EEE116 Multimedia Systems

Topic 5 – Digital Audio

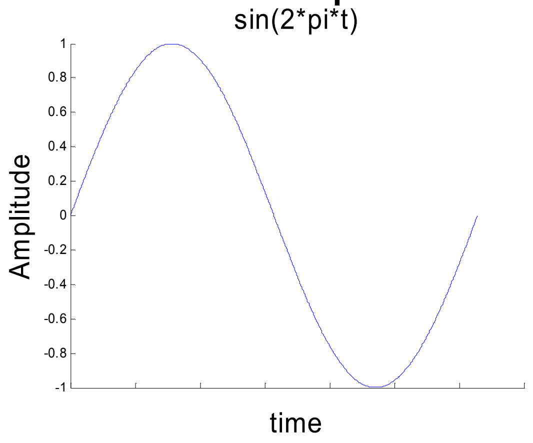
- Audio Digitisation
  - Sampling, quantisation
  - data rates & audio quality.
- Speech
  - Linear predictive coding
- Perceptual audio coding
  - Temporal masking
  - Frequency masking
- Audio coding standards.
  - MPEG family of codecs (of course this includes MP3)

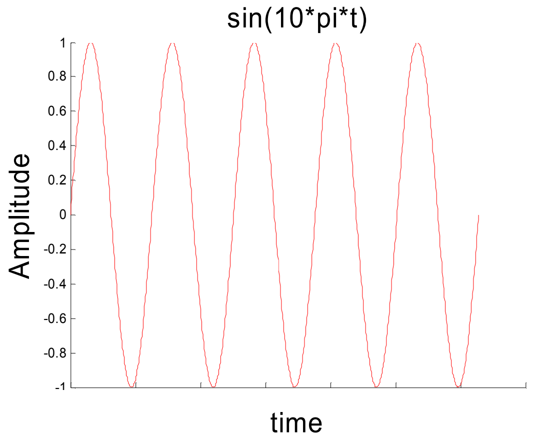
Dr. Charith Abhayaratne
c.abhayaratne@sheffield.ac.uk

Signals, Frequency, Bandwidth and Noise

- Frequency - A measure of the rate of change of a signal.

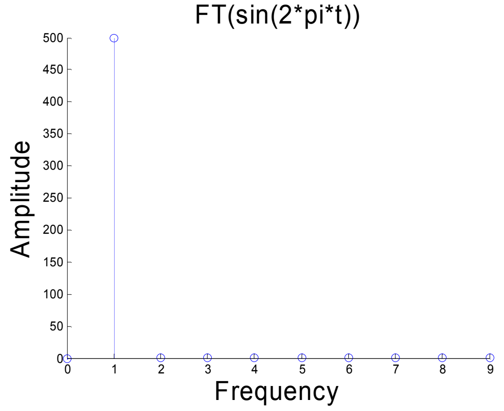
- Time domain representation

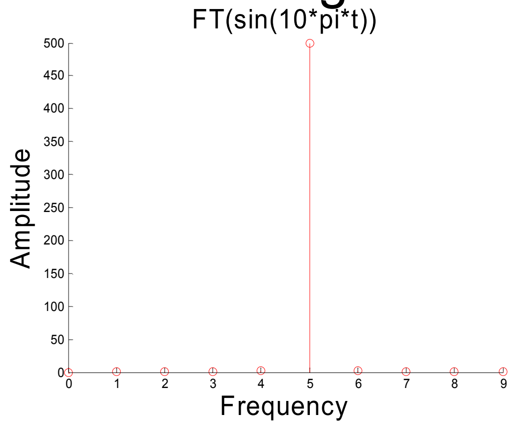


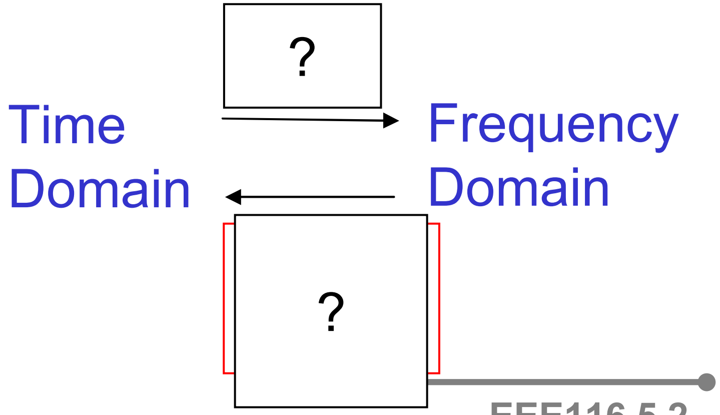


Frequency ? ? ?

- Frequency domain representation – using the Fourier transform (FT)







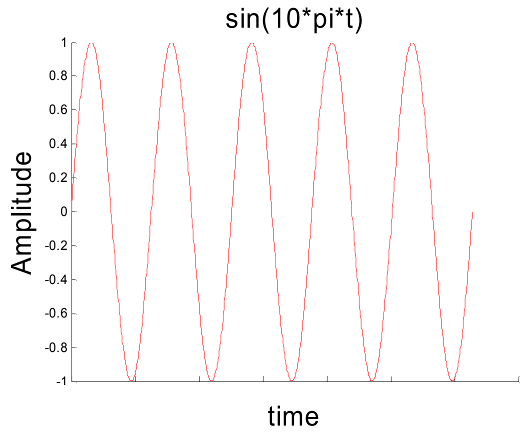
?

Time Domain

Frequency Domain

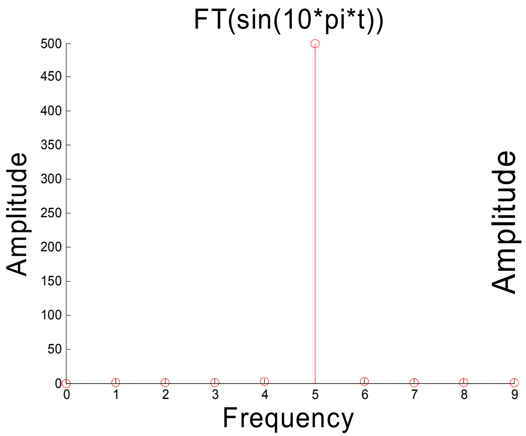
?

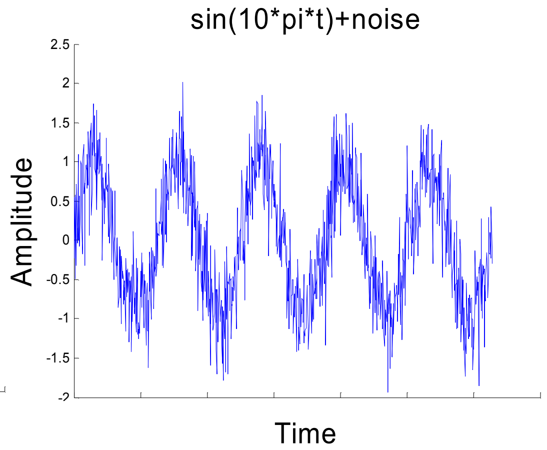
Signals, Frequency, Bandwidth and Noise



Original signal

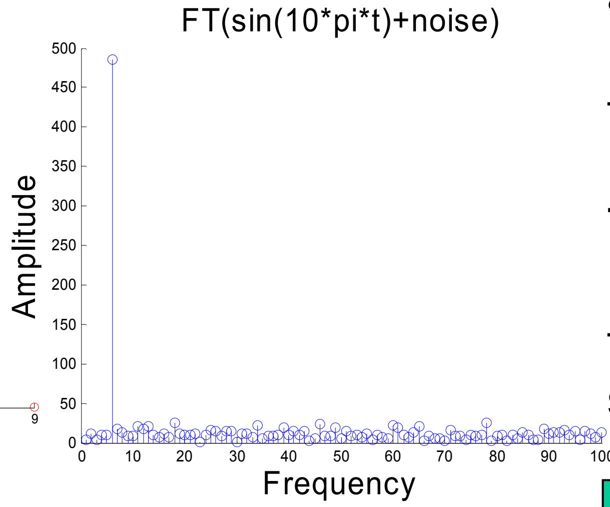
↓ FT





Noisy signal

↓ FT

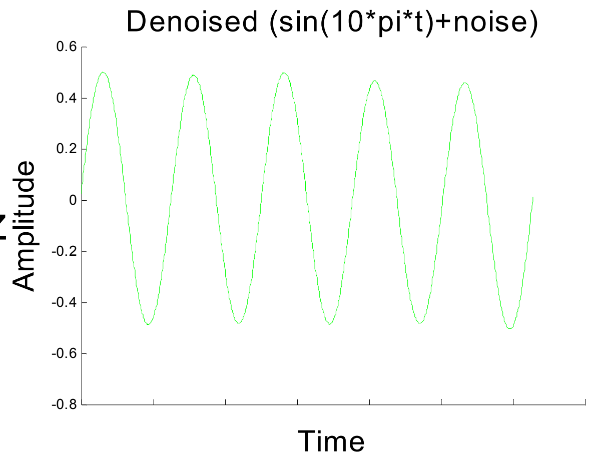


Noise adds high frequency components. The original signal contains only 5Hz in this case.

A low pass filter is used to keep frequencies only up to a specified frequency and to attenuate the higher frequencies.

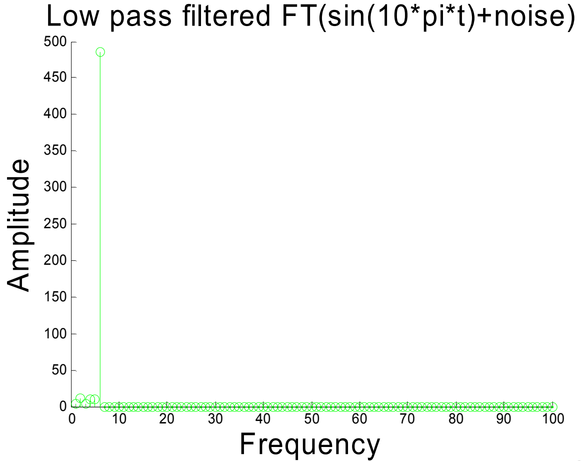
The range of important frequencies in a signal is called Bandwidth. E.g.

?



Denoised signal

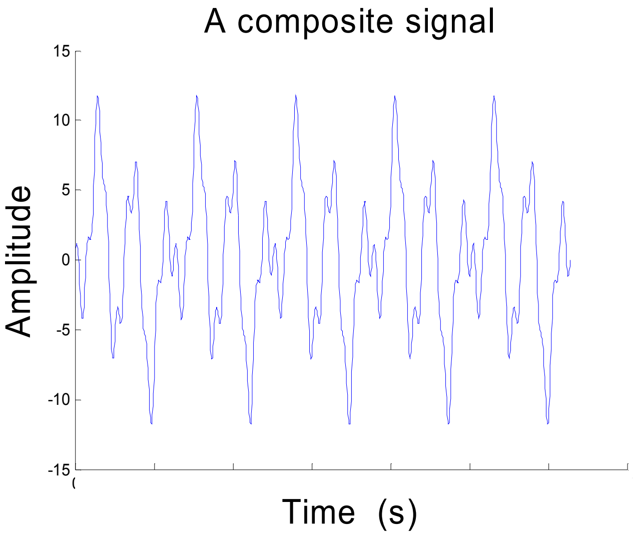
↑ Inverse FT

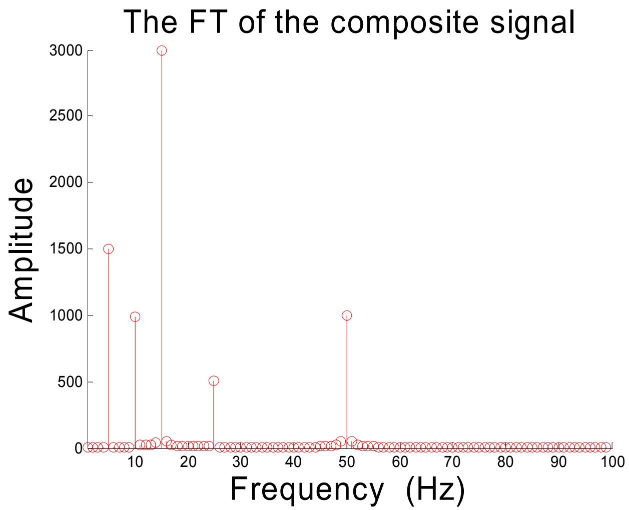


Signals, Frequency, Bandwidth and Noise

Example:

A composite signal in time domain and in frequency domain are shown as follows:





There are large amplitude components at frequencies 5Hz, 10Hz, 15Hz, 25Hz, 50Hz.

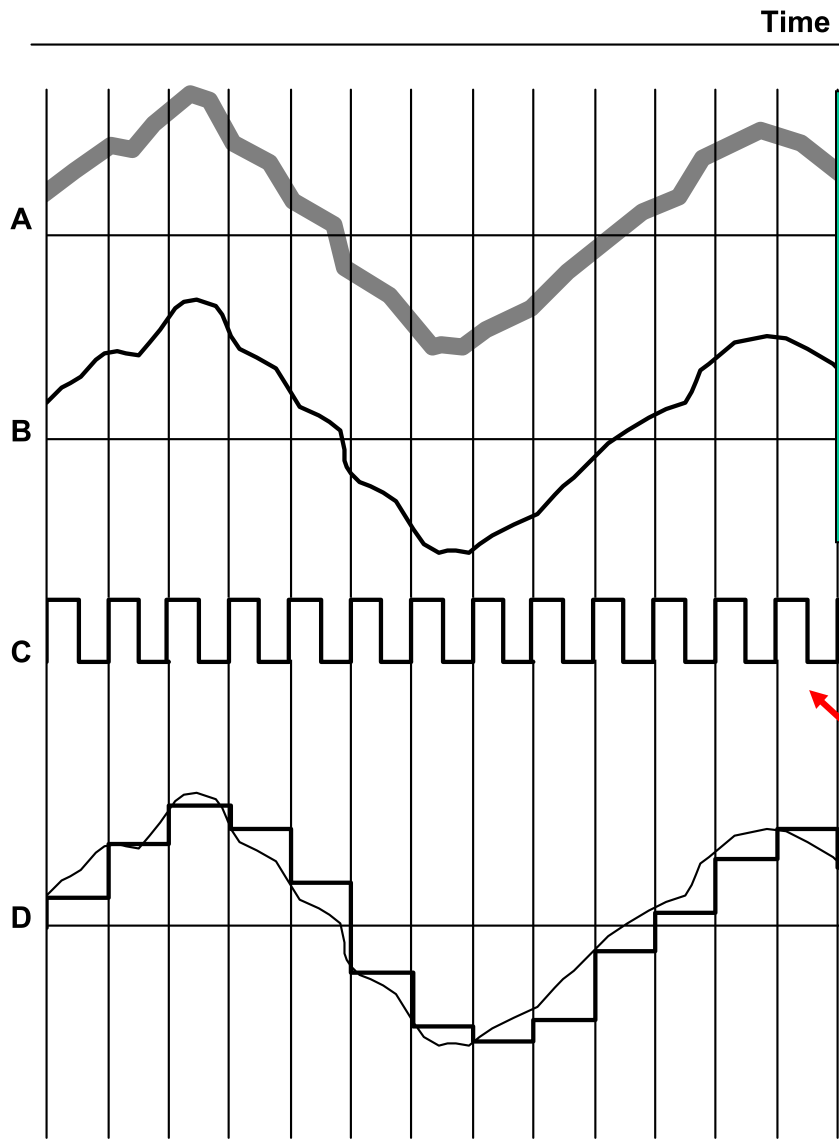
What is the signal bandwidth? ?

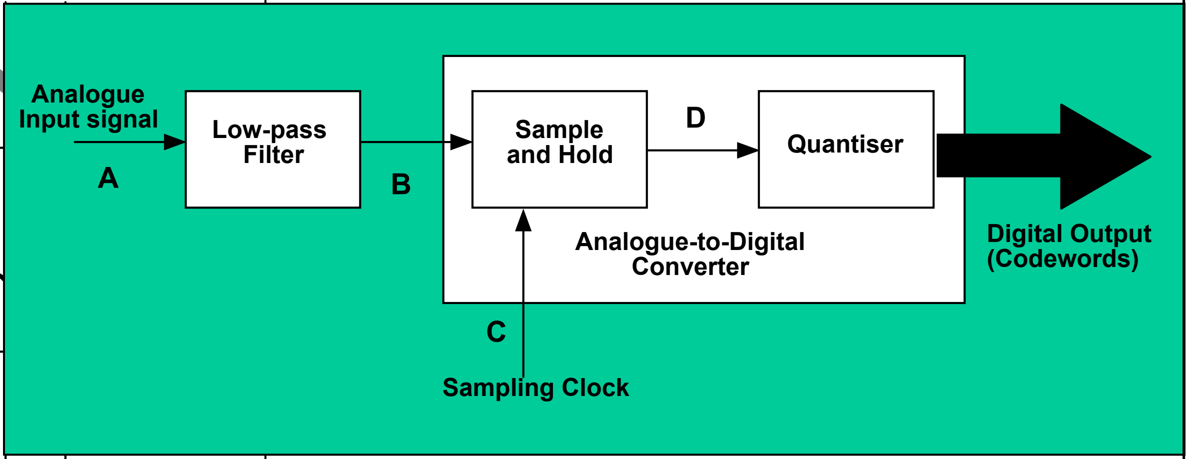
What is the bandwidth (or cut-off frequency) of the low pass filter required to denoise this signal (i.e., to remove unimportant high frequency components) ? ?

What is the suitable sampling frequency for this signal? ? ? z

Audio Signal Digitisation

Signal Encoder (digitizer)



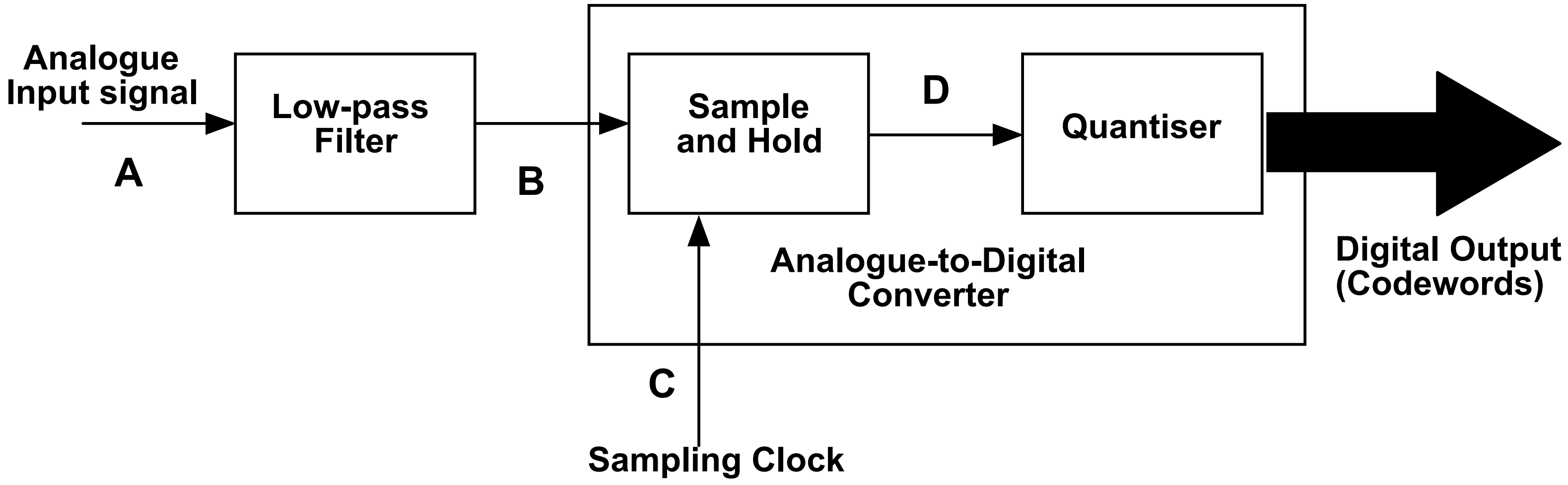


All signal encoders consist of low-pass anti-aliasing filter and an analogue-to-digital converter (ADC)

What is the lowest frequency that can be used of this sampling clock?

?

Signal Encoder



Low-pass Filter: passes low frequencies and attenuates high frequencies. Reduces noise and helps prevent aliasing. (Limits the signal to its maximum usable frequency component).

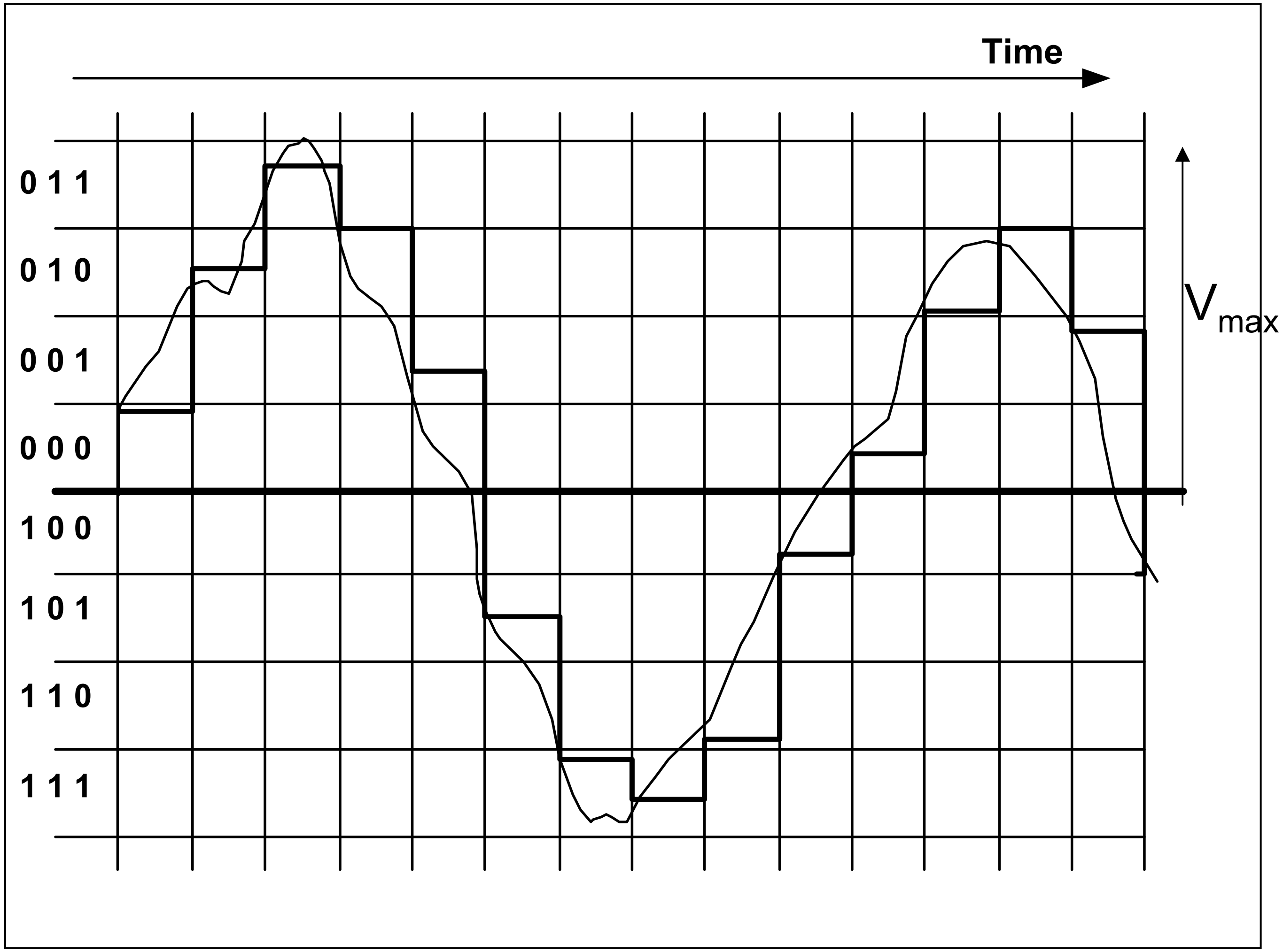
Sample and Hold: samples instantaneous amplitude of signal and holds this value until the ADC has converted value to digital representation.

Sampling Clock: sets sampling rate : minimum rate = Nyquist rate

Quantiser: converts analogue value to digital representation.

This is pulse coded modulation (PCM) as the magnitude values are encoded.

Revision from Topic 2



Output = 🟩🟩🟩🟩🟩🟩🟩 etc

Quantisation interval = d
Different symbols =c
c= $[(V_{max}-V_{min})/d] +1$

Using N-bits we can represent 2^N symbols.

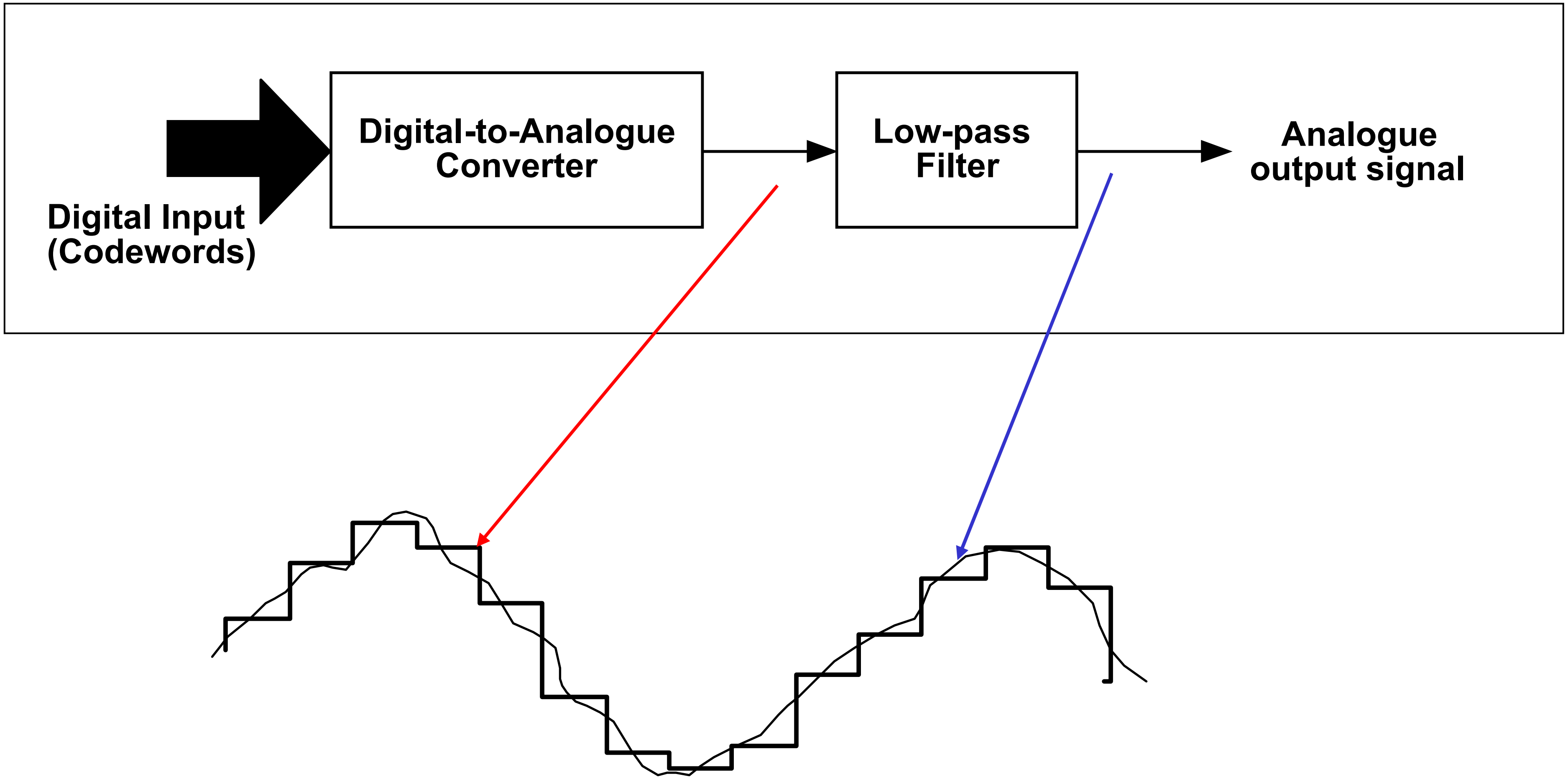
$c=2^N$

How to compute “d” for an N-bit representation?

Quantisation causes an error (difference between actual value and digitised one)

What is the maximum error?

Signal Decoder



Audio Content

Two types of audio signals:

Speech:
Typical bandwidth: 50 Hz - 10 kHz
Quantisation: 12 bits

What is the bit rate?

?

Music
Typical bandwidth: 15 Hz - 20 kHz
Quantisation: 16 bits

What is the bit rate?

?

Audio digitisation

- Most common sample rate is 44.1 k samples/sec.
- 16-bit linear quantisation
- 2 channels for stereo
- Hence total bit rate = ?

🔊 44.1 kHz sampling at 16 bits - CD quality

🔊 44.1 kHz sampling at 8 bits -

🔊 22 kHz sampling at 8 bits –

🔊 11 kHz sampling at 8 bits –

Audio Compression

Example: What is the file size of a 3 minute song with CD-quality stereo recorded using 44.1 kHz sampled at 16 bits ?

? ?

How to reduce the data rate?

1. Change sampling parameters:
    1. Reduce Bandwidth -- ?
    2. Reduce the sample size -- ?
    3. Reduce number of channels – ?

2. Pulse Coded Modulation (PCM): (just sampling & quantisation)
    1. PCM
    2. DPCM and Adaptive DPCM (Data modelling techniques)

3. Advanced Compression Schemes
    1. For speech encoding - model human speech production
       Linear Predictive Coding (LPC)
    2. For general audio encoding - model human hearing

PCM Speech

Pulse Code Modulation

For **Public Switched Telephone Networks** (PSTN) – the basis of all modern digital phone systems
- Uses 8 bits per sample and 8 kHz sampling rate.

This gives 64 kbps. Standard ITU-T G.711

Analogue speech bandwidth is 200 Hz - 3.4 kHz

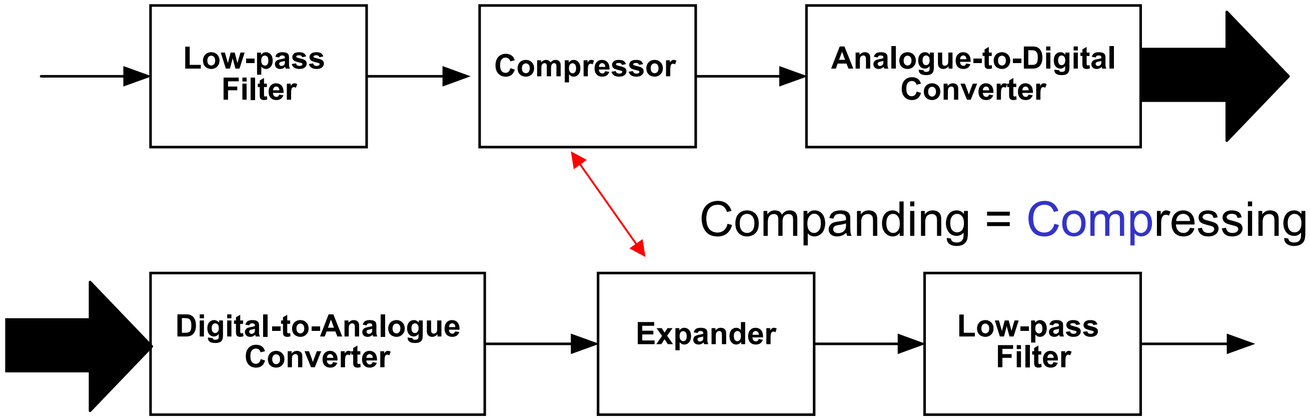
Because limited to 8 bits, We are more sensitive to background noise during quiet passages rather than in loud passages

Solution is using non-uniform bit allocation – i.e., non-uniform quantisation.

i.e., expand number of bits for low amplitude and reduce at higher amplitudes.

This is achieved by “μ -law” correction.

Non-uniform quantisation = multiply with a non-linear function followed by uniform quantisation



Companding = Compressing + expanding process

Effectively compresses the signal range, by allotting fewer bits at the high amplitudes. The exact details need not concern us - called "μ -law" correction

Results in non-uniform quantisation as opposed to uniform quantisation in PCM.

Differential Pulse Code Modulation (DPCM)

Difference in amplitude of successive sound samples is less than actual amplitudes, so encode differences - fewer bits required.

On average typical savings from an order-1 DPCM are limited to 1 bit
– hence 64 kps –> 56 kps.

Accuracy of each computed difference signal (called the residual signal) depends on accuracy of previous value decoded.

The A/D conversion errors - quantisation errors - can accumulate.

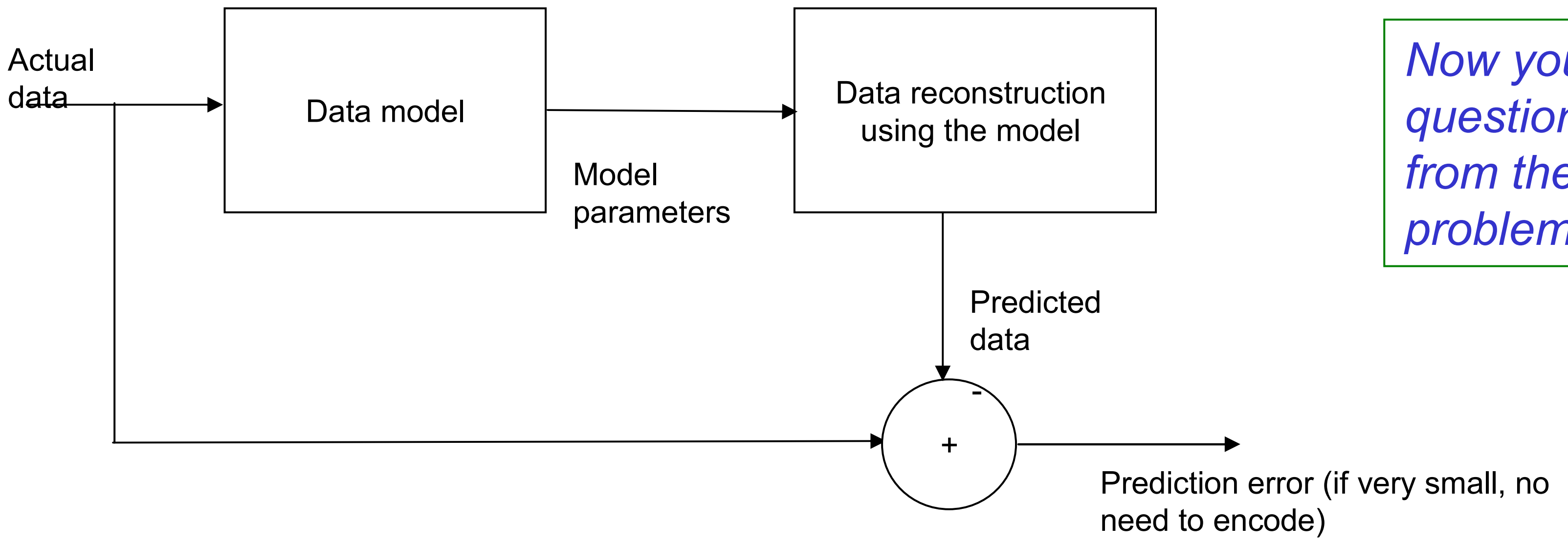
Hence use a more complex scheme - where (say) three sample differences are held in different registers (buffers) and proportions of each one make up the residual signal. This is the order-3 DPCM. Can get to 32 kbps

Further improvement using fewer bits to encode smaller difference signals - called Adaptive Differential PCM - can get to 16 kbps

- So far considered speech encoders that either transmitted digitised samples directly (e.g., PCM) or differences between samples (e.g., DPCM).

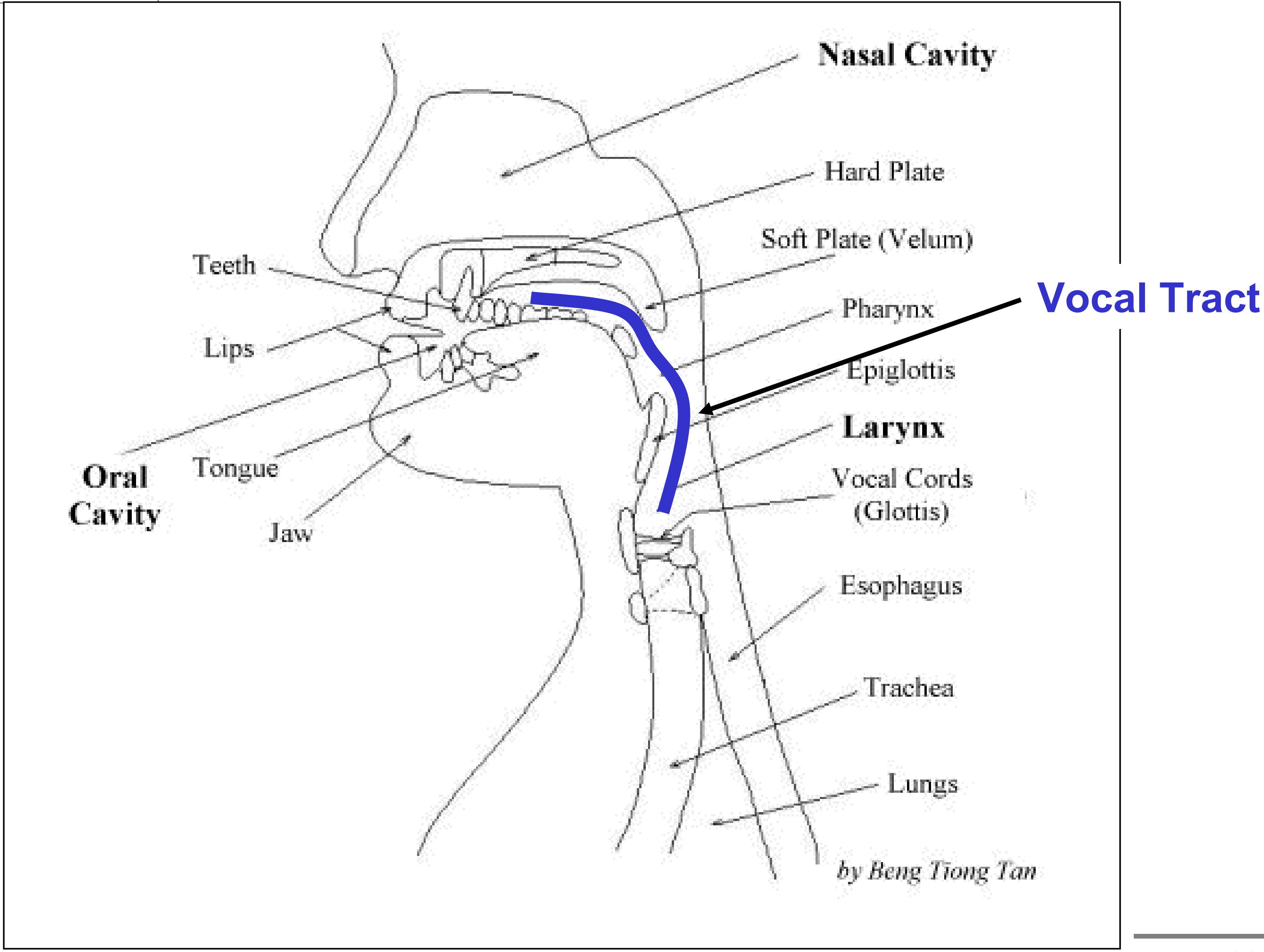
- Possible to achieve much greater compression if determine the perceptually important parameters of the speech and transmit these. Then synthesise speech at the decoder using these parameters. (Revisit: Data modelling in Topic 4)

- This is the basis of **Linear Predictive Coding**. The speech can sound a little synthetic but high compression is possible – basis of all current low bit-rate coding.


Actual data
Data model
Model parameters
Data reconstruction using the model
Predicted data
-
+
Prediction error (if very small, no need to encode)

*Now you can attempt questions **Q9 and Q10** from the tutorial problem sheet.*

Human Vocal System



When you speak:

- Air is pushed from the lungs through the vocal tract and out of the mouth.

- For certain *voiced* sounds, the vocal cords vibrate (open and close). The rate at which the vocal cords vibrate determines the *pitch* of the voice. Women and young children tend to have higher pitch (faster vibrations) while adult males tend to have lower pitch (slower vibrations).

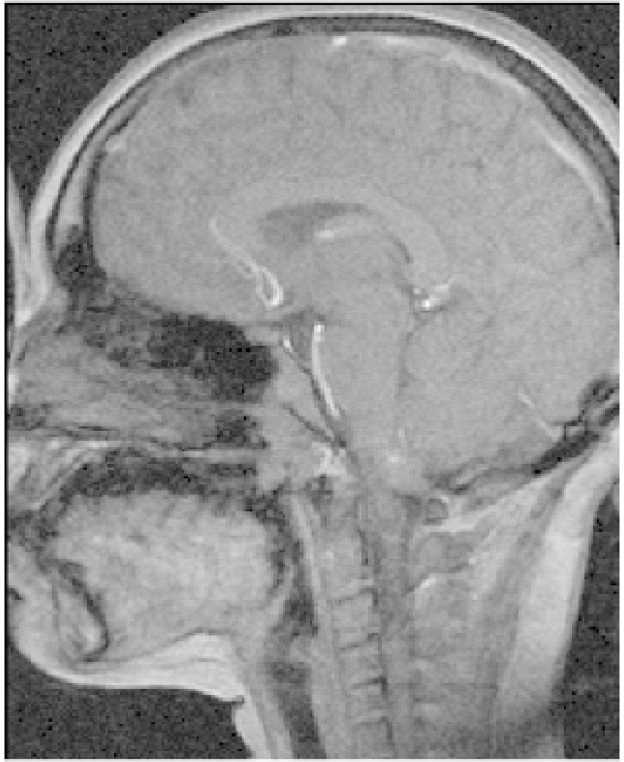
- For certain *fricatives and plosive (or unvoiced)* sound, the vocal cords do not vibrate but remain constantly open.

- The shape of the vocal tract determines the sound that is made.

- As you speak, the vocal tract changes its shape producing different sounds.
- The shape of the vocal tract changes relatively slowly (on the scale of 10 ms to 100 ms).

- The quantity of air coming from the lungs determines the loudness of your voice.

Magnetic Resonance Imaging Video



Greg Kochanski (University of Oxford Centre for Clinical Magnetic Resonance Research)

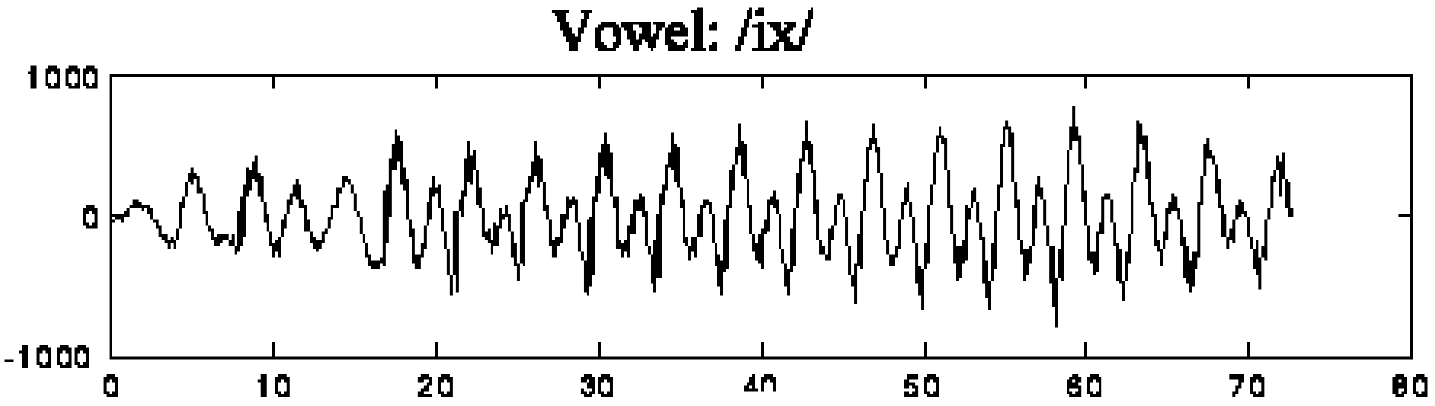
Repeated utterances of "I'm a spotted chicken".

Voiced sounds are produced by forcing the air stream through vocal cords while vocal cords are forced to open and close rapidly to produce a series of periodic puffs which has a ***fundamental frequency -*** same as the vocal cord vibration frequency.

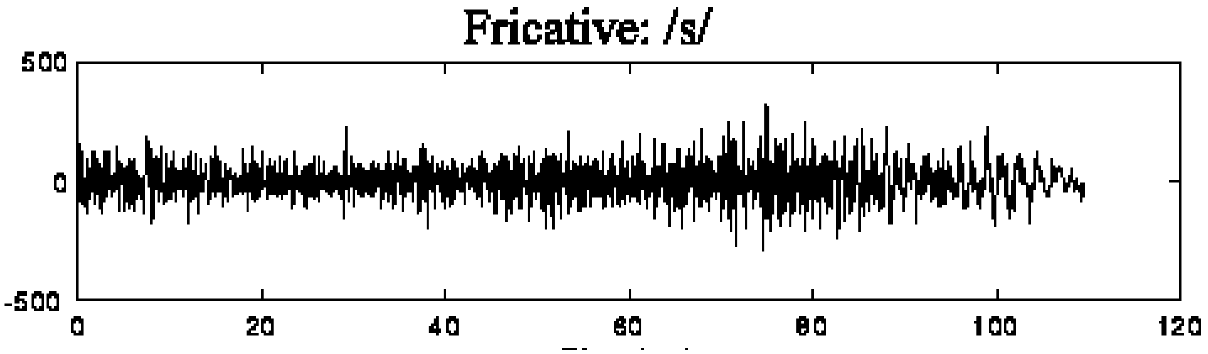
The vocal cord frequency is dependent on the massiveness, tension and length of the vocal cords and the effect of low air pressure created in the glottis - a space between the vocal cords.

Any sound which is produced without vibration of the vocal cord is called ***unvoiced***.

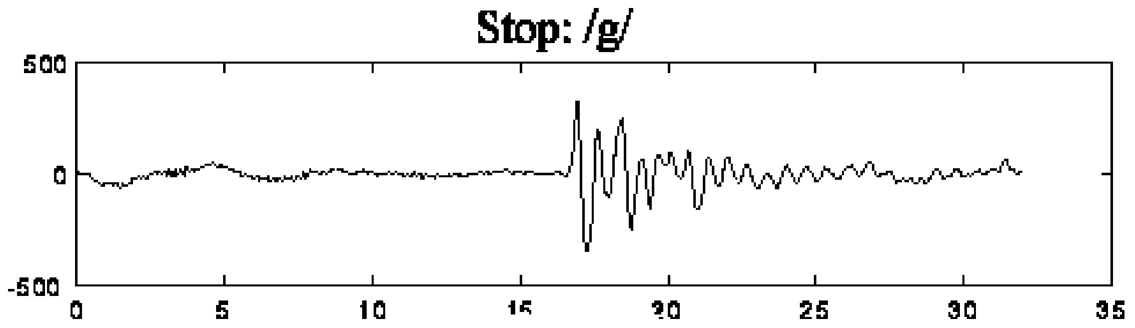
The figure shows a segment of vowel /ix/. The ***quasi-periodicity*** (almost periodic) of voiced speech can be observed.



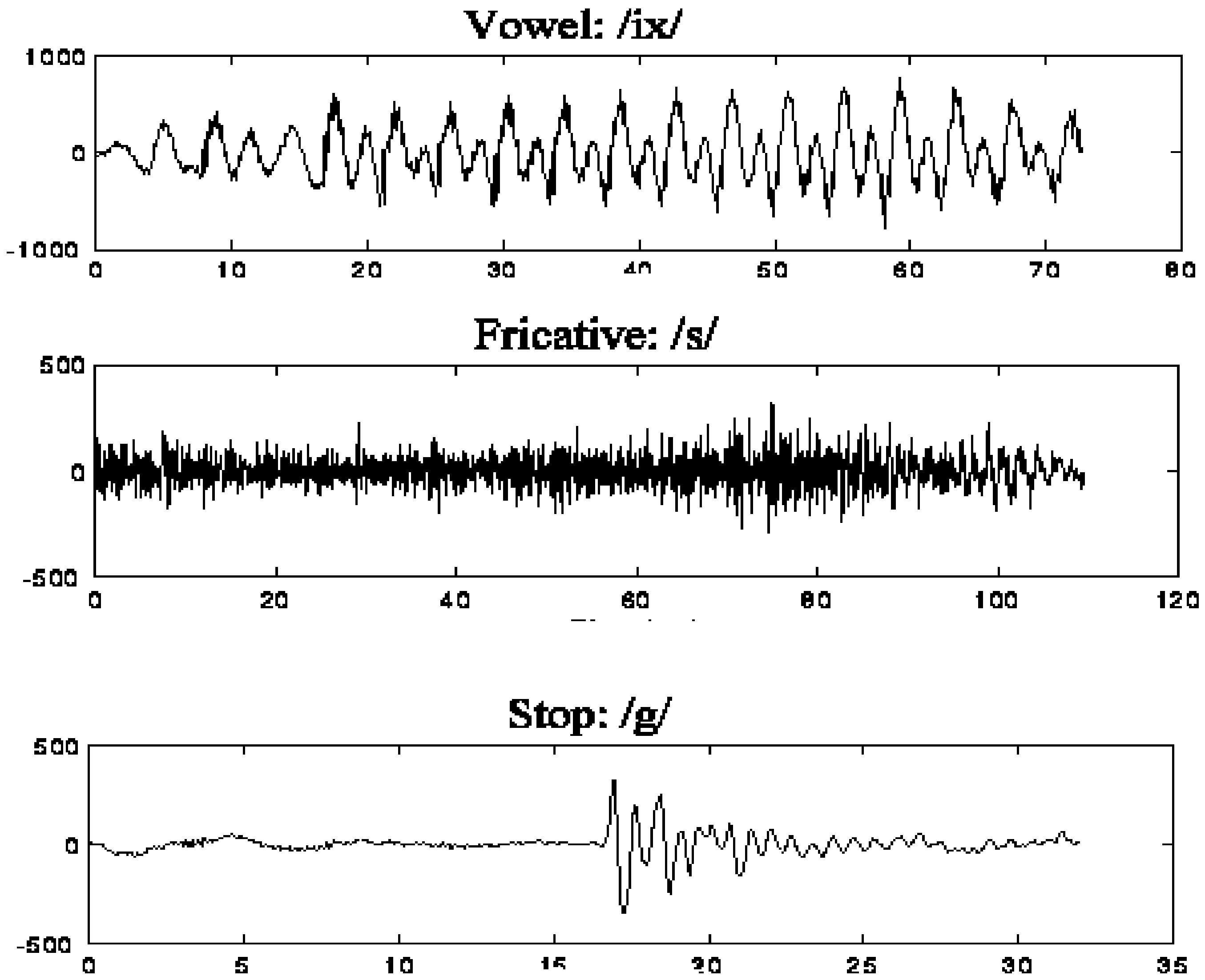
Fricative sounds are generated by constricting the vocal tract at some point along the vocal tract and forcing the air stream to flow through at a high enough velocity to produce turbulence. This turbulent air sounds like a hiss e.g. /hh/ or /s/.


Fricative: /s/

Plosive or ***stop*** sounds are resulted from blocking the vocal tract by closing the lips and nasal cavity, allowing the air pressure to build up behind the closure, and following by a sudden release of it. This mechanism produces sounds like /p/ and /g/. The following figure shows the stop /g/. The silence before the burst is the stop closure.


Stop: /g/

What parameters can we estimate?



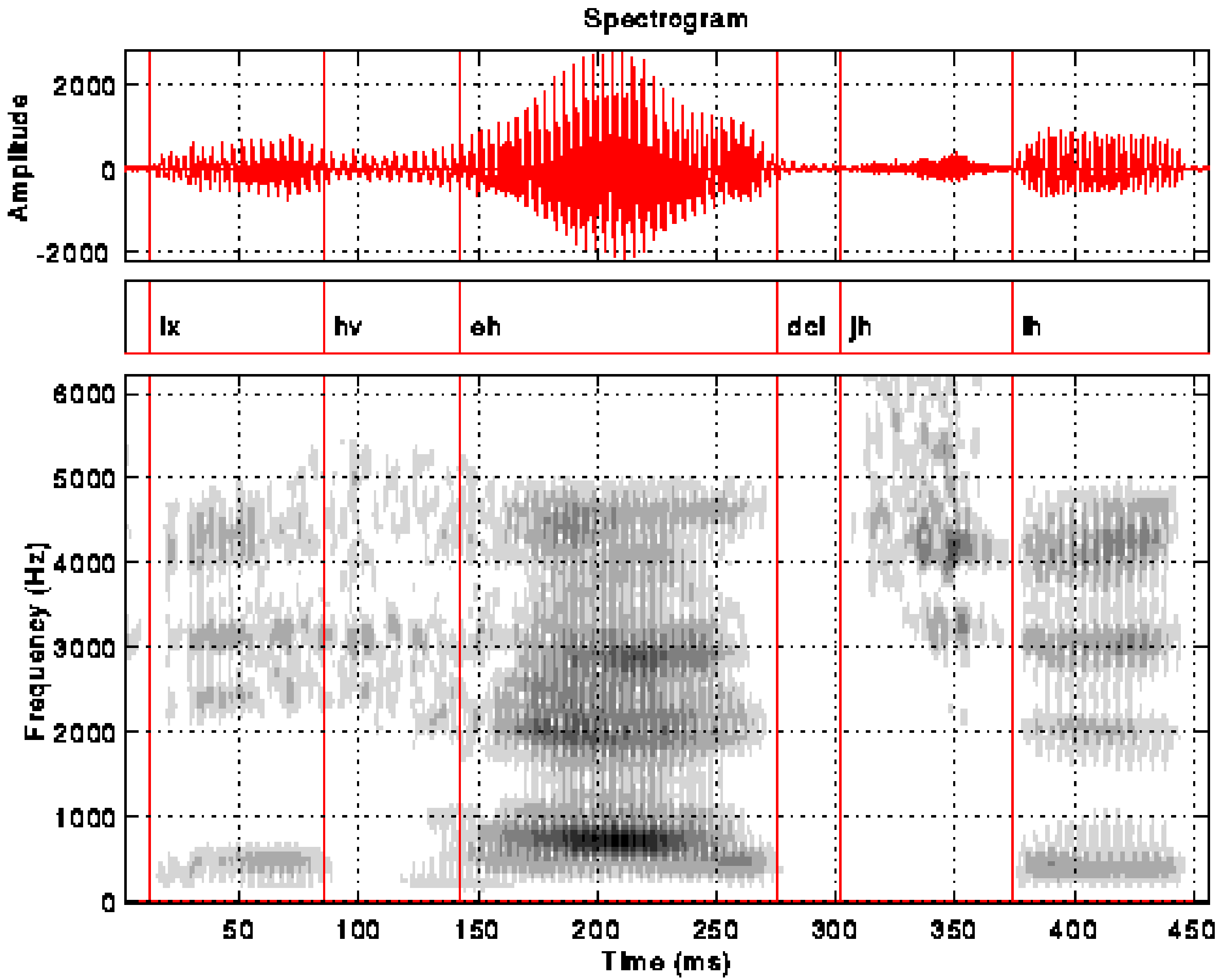
Formant frequency is a resonance frequency of the vocal tract. The formant frequencies of a vowel are determined by the parameters of the vocal tract configuration such as

the length of vocal tract,
the position of tongue, and
the shape of lips.

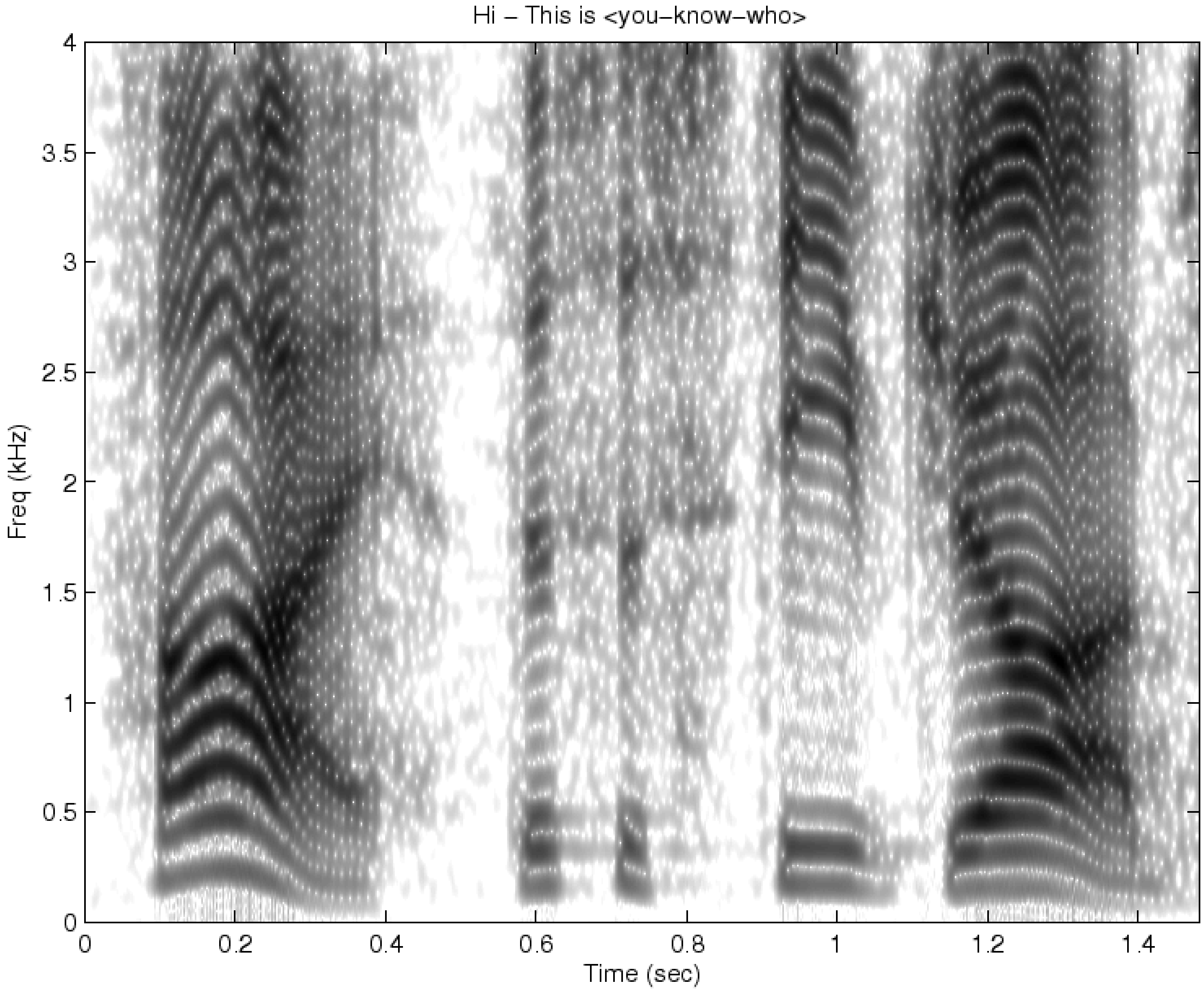
The first three formants are the primary cues in recognising English vowels.

For unvoiced sounds, such as /hv/ and /jh/ in a spectrogram, the primary cues for recognition is the energy distribution along the frequency axis.

Speech Spectrogram



Speech Spectrogram

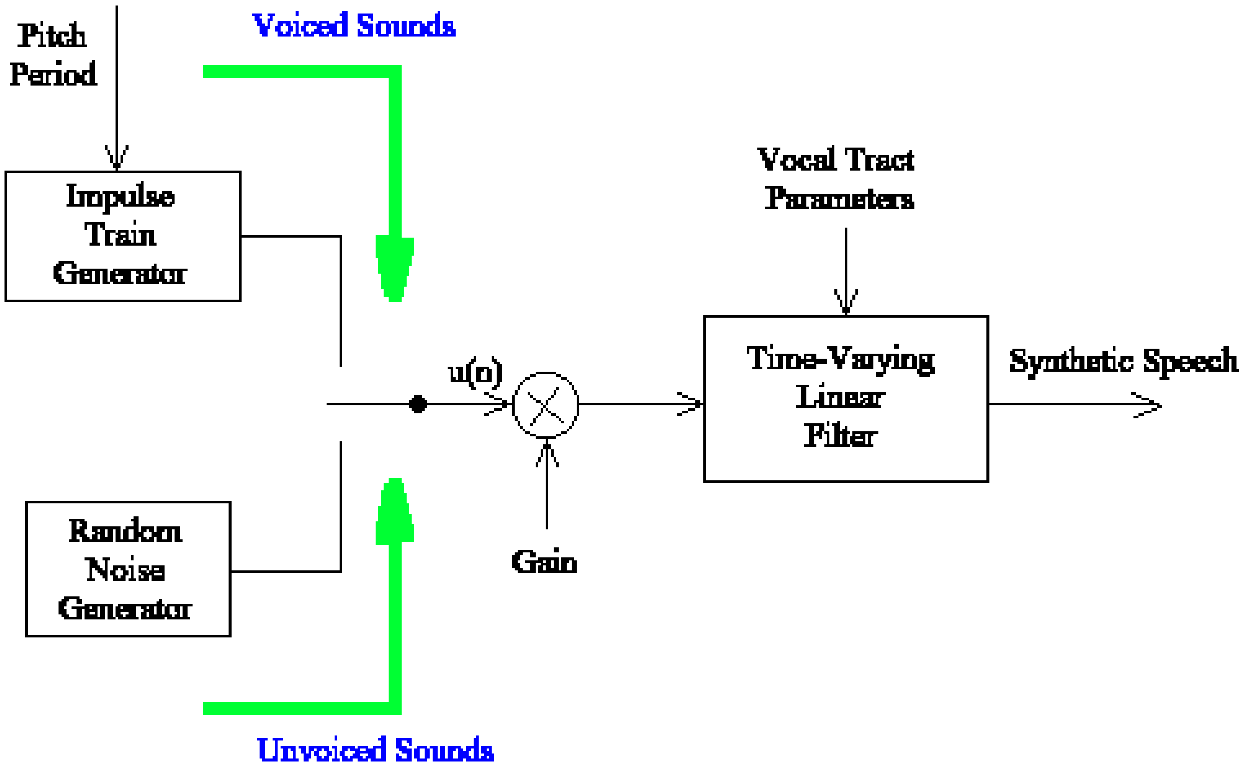


Linear Predictive Coding (LPC) Model

All vocal tract parameters are represented by a set of LPC coefficients, which are calculated automatically from the natural speech signals.

Number of coefficients is typically 10 to 20, i.e. 2 for each formant in the 4-5 kHz bandwidth of the speech signal,

plus a few others - e.g., the source of vocal tract excitation.



Speech Synthesis model based on LPC model

Example

"A lathe is a big tool. Grab every dish of sugar."

Original (64 kbps) This is original speech signal sampled at 8,000 samples/second and μ -law quantized at 8 bits/sample. Approximately 4 seconds of speech

LPC10 (2400 bps) This is speech compressed using a Linear Predictive Coding (LPC10) scheme.

The effective bit resolution is $\boxed{?}$ bits/sample

(I.e., a compression ratio of $\boxed{?}$)

🔊 64 kbps - μ-law corrected

🔊 32 kbps - Adaptive differential PCM

🔊 4.8 kbps - Linear Predictive Coding

🔊 2.4 kbps - Linear Predictive Coding

🔊 1.4 kbps - Linear Predictive Coding

- **For speech encoding -** model human speech production
- **For general audio encoding -** model human hearing

- It is often difficult to hear one sound when a much louder sound is present. This effect seems intuitive, but at the psychoacoustic and cognitive levels it becomes very complex.

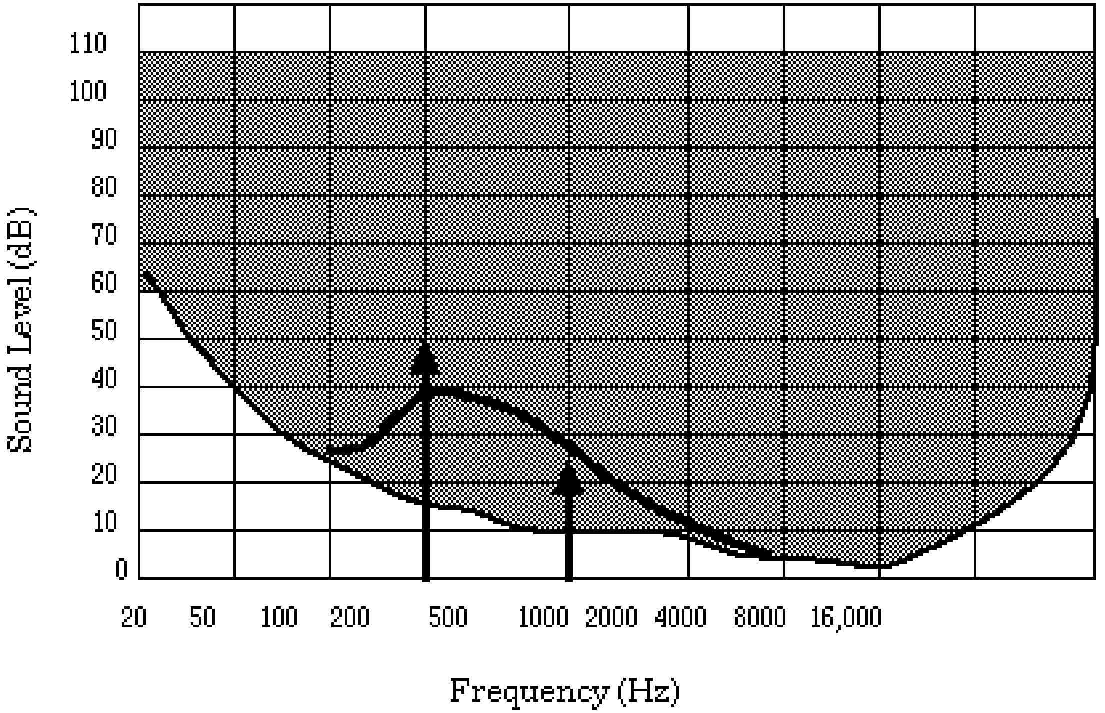
- The term for this process is masking. (i.e., reducing the sensitivity)

Masking

- Human ear is sensitive to signals in the range 15 Hz to 20 kHz, but level of sensitivity to each signal is non-linear – i.e. Ear is more sensitive to some signals than others.
- A signal may mask (reduce the sensitivity) other signals.
- When ear hears a signal at one frequency it may reduce the level of sensitivity of another signal at a similar frequency - called **frequency masking**.

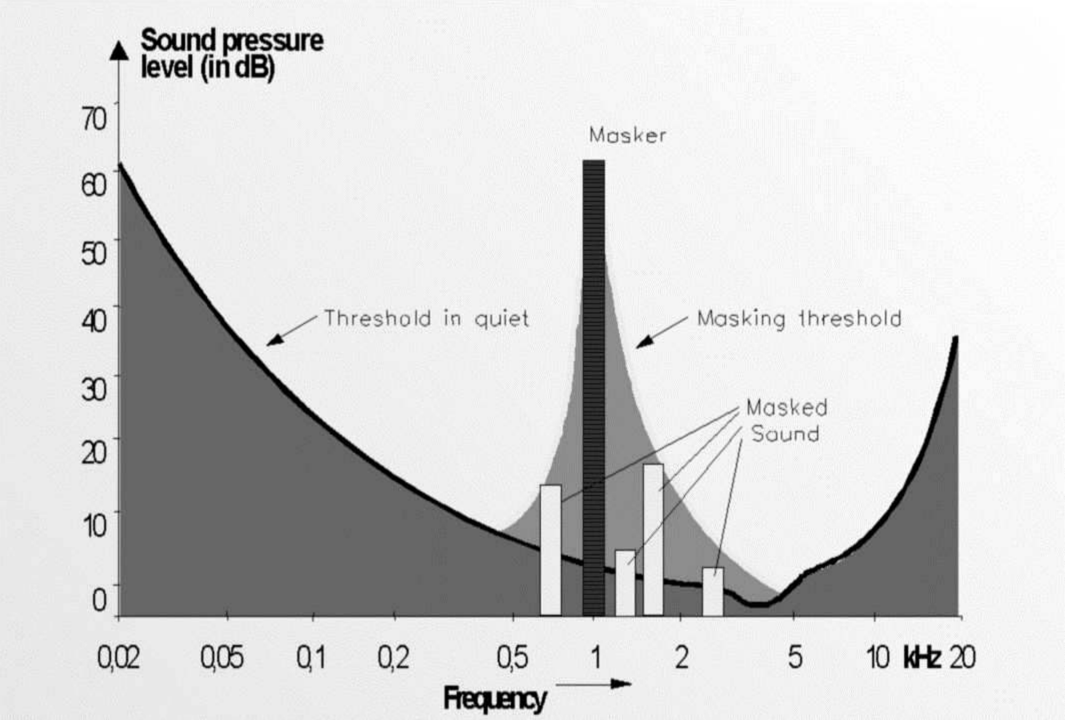
- For a strong signal, there is a short time afterwards while we cannot hear a quieter sound - called **temporal masking**.

- Ear is most sensitive in the range 2k - 5k Hz.



Frequency Masking

If a tone played at (say) 0.7 kHz then normally perceptual. But if louder tone played at (say) 1 kHz, then the 0.7 kHz tine is inaudible

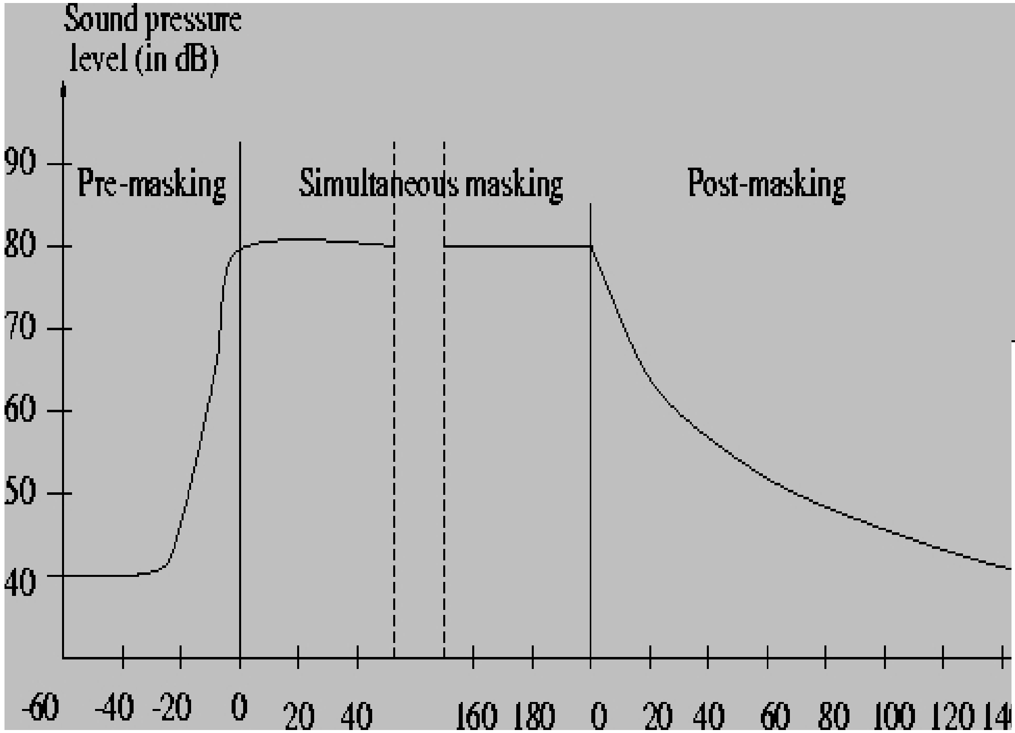


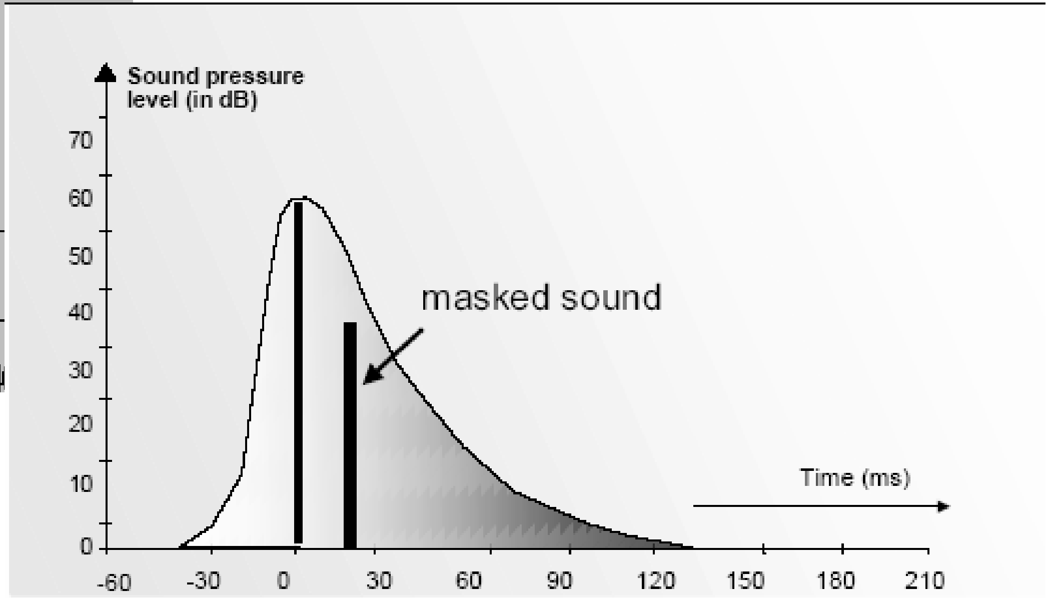
Hence decompose the audio spectrum into frequency bands and if level in one band is high enough, can omit the lower level signal in a neighbouring band.

Hence *compression*.

Temporal Masking

A sudden loud signal will temporally raise the threshold of audibility





Masking: Example

After analysis, the first 16 levels of the 32 bands are as follows:

Band	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Level (dB)	0	8	12	10	6	2	10	60	35	20	15	2	3	5	3	1

Level in Band 8 is 60 dB

- This can cause a masking of
  - 12 dB in band 7
  - 15 dB in band 9

How can this be exploited in audio compression?

????

Multimedia compression standards

Why Standards?

- An internationally consistent formats for representing multimedia.
- Multimedia (digital content) can be used effectively. (No conversions necessary)

- Standards are developed by calling for proposals – leads to international competition – good for the advancement of the technology.
- Only the decoder is standardised. Decoders are available in players (image and video displays). The consumer electronic device designers face the challenging problem of designing "good" encoders for the standardised decoder, so that the encoder-decoder pair gives good performances.

Major Standards Organisations:

International Organization for Standardization (ISO)

e.g., MPEG standards (Motion Picture Expert Group)

International Telecommunication Union (ITU-T)

e.g., H.26x standards by VCEG (Video Coding Experts Group)

Jointly by above both groups

e.g. Joint Picture Expert Group (JPEG), Joint Video Team (JVT)

MPEG Audio Compression

- Standard compression used on CD, MPEG players, digital TV.
- Range of standards - different qualities - different bit rates

- Decompose the PCM encoded audio bit stream into 32 frequency bands (Use the Discrete Fourier Transform)
- Use a psychoacoustic model (simulates frequency and temporal masking) to assign number of bits to each of these 32 sub-band channels.

- Combine sub-bands into a single frame (unit) - causes delay (typically 20 - 60 ms)

- At decoder, de-multiplex the frames and re-create the 32 sub-band channels and re-synthesise the audio

Audio Coding Standards

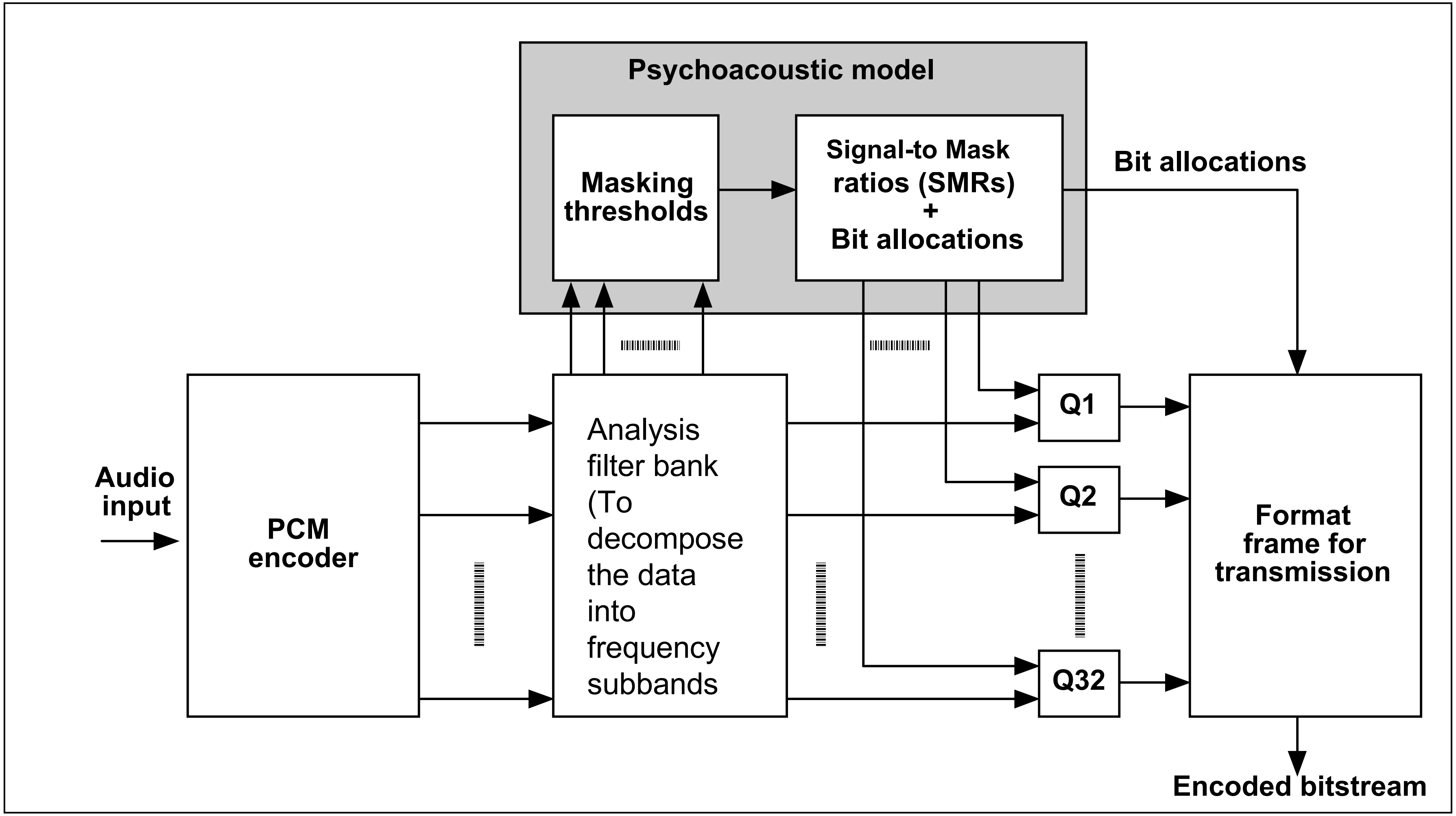
MPEG-1 Audio layers

Layer -I: Supports data rates: 32 kbps - 384 kbps → compressed to 4:1
- Applications: @384 kbps digital compact cassette (DCC)

Layer -II: Supports data rates 32 kbps - 384 kbps → compressed to 6:1
-Applications: @224 kbps, direct broadcast satellite (DBS)
@256 kbps, Eureka 147 DAB

Layer –III (1987): Supports data rates 32 kbps to 320 kbps → compressed to 10:1 to 12:1
- Applications : MP3 music
-same as layers I/II but additional inclusion of a Modified Discrete Cosine Transform (MDCT) for further frequency band decomposition.

MPEG Encoder

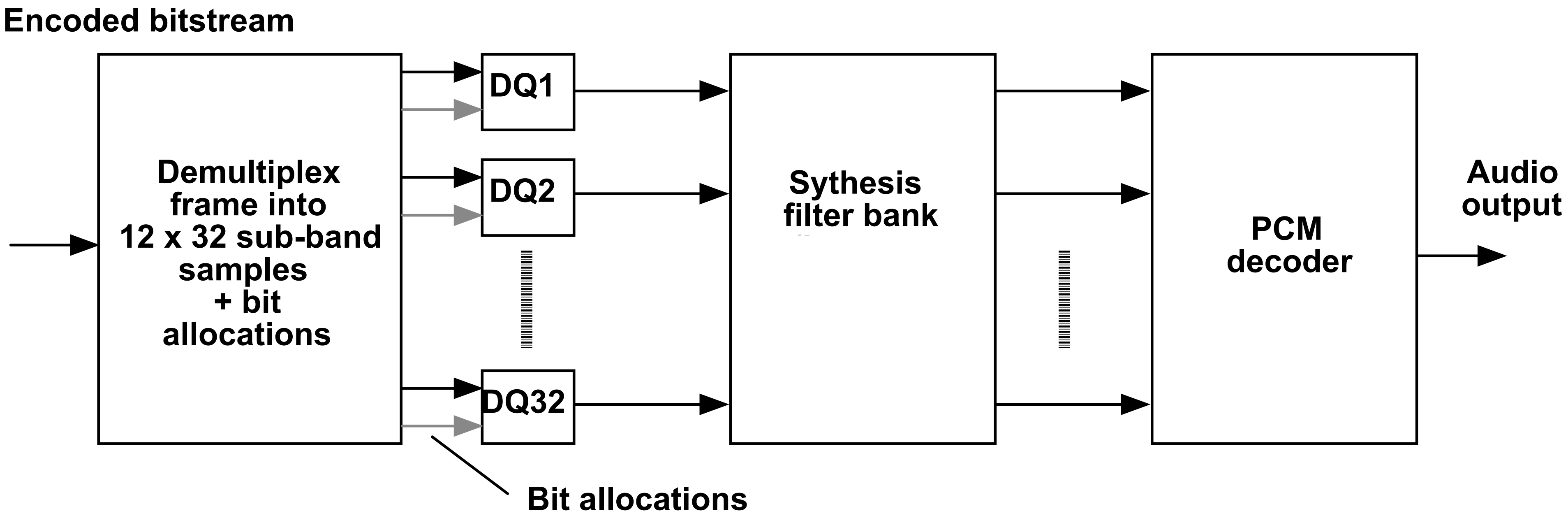

Psychoacoustic model
Masking thresholds
Signal-to Mask ratios (SMRs) + Bit allocations
Bit allocations
Audio input
PCM encoder
Analysis filter bank (To decompose the data into frequency subbands
Q1
Q2
Q32
Format frame for transmission
Encoded bitstream

Q1…Q32 represents quantisation to represent larger values with less accuracy (similar to companding)

Quantised values are Huffman coded.

This block diagram represents a basic MP3 encoder.

MPEG Decoder



Encoded bitstream

Demultiplex frame into 12 x 32 sub-band samples + bit allocations

DQ1

DQ2

DQ32

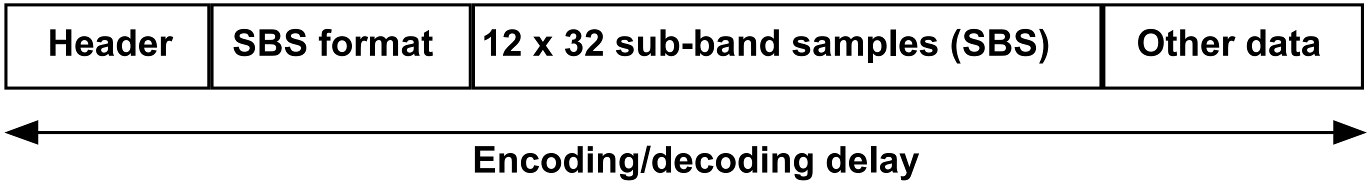
Sythesis filter bank

PCM decoder

Audio output

Bit allocations

MPEG Frame Format



Header

SBS format

12 x 32 sub-band samples (SBS)

Other data

Encoding/decoding delay

Audio Coding Standards

MPEG-2 BC/LSF (1994)

- BC: backward compatible
- LSF: low sampling frequency
- Mono, stereo, supports 16, 22.05, 24, 32, 44.1 and 48 kHz
- supports data rates: 32-640 kbps
  - Applications: Cinema, Digital Television

MPEG-2 NBC/AAC (1996)

- NBC/AAC: Non-backward compatible/Advanced audio coding
- Profiles: Main/ Low complexity (LC)/ Scalable sample rate (SSR)
- 5-channel: left, right, centre, surround left, surround right
- Supports 32, 44.1 and 48 kHz sampling rates
- Supports data rates: 8-64 kbps /channel
  - Applications : Internet audio

Audio Coding Standards

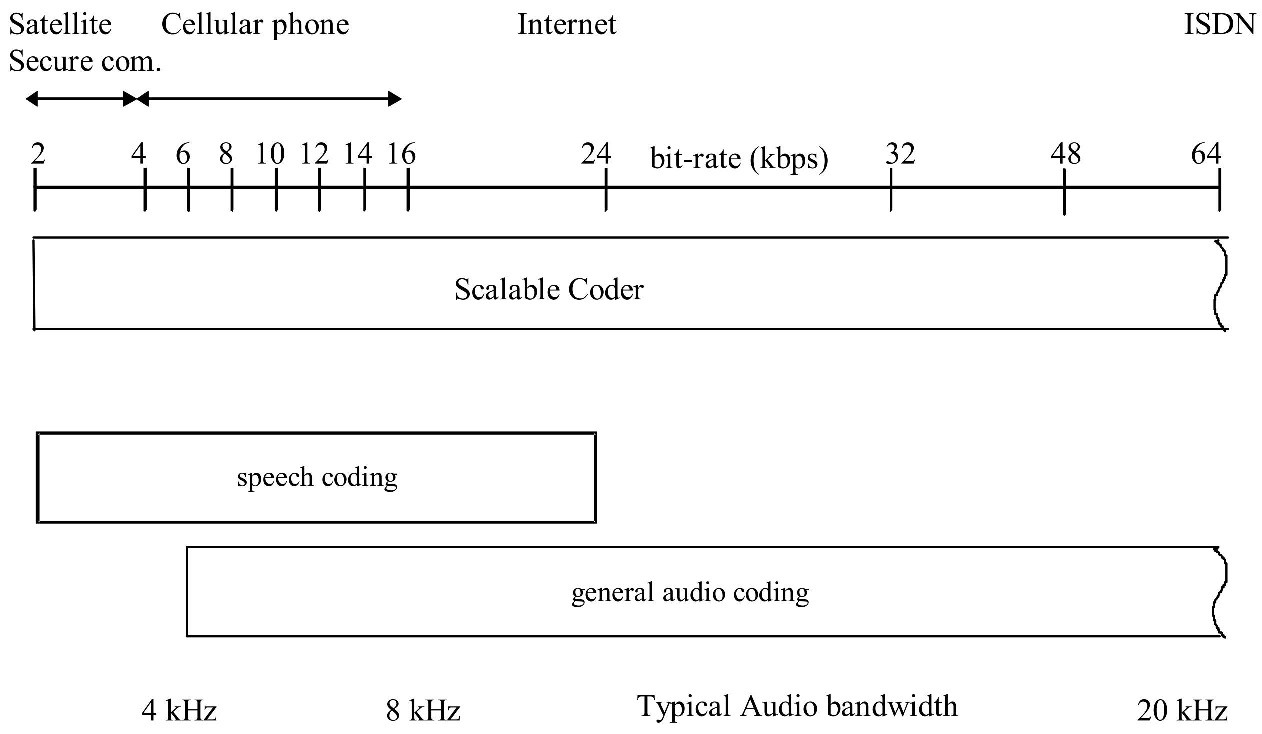
MPEG-4 (1999)

- Built on top of the MPEG-2 AAC
- by adding/removing extra tools.

- Supports data rates of 200 bps – 64 kbps.

- Target Applications:
  - very low-bit rate streaming, mobile phone, Internet audio

MPEG-4 Audio Usage



Mid term test

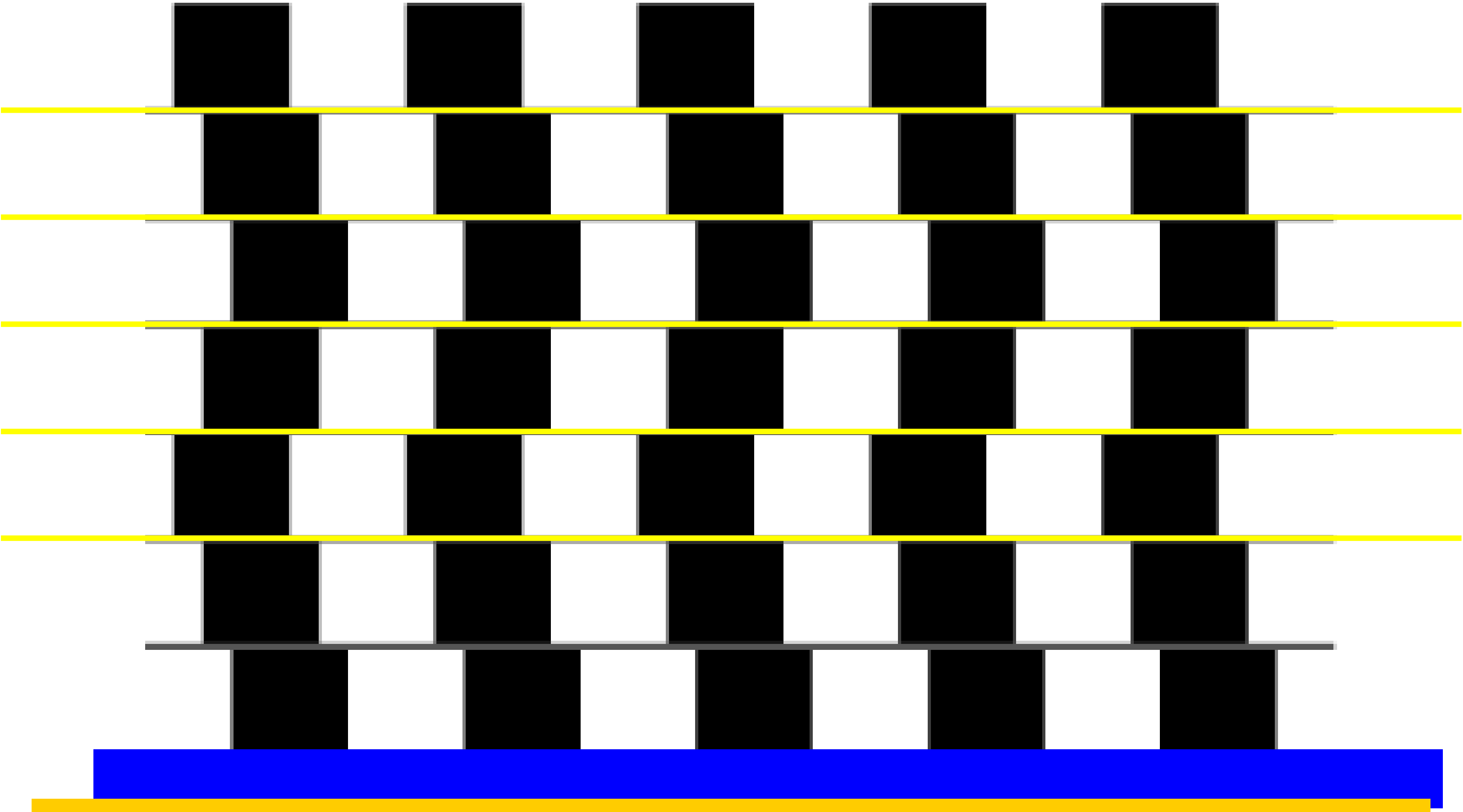
- Friday 27th March 2009 @ **9.00 am** in the 1st year labs (Portobello centre)

- 30 minutes – 12 short questions – 1 mark per each correctly answered question.
- Contributes to 9.1% of final marks for EEE116.

- All materials covered so far Topic 1 - 5, Tutorial problems Q1-Q10.
- Self Assessment Quizzes – Topic 2, Topic 3 and Topic 4 - 5
- Last year's mid term test + feedback in MOLE.

- Make sure you write "units" for all your answers.
- Bring
  - U Card
  - Calculators

A sneak preview of the next lecture



- Human Visual System studies will explain why
- Topic 6