking thresholds add. *P* is chosen as 5**

ELEC9344:Speech & Audio Processing

Chapter 14 (week 14)

Wideband Audio Coding

Introduction

- Reduction in bit rate requirement for high quality audio has been an attractive proposition in applications such as multimedia, efficient disk storage, and digital broadcasting.
- A number of audio compression algorithms exists
- Among them, the most notable is the ISO/MPEG standard, which is based on Modified Discrete Cosine Transform method and provides high quality at about 64 kb/s.

Wideband Audio Coding

- The data rate of a high fidelity stereophonic digital audio signal is about 1.4 Mb/s for 44.1 kHz sampling rate and 16 bits/sample uniform quantisation.

- This rate is simply too high for many transmission channels and storage media.

- It severely limits the application of digital technology at a time when high quality audio is becoming increasingly important.

- As a result, data reduction of digital audio signals has recently received much attention.

- However, low bit-rate coding can introduce distortion such that listeners may deem the sound quality of the decoded signal unacceptable.

- The masking properties of the human ear can provide a method for concealing such distortion.

- The most successful of the current low bit-rate wideband coders is ISO/MPEG which is based on subband coding and use psychoacoustic models to determine and to eliminate redundant audio information.

- This coder gains in efficiency by first dividing the frequency range into a number of bands, each of which is then processed independently.

- The algorithm results in data rates in the range of 2 - 4 bits/sample.

- If more than one channel sound is to be processed then samples from each channel are treated independently.

- First, for each channel the masking threshold is determined.

- Then redundant, masked samples, are discarded and the remaining samples are coded using a deterministic bit allocation rule.

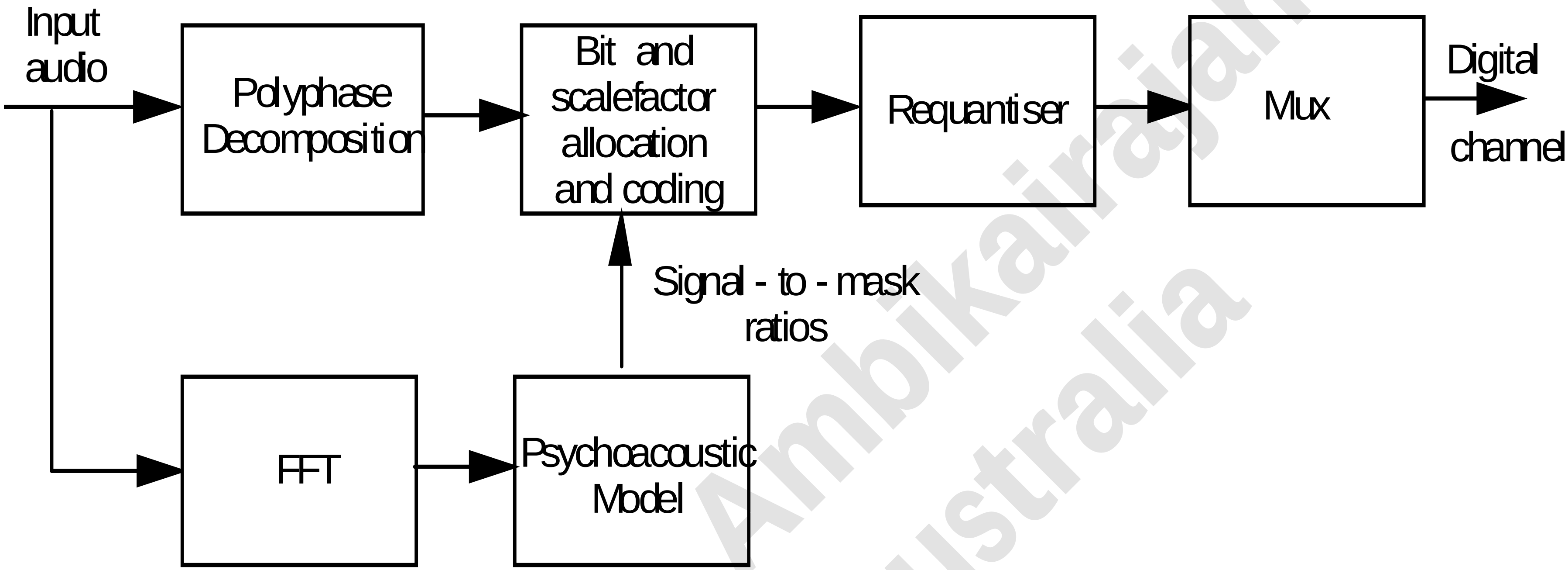
ISO/MPEG Layer -I

- In ISO/MPEG Layer -I model the filterbank decomposes the audio signal into 32 equal bandwidth subbands.

- Efficient implementation is achieved by a polyphase filterbank, which however, cannot provide the resolution required the psychoacoustic model.

- Therefore, the ISO/MPEG coder employs an FFT analyser which further increases the overall computational load.

- Figure 1 shows the main functional elements used by the ISO/MPEG coder.



Block Diagram of the ISO/MPEG Layer -I coder

We can show that the sub band decomposition carried out using Wavelet Packet (WP) decomposition provides sufficient resolution to extract the time-frequency characteristics of the input signal thus eliminating the requirement for a separate FFT analysis to derive a psychoacoustic model.

Wideband Audio Coding Algorithms

- Some of the important algorithms and standards for wideband speech and audio coding is reviewed in this section. There are two fundamentally different techniques are available for the compression of PCM audio data:

- Time domain coding
- Frequency domain coding

- **Time domain coders** exploit temporal redundancy between audio samples such that one can maintain the same Signal-to-Noise ratio at a reduced bit rate (e.g. Differential PCM coders).

- Frequency domain coders are designed to identify and remove redundancy in frequency domain.
- A common features of all frequency domain coders is the time-frequency transform, which maps a nonstationary signal onto the time-frequency plane.
- This mapping may be achieved by a transform, resulting in a transform coder or by subband decomposition, resulting in a subband coder.
- The time-frequency representation lends itself to the identification and removal of perceptually redundant signal components.
- The subband samples are quantised with the minimum resolution necessary to ensure that the quantiser noise is below the threshold of perceptible distortion.

- Powerful algorithms and standards for wideband speech and audio coding enhance service in communication and other applications.

- **Wideband speech** covers 50 Hz to 7 kHz frequency band and **wideband audio** covers 10 Hz to 20 kHz frequency band.

- These two signals differ not only in bandwidth, but also in listener expectation of offered quality.

- Table 1 provides an overview of wideband speech and audio coding algorithms.

Standard	Input	Coder	Rate (kb/s)
CCITT G.721	Toll-quality Speech	ADPCM	32
CCITT G.722	Wideband Speech	SB, ADPCM and QMF	48, 56, 64
LD-CELP	Wideband Speech	LP and VQ	8, 16, 32
ISO/MPEG	Wideband Audio	SB, TC, EC and PaM	32 - 192
MUSICAM	Wideband Audio	SB and PaM	64 - 192
PASC	Wideband Audio	SB and PaM	128 - 192
ASPEC	Wideband Audio	TC, EC and PaM	64 - 192

Wideband speech and audio coding techniques

ADPCM:	Adaptive differential pulse code modulation
EC:	Entropy coding
LP:	Linear prediction
PaM:	Psychoacoustic model
QMF:	Quadrature mirror filter
VQ:	Vector quantisation
SB:	Subband coding
TC:	Transform coding

Wavelet Packet based scalable audio coder

- The objective is to use wavelet packet decomposition as an effective tool for data compression and to achieve the high quality low complexity scalable wavelet based audio coding.

The proposed features:

□ The bit rate can be scaled to any desired level to accommodate many practical channels

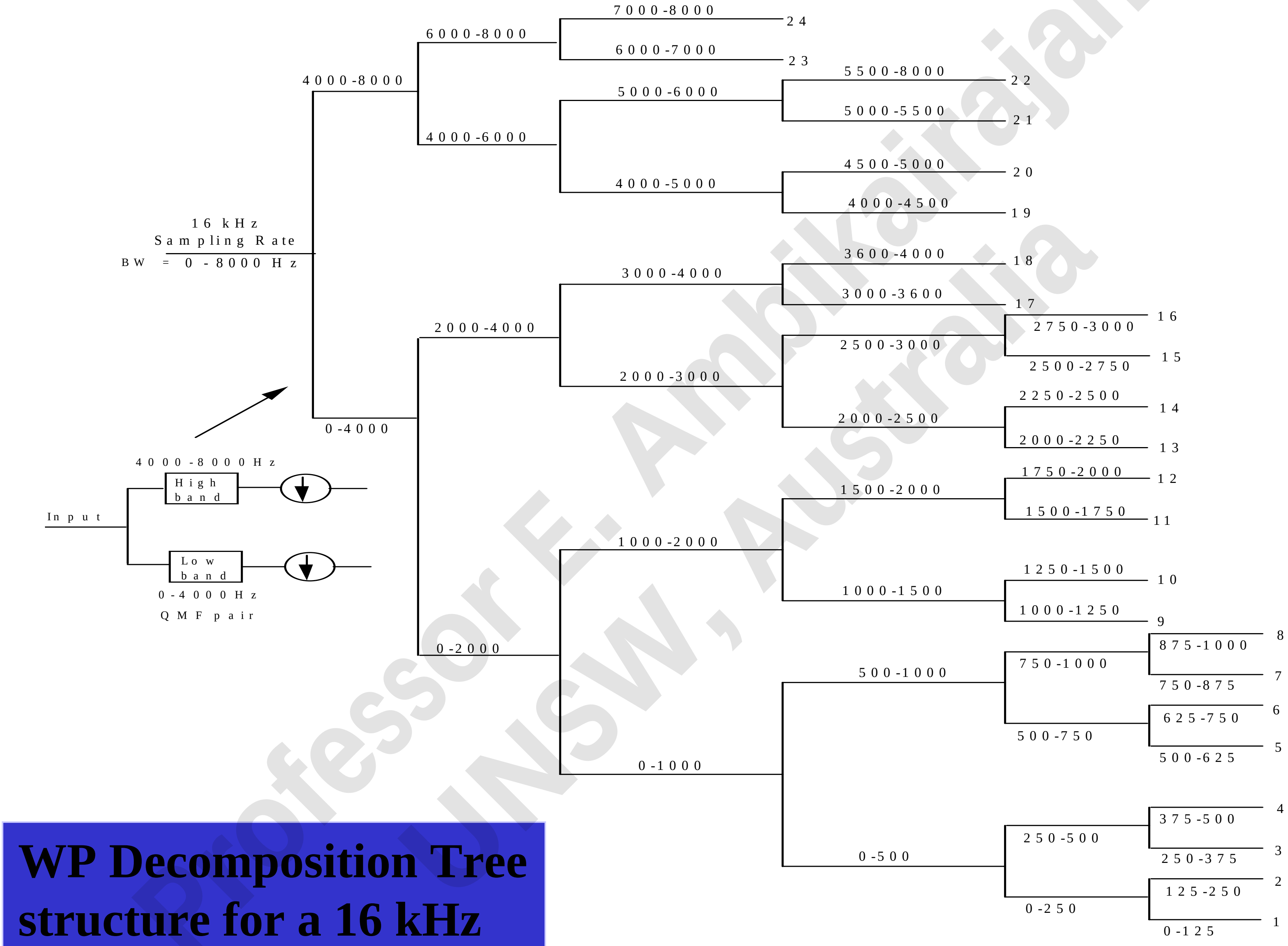
□ Most industrial standard sampling rates can be supported (e.g. 44.1 kHz, 32 kHz, 22 kHz, 16 kHz and 8 kHz)

- An example of a 24-band WP representation is shown in the next slide where the sampling rate is 16 kHz.

- This filterbank structure is identified because it has sufficient resolution for direct implementation of the psychoacoustic model.

- Also the subband bandwidths and centre frequencies closely approximate the critical bands.

- The subband numbering (see figure) does not take into account the switching of the highpass and lowpass spectra as the output of each highpass branch in the decomposition tree is decimated.



WP Decomposition Tree structure for a 16 kHz sampling rate

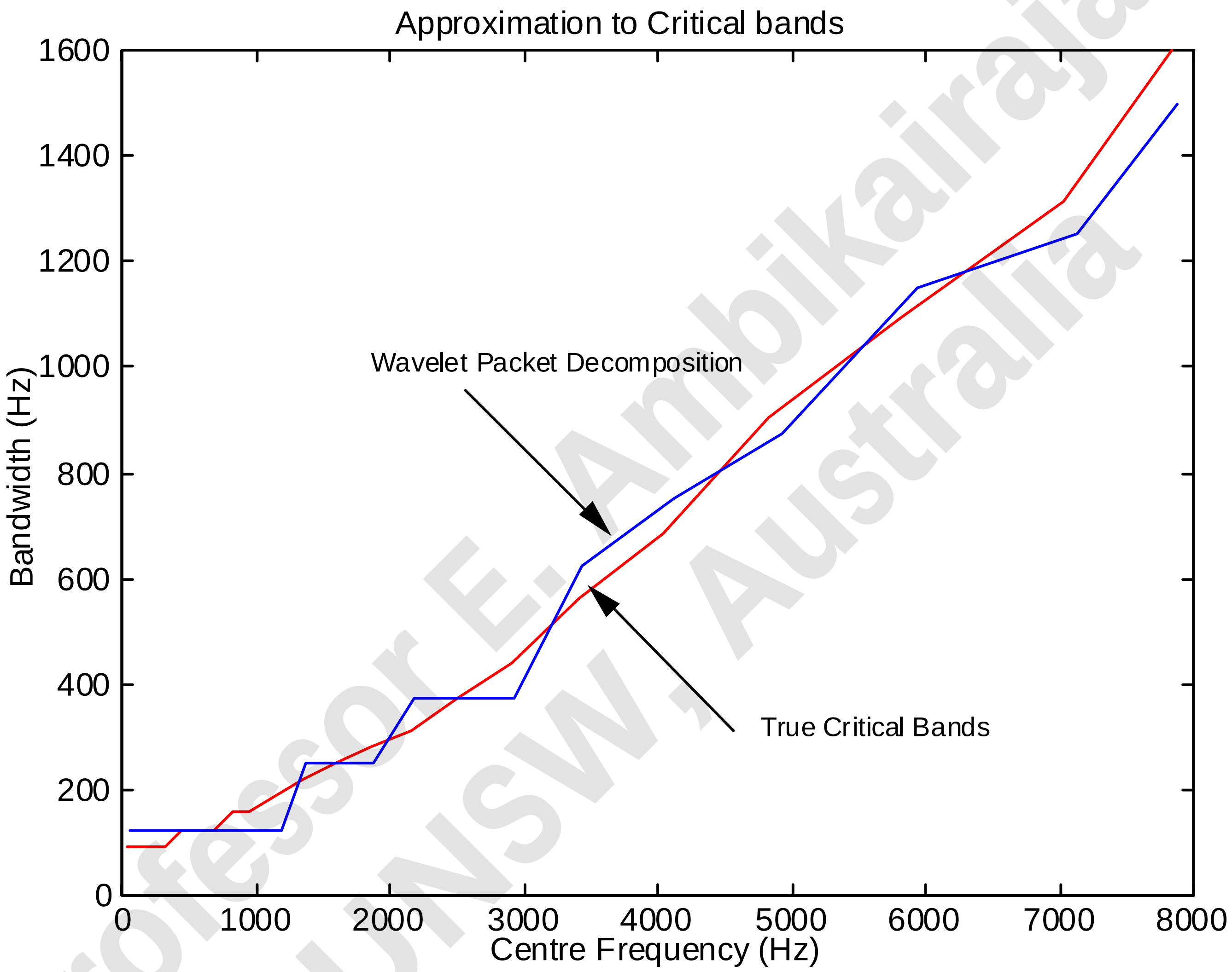
- Appropriate numbers for reordering the spectra can be illustrated, for example, using a 4 level Wavelet Packet decomposition tree as shown in the Table below:

Band No: ->	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**		**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
				L									**H**				
Level 1	1	2	3	4	5	6	7	8		16	15	14	13	12	11	10	9
			L			**H**						**L**			**H**		
Level 2	1	2	3	4	8	7	6	5		16	15	14	13	9	10	11	12
	L		**H**		**L**		**H**			**L**		**H**		**L**		**H**	
Level 3	1	2	4	3	8	7	5	6		16	15	13	14	9	10	12	11

L - Lowpass subband; H - Highpass subband

- The diagram (see next slide) displays the bandwidths of the critical band filters versus their respective centre frequencies.

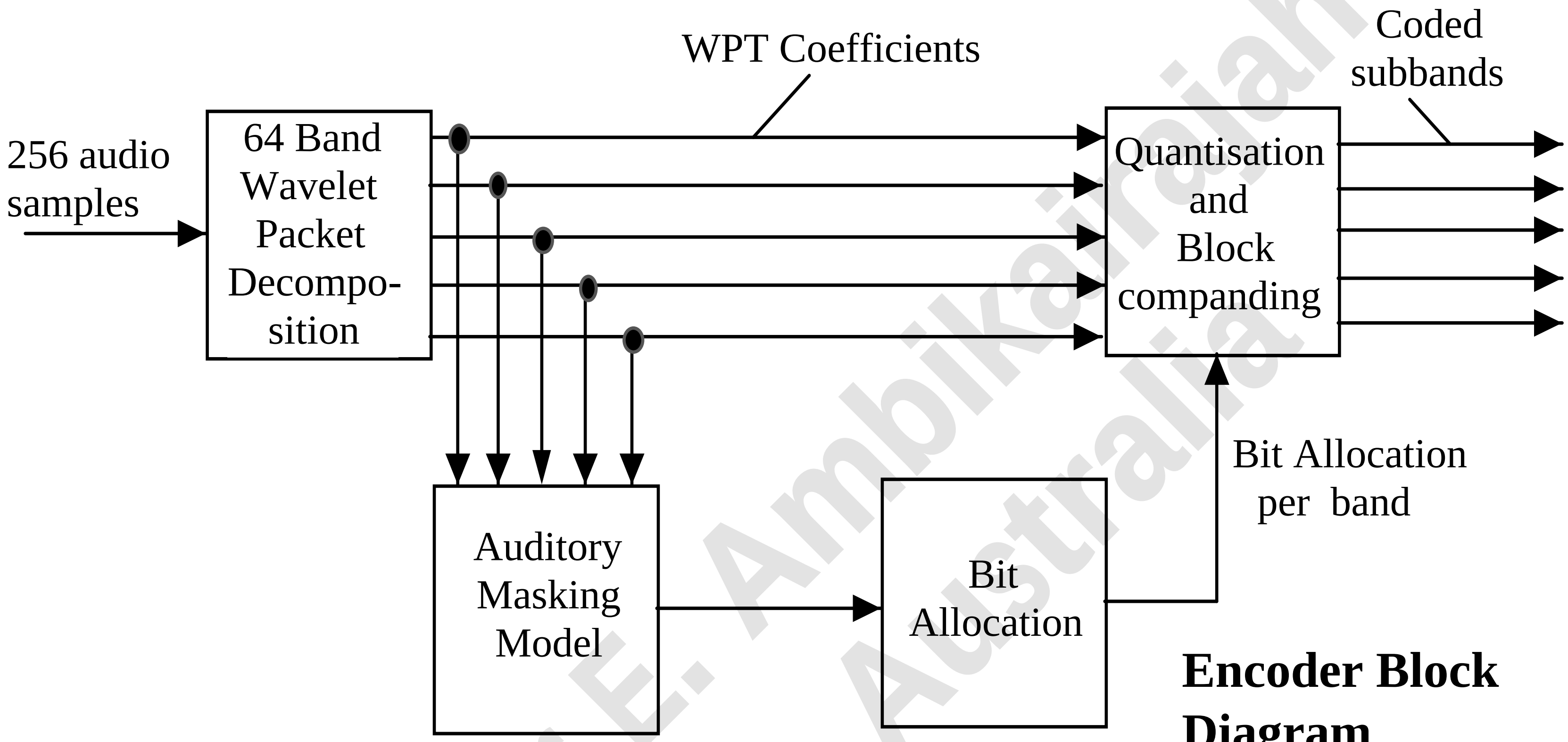
- The WP decomposition closely approximates the critical bands, allowing the output of the WP expansion to directly drive the psychoacoustic model thereby eliminating the need for an FFT, and reducing the computational effort.



Comparision of resolution resulting from WP decomposition and True Critical bands

Coder Structure

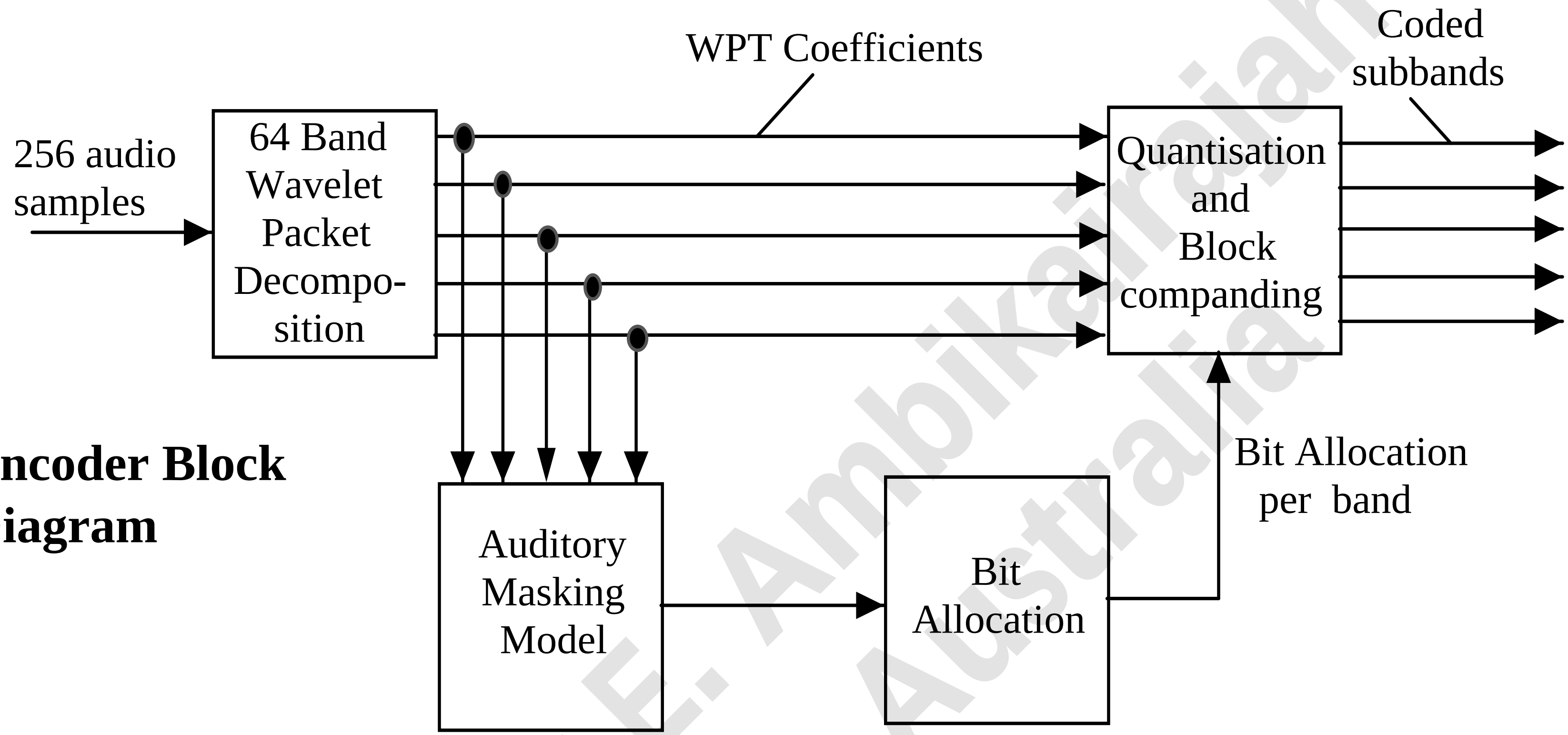
- A block diagram of a Wavelet Packet decomposition based audio coder is shown in the next slide where the sampling frequency of the audio signal is 16 kHz.
- A six-level decomposition is carried out thus resulting in a 64 band WP decomposition.
- Psychoacoustic auditory masking is a phenomenon whereby a weak signal is made inaudible by a simultaneously occurring stronger signal.
- Most progress in audio compression in recent years can be attributed to successful application of auditory masking model.



Encoder Block Diagram

- In a psychoacoustic model, the signal spectrum is divided into a number of critical bands.
- In the above implementation, the 64 band WP decomposition are grouped together in a particular manner to obtain 22 critical bands and an auditory masking model could then be directly applied in the wavelet domain.

Encoder Block Diagram



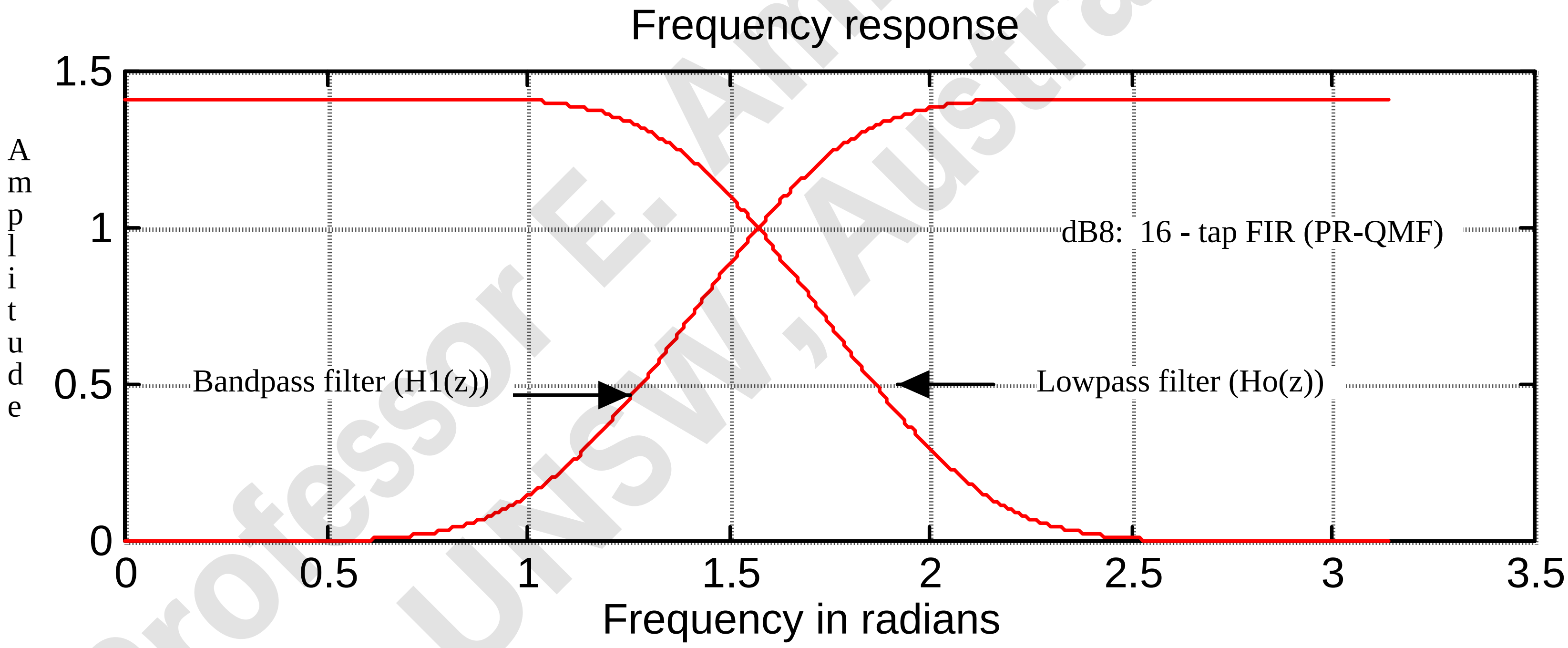
- The maximum signal energy and the masking threshold in each band can be calculated (see later on)
- The masking model output can be used to determine the bit allocation per subband for perceptually lossless quantisation.
- The samples are then scaled and quantised according to the subband bit allocation.

Wavelet Function

- For the Wavelet Packet Decomposition, an FIR Perfect Reconstruction-Quadrature Mirror Filters (PR-QMF) can be utilised.
- In this study, a 16-tap FIR lowpass filter derived from the Daubechies wavelet is used.
- Daubechies wavelet has the desirable regularity property as it generates a lowpass filter with transfer function $H_o(z)$ with the maximum number of N/2 zeros at $\omega = \pi$, where N is length of filter impulse response such that $|H_o(\theta)|$ is maximally flat.
- The diagram (see next slide) shows the magnitude response of the $\{H_o(z), H_1(z)\}$ QMF pair used as the basis of the decomposition filterbank.

Wavelet Function

- The magnitude response of the 16-tap lowpass filter based on the Daubechies wavelet (‘dB8’) provides an acceptable compromise between the subband separation and increased computational load.



Magnitude Response $H_o(z)$ and $H_1(z)$

- Although aliasing effects between neighbouring bands can be reduced by using filters with narrow transition bands, such effects will inevitably exist since any practical filters have to be of finite length.
- The length of the filter impulse response determine the width of the transition band which in turn specifies the overlap of the subband filter frequency responses.
- A longer filter impulse response results in a sharper transition between the subbands.
- However, any increase in the length of the filter impulse response is also accompanied by a corresponding increase in the computational load which therefore has to be weighted against the gain in coding efficiency due to narrower transition bands.

Implementation of the Auditory Masking Model

- Masking is the process where a number of least significant bits (LSBs) are removed from the binary representation of each sample which are deemed to be imperceptible by the auditory masking model.

- Identifying the LSBs that can be safely removed from the subband samples is a difficult task.

- However, it is possible to identify the imperceptible LSBs by calculating the masking threshold from the subband signal power.

- The auditory model used here determines only the noise masking properties of the subband signals.

Implementation of the Auditory Masking Model ....

- Implementation of tonal masking requires the detection of tonal components and the identification of the frequency and power of each tonal component.

- This, in turn, require a high resolution subband decomposition, causing a significant increase in the total computational effort.

- The auditory model used in this study is similar to the one used by Black and Zeytinoglu (1995).

- The steps involved in calculating the masking threshold per critical band are as follows:

- Calculate the maximum power per critical band (i.e. maximum squared coefficient in each band)

$$P(k) = 10\log_{10}(\max\{C_k(1)^2, C_k(2)^2, C_k(3)^2, \ldots\ldots C_k(L)^2,\})$$

where $C_k(1)$, $C_k(2)$, $C_k(3)$, ... $C_k(l)$ are WP coefficients in subband k and L is the number of coefficients per band.

- It is also possible to use power per critical band by calculating the average sum-square of the coefficients. Also using the maximum squared coefficient in each band would provide a sufficiently accurate measure of power in that band, whilst also lowering the complexity and computational load.

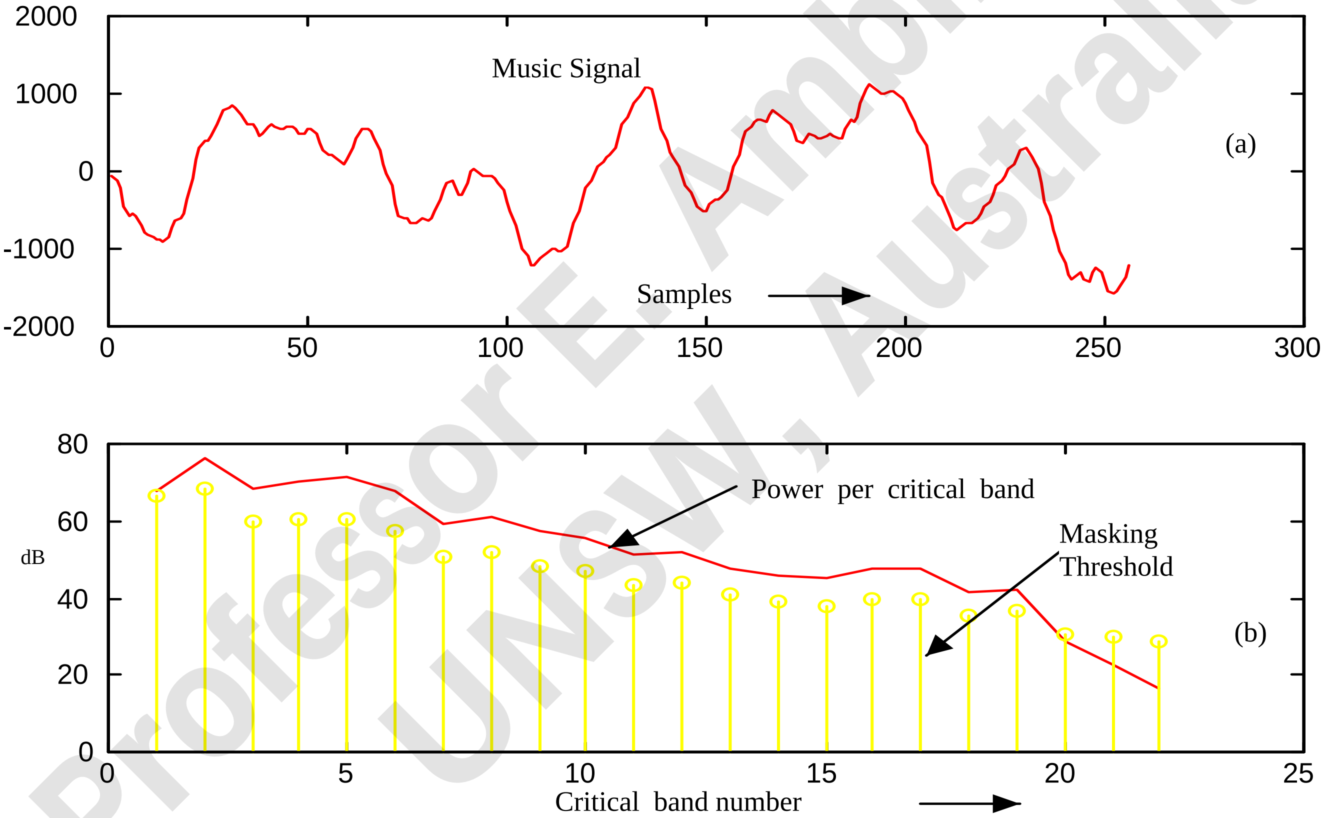
- Calculate the centre frequency in Barks.

- Identify the masker in a critical band and calculate the amount of masking it introduces other critical bands. This can be calculated using the piecewise linear approximation equation provided by Black(1995) for the masking shape of the masker at different power levels.

- Calculate the value of self masking (i.e. Spectral components within a critical band can be masked by other components within the same critical band.)

- Calculate the total masking level by summing the masking contribution from all the subband signal components.

Figure (a) below shows one frame of the music signal that was decomposed using WP decomposition . Figure (b) shows the maximum energy per critical band and the estimated global masking threshold for each critical band for the same frame of music signal sampled at 16 kHz.



Critical band energy levels and masking thresholds

Bit Allocation

- From the global masking thresholds the bit allocation per band is then determined.

- Figure (next slide) shows the parameters related to auditory masking.

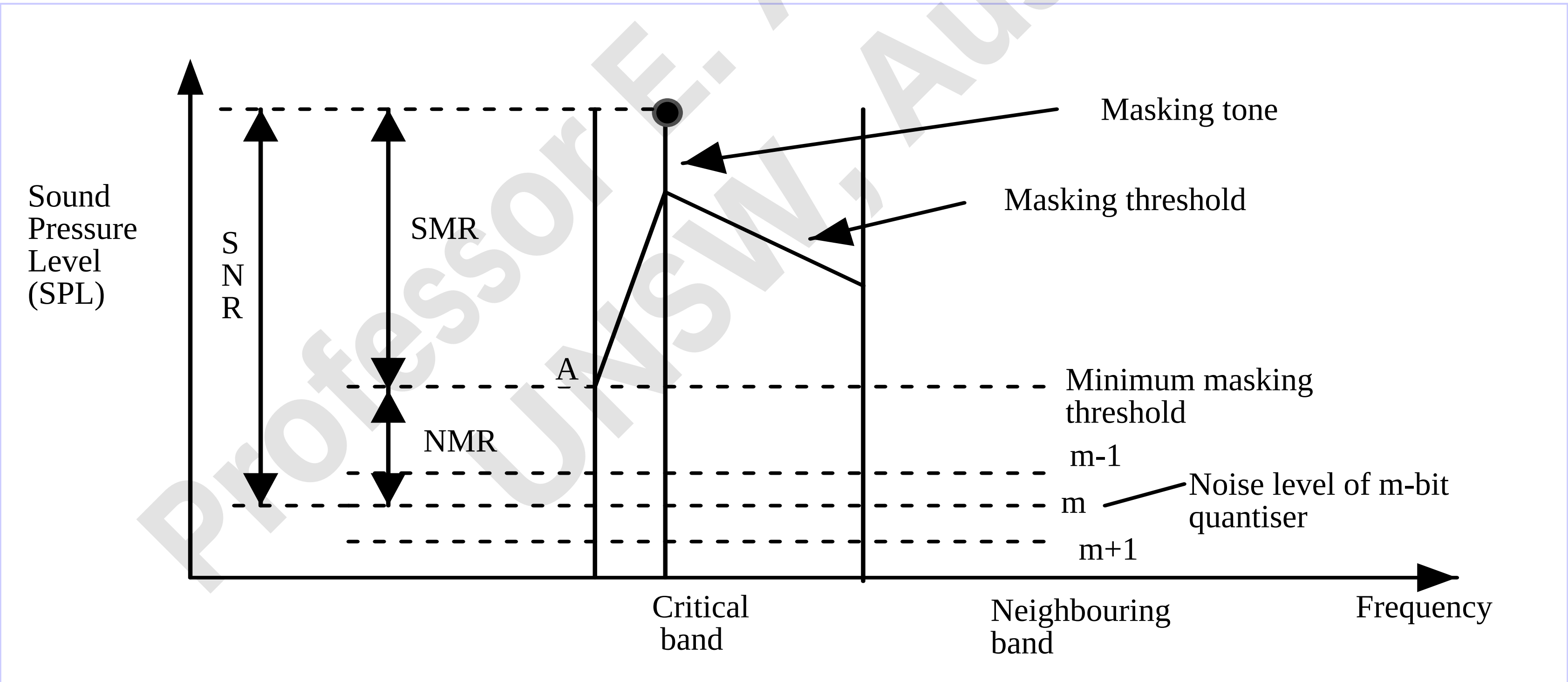
- The distance between the level of masker (shown as a tone in Figure) and the masking threshold is called Signal-to-Mask Ratio (SMR). Its maximum value is at the left border of the critical band (point A).

- Within a critical band, coding noise will not be audible as long as its SNR is higher than its SMR.

- Let SNR(m) be the signal-to-noise ratio resulting from m-bit quantisation, the perceivable distortion in a given subband is then measured by NMR(m) = SNR(m) -SMR

- NMR(m) describes the difference between the coding noise in a given subband and the level where a distortion may just become audible. The above discussion deals with masking by only one masker.

- If the source signal consists of many simultaneous maskers, a global masking threshold is calculated as discussed and the bit allocation can be determined by using the SMR.



Unconstrained number of bits to be allocated for each frame

- Firstly the number of bits per subband set to zero and the SMR for each band is calculated: {i.e signal power – auditory masking threshold}

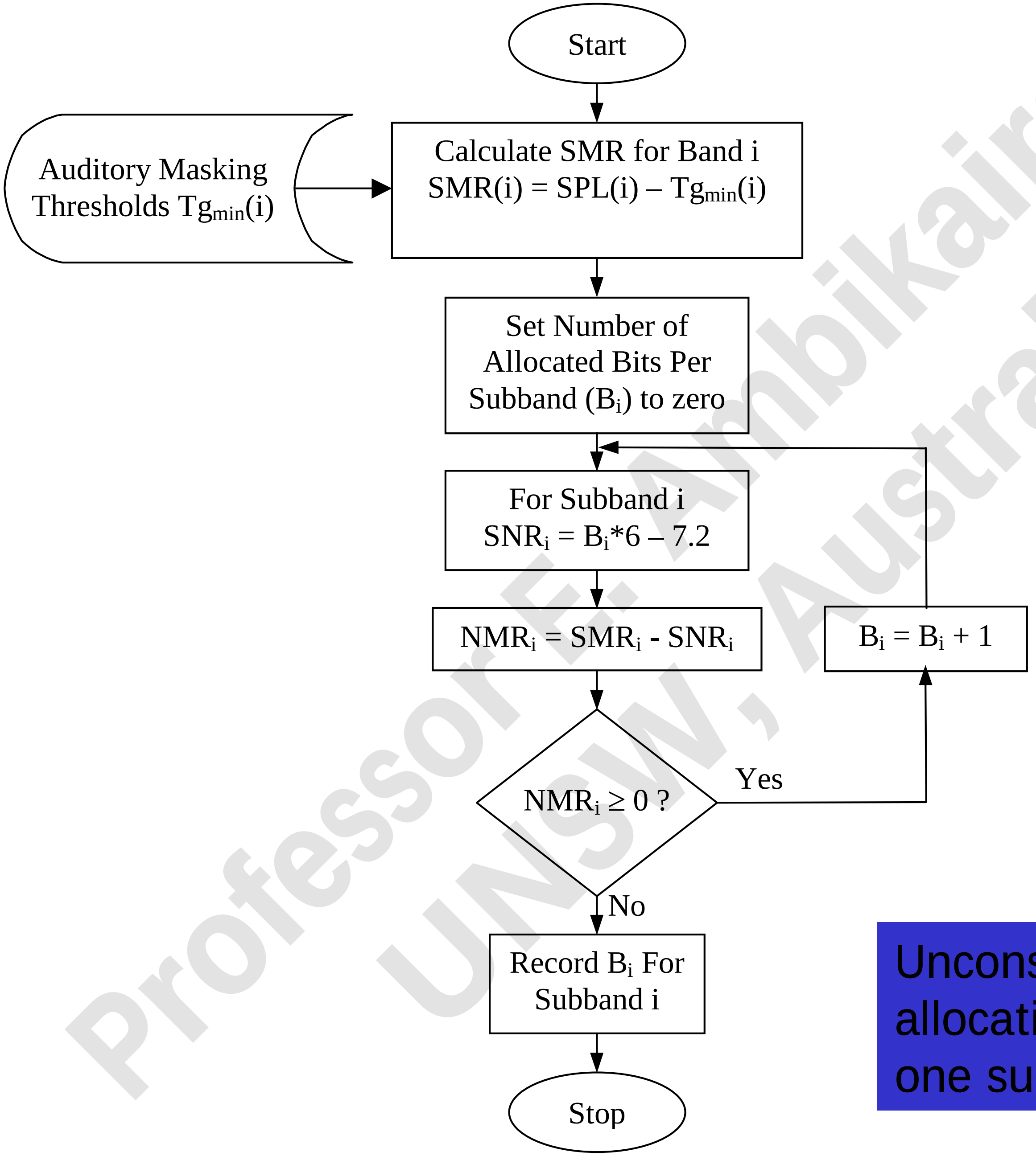
- Then for each subband the SNR is calculated by :

$$\text{SNR} = 6.02B - 7.2 \quad \text{dB}$$

- The NMR per band is then calculated as

$$\text{NMR} = \text{SMR} - \text{SNR}$$

- If the NMR for a band is greater than zero the number of bits allocated to that band is increased by one. This procedure is repeated until the NMR is zero, i.e. the quantisation noise is imperceptible.


Start
Auditory Masking Thresholds $Tg_{min}(i)$
Calculate SMR for Band i
$SMR(i) = SPL(i) – Tg_{min}(i)$
Set Number of Allocated Bits Per Subband (B_i) to zero
For Subband i
$SNR_i = B_i*6 – 7.2$
$NMR_i = SMR_i - SNR_i$
$B_i = B_i + 1$
$NMR_i \geq 0$?
Yes
No
Record B_i For Subband i
Stop

Unconstrained Bit allocation procedure for one subband

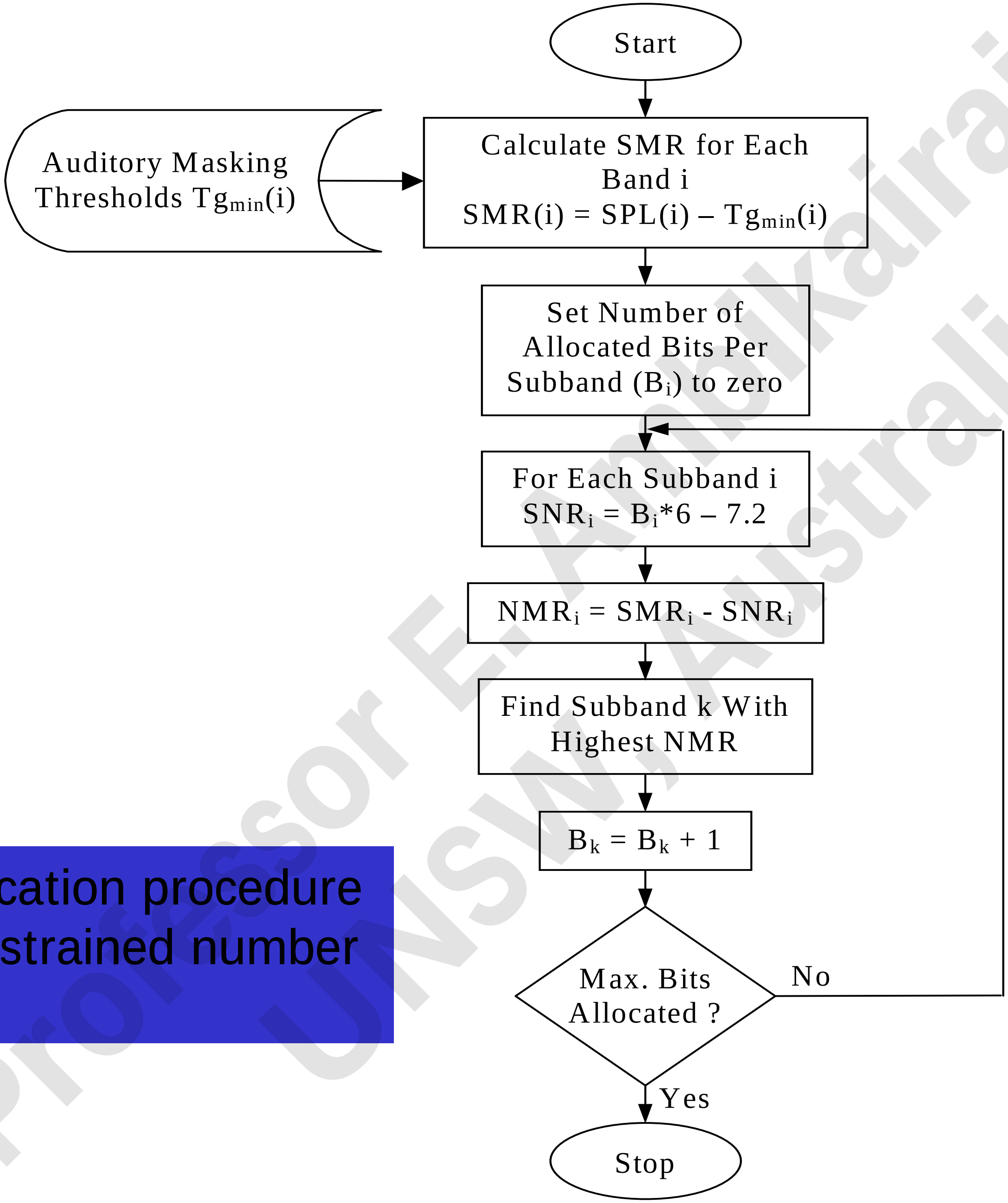
Bit Allocation procedure for constrained number of bits per frame

- For the allocation of a constrained number of bits the SMR for each band is again calculated and initial number of bits per subband set to zero as before.

- Then the subband with the highest NMR is found and an extra bit allocated to that band.

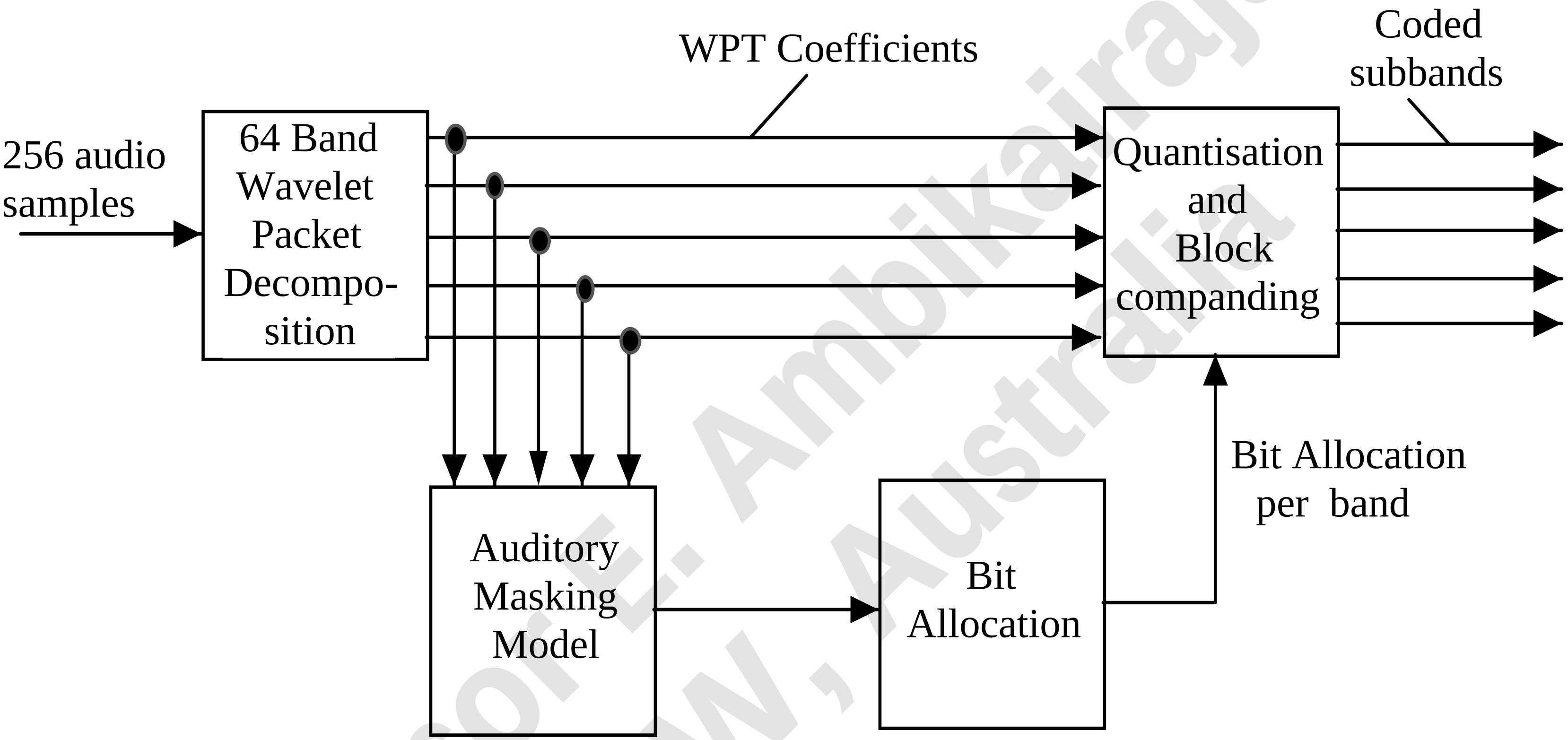
- This search and allocate procedure is repeated until the total number of bits allowed have been allocated.

- A flowchart for this procedure is given in the next slide.


Start
Calculate SMR for Each Band i
SMR(i) = SPL(i) – Tg$_{min}$(i)
Auditory Masking Thresholds Tg$_{min}$(i)
Set Number of Allocated Bits Per Subband (B$_i$) to zero
For Each Subband i
SNR$_i$ = B$_i$*6 – 7.2
NMR$_i$ = SMR$_i$ - SNR$_i$
Find Subband k With Highest NMR
B$_k$ = B$_k$ + 1
Max. Bits Allocated ?
No
Yes
Stop

Bit allocation procedure for constrained number of bits

Scaling and Quantisation



- Once the bit allocations per subband have been determined, the WP coefficients in each subband are scaled and quantised. Coefficients are scaled so that the maximum absolute value is one in each subband and the scalefactors are recorded for decoding.

- The scaling reduces the amount of bits required since the coefficients now only have to be quantised to a level in the range –1 to +1.
- Scaling is similar to block companding (See next few slides)

Block Companding

- In block companding the number of bits required to encode a subband block of samples can be reduced by removing redundant most significant bits (MSBs).

- For this description of block companding an assumption will be made that the samples of the signal in question are all positive.

- If the signal has been digitised using a uniform analogue-to-digital converter with a resolution of B bits, then there are 2^B quantisation levels available and the levels are 0, 1, 2,..., $2^B – 1$, i.e. $2^B – 1$ is the maximum amplitude available.

- If a sample is at the maximum value then bit B will be set to 1.

- For low amplitude samples one of the lower bit positions will be a leading 1 and all of the more significant bit positions will be 0.

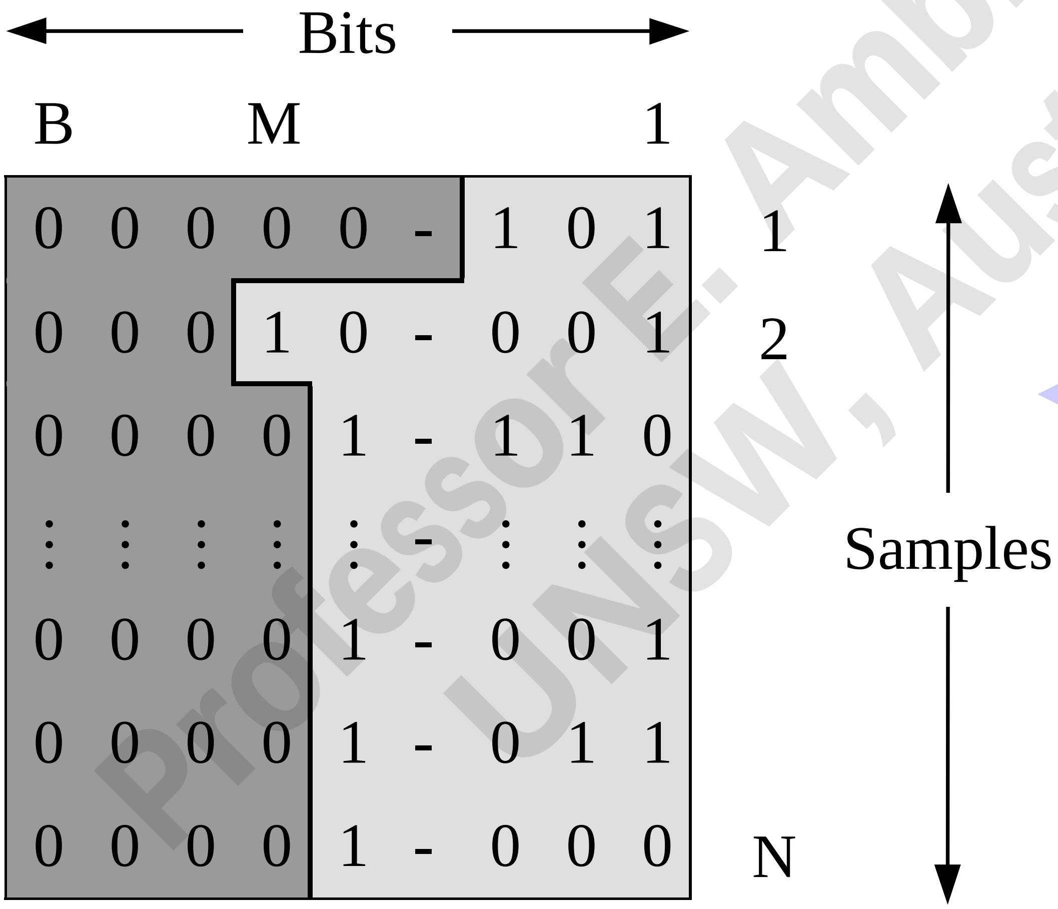
- These zeros can be removed (and only the lower bits stored) and be replaced without altering the signal, reducing the amount of storage space required for the sample.

- Block companding refers to the fact that the samples are grouped together into a block.

- Such a block would be a set of samples from the same subband.

- Companding a block, as opposed to each sample individually, reduces the amount of sideband information (i.e. the number of bits discarded) that has to be stored. Consider such a block of N samples with B bit resolution.

➢If the highest position of a leading “1” is bit M in the block, then we can discard bits M+1 to B before storage, and replace them later, without altering the signal stored. This process is indicated below:



B			M					1	
0	0	0	0	0	-	1	0	1	1
0	0	0	1	0	-	0	0	1	2
0	0	0	0	1	-	1	1	0	
⋮	⋮	⋮	⋮	⋮	-	⋮	⋮	⋮	
0	0	0	0	1	-	0	0	1	
0	0	0	0	1	-	0	1	1	
0	0	0	0	1	-	0	0	0	N

Block Companding

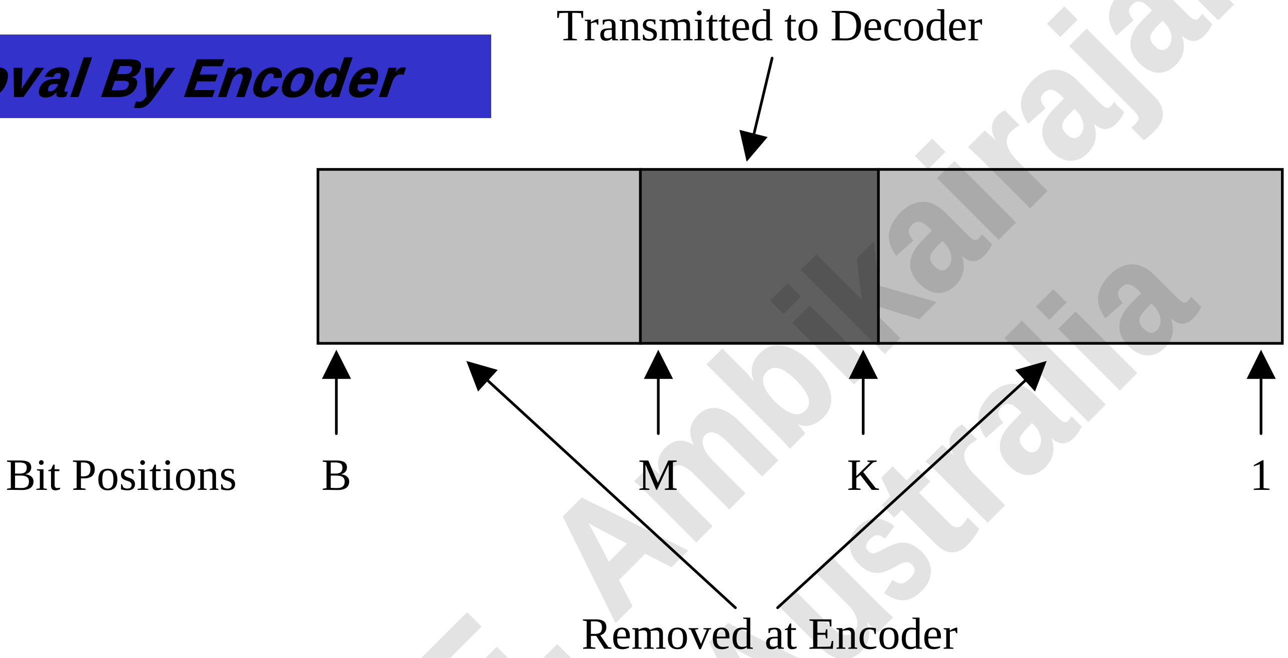
- As can be seen, the MSBs that are shaded dark are all zero and so can be discarded.
- However, due to the position of the leading “1” in sample 2, M bits are required for each sample in the block.
- So for this block a total of N × M bits are required for storage, a saving of N(B - M) bits.
- For each block the number M also has to be stored in order for the decoder to reconstruct the companded block.
- The decoder will place M leading 1s or 0s in front of each sample, depending on the sign.

- This data is part of the *sideband* information that has to be stored along with the data itself.

Quantisation by Masking of Least Significant Bits

- To consider the masking by least significant bit (LSB) removal, consider a sample from a subband that has an allocation of L bits per sample.
- If M bits remain after block companding, then only bits K to M must be stored, where K=M-L.
- This is shown in the next slide for a sample with B bits originally.

Bit Removal By Encoder



As can be seen the encoder only needs to transmit bits K to M, which are shaded in dark grey. All remaining bits can be discarded. At the decoder the missing MSBs and LSBs are replaced either by 1s or 0s depending on the sign of the sample.

Note that the number of bits per sample for each subband must also be stored as part of the sideband information.

Results

- The audio coder described in this chapter was implemented in Matlab on several short pieces of music.
- Almost transparent coding was achieved with an average of 3 to 4 bits per sample with unconstrained bit allocation.
- Experimental data shows that the coder operates well, significantly reducing the bit rate of the signal with little perceptible distortion introduced.
- The coder performs almost equally well for several types of music, with approximately the same bit rate required.
- Due to the nature of the WP tree used for the audio coder it can be adapted to operate at most of the industrial sampling rates which is another important feature for a real time audio coder i.e. it is *scalable*.