

Brunel University

Department of Electronic and Computer Engineering

Final Examination June 2002

EE3052B - MultiMedia Signal Processing

ANSWERS

Time allowed 3 Hours

Answer five out of eight questions

Ensure that your registration number is written clearly on the front cover.

Q1

(a)

(i) MPEG is the acronym for moving picture experts group established in 1988 to develop open standards for development of coders for moving pictures and audio. [1 mark]

(ii) MPEG 1, MPEG 2, MPEG3 (rolled into MPEG2), MPEG4, MPEG7.

An “open standard coder” defines functionality and parameters such as bit rate, signal to quantisation noise and model orders but leaves other details of implementation and complexity to the developers.

[1 mark]

(iii)

MPEG-1 consists of three different operating modes, called layers. MPEG-1 defines audio compression at 32 kHz, 44.1 kHz and 48 kHz. It works with both mono and stereo signals. MPEG-1 Layer-3 provides high quality audio at about 128 kbps for stereo signals.

MPEG-2 introduced new concepts for video coding and digital TV. It also extends MPEG-1 audio sampling rates to half rates to include 16 kHz, 22.05 kHz, and 24 kHz.

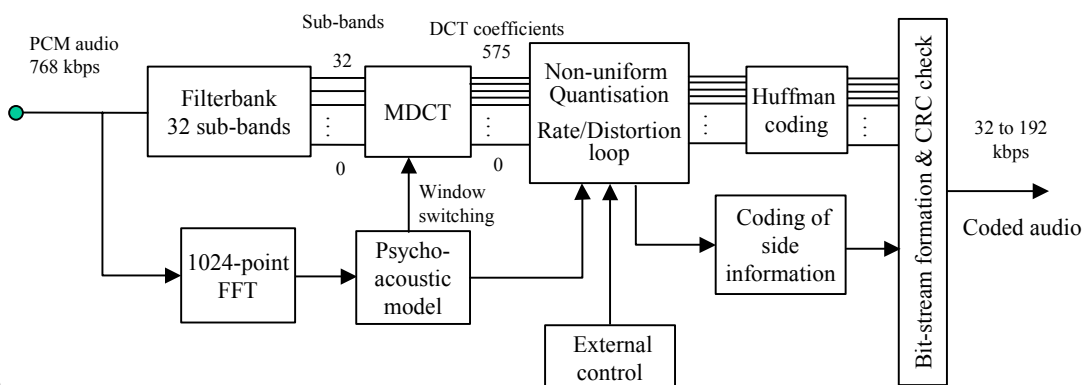
MPEG-3 was to define video coding for high definition television (HDTV) applications. However as MPEG-2 contains all that is needed for HDTV, MPEG-3 was rolled into MPEG-2.

MPEG-4 is more concerned with new functionalities than better compression efficiency. The major applications of MPEG-4 are mobile and fixed terminals, database access, communications and interactive services. MPEG-4 audio consists of audio coders spanning the range from 2 kbps low bit rate speech coding, up to 64 kbps/channel high quality audio coding.

MPEG-7 is a content representation standard for multimedia information search engines, filtering, management and processing.

[3 marks]

(b)



(b)

The detailed block diagram of an MPEG1 layer 3 (also known as MP3) music coder.

[5 marks]

(c)

(i) The filter bank unit consists of 32 poly-phase filters. Quadrature mirror filters are used for their good anti-aliasing properties.

[1 mark]

(ii) Each filter is followed by a modified discrete cosine transform (MDCT). The modified discrete cosine transform splits the signal in each band into finer resolution and also compresses most of the signal energy into a few coefficients.

[1 mark]

(iii) The Huffman coder is a probabilistic coding method that achieves coding efficiency through assigning shorter length codewords to the more probable signal values and longer length codewords to less frequent values. Consequently, for audio signals smaller quantised values, which are more frequent, are assigned shorter length codewords and larger values, which are less frequent, are assigned longer length codewords.

[2 marks]

(iv) Psychoacoustic models are used to shape the quantisation noise so that it is below the masking threshold. If the quantisation noise in a band exceeds the masking threshold then the scale factor for that band is adjusted to reduce the noise below the masking threshold. However this adjustment and control of distortion noise can result in a higher bit rate than that allowed. So the rate adjustment loop has to be repeated each time scale-factors are adjusted. Therefore the rate loop is nested in the distortion loop.

[2 marks]

(v)

Quantisation and coding is achieved through an iterative two-stage optimisation loop. A power law quantiser is used so that large values are coded with less accuracy, as higher signal energy would mask more noise. The quantised values are then coded by Huffman coding. To adapt the coder to the local statistics of the audio signal the best Huffman coding table is selected from a number of choices. The relative proportion of smaller and larger sample values is controlled using a global gain and set of subband scalefactors. If the number of available bits is not enough to encode a block of data then the global gain can be adjusted to result in a larger quantisation step size and smaller quantised values. This is repeated with different values of quantisation step size until bit requirements is no more than the available bit resources.

[4 marks]

(a)

(i)

Information gives/predicts the state(s) of a multi-state random variable.

[2 marks]

(ii) Information is carried by random variables, and a probability function is used to model a random variable. \log_2 probability is used because logarithm has the desired properties such as $\log(1)=0$; also base 2 reflects the fact that information is ultimately binary in nature.

[2 marks]

(iii) The information content of a variable is given by

$$I(x_i) = -P_X(x_i) \log P_X(x_i)$$

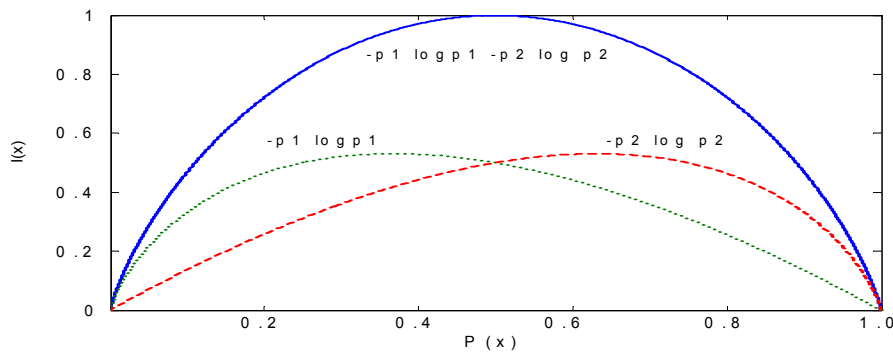


Illustration of $I(x_i)$ vs $P(x_i)$, For a binary source the maximum information content is one bit, when each state has a probability of 0.5.

[4 marks]

(iv) Entropy gives a measure of the information content of a random (source) variable in terms of the minimum number of bits required to encode the variable. Entropy can be used to calculate the theoretical minimum capacity or bandwidth required for the storage or transmission of an information source. Consider a random variable X with M states $[x_1, x_2, \dots, x_M]$ and state probabilities $[p_1, p_2, \dots, p_M]$ where $P_X(x_i) = p_i$. The entropy of X is defined as

$$H(X) = - \sum_{i=1}^M P(x_i) \log P(x_i)$$

where the base of the logarithm is 2.

[4 marks]

(b)

(i) Written English text composed of a combination of 26 letters, 10 numbers and 30 symbols. Total of 66 symbols, assuming uniform probability distribution for there we have

$$H(X) = -\sum_{i=1}^M P(x_i) \log P(x_i) = -\sum_{i=1}^M \frac{1}{66} \log\left(\frac{1}{66}\right) = 6.04 \text{ bits / symbol} \quad [2 \text{ marks}]$$

(ii) Spoken English speech is composed of 40 phonemic symbols. Assuming uniform probability distribution for the we have

$$H(X) = -\sum_{i=1}^M P(x_i) \log P(x_i) = -\sum_{i=1}^{40} \frac{1}{40} \log\left(\frac{1}{40}\right) = 5.32 \text{ bits / symbol} \quad [2 \text{ marks}]$$

(c)

(i) $(3000 \times 5 + 50) \times 6.04 = 91 \text{ kbits}$

[2 marks]

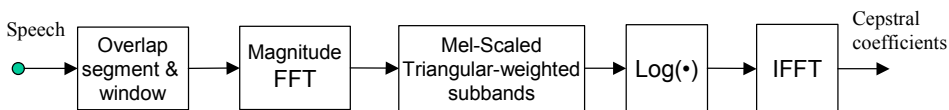
(ii) $(30 \times 120 \times 5) \times 5.32 = 95.8 \text{ kbits.}$

[2 marks]

Q3

(a)

(i)



Block diagram of a typical cepstral feature extraction system.

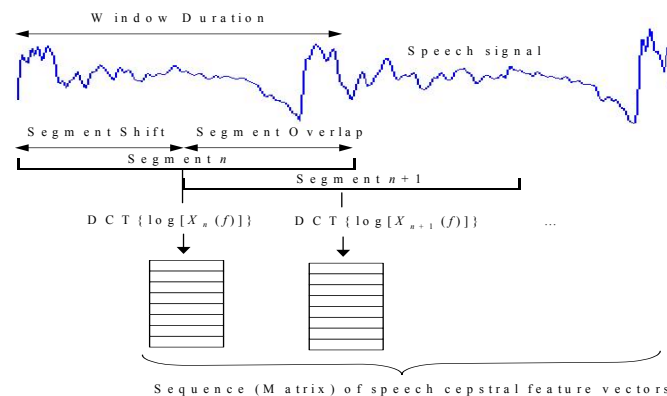


Illustration of Speech Feature Extraction

[2 marks]

(ii)

$$c(m) = \sum_{k=0}^{N-1} \ln[|X(k)|] e^{\frac{j2\pi mk}{N}}$$

where $X(k)$ is the FFT-based mel-scaled triangular-weighted spectrum of speech $x(m)$.

$$c(m) = DCT \left\{ \ln[|X(k)|] \right\}$$

[2 marks]

(iii) The use of difference features improve the accuracy in speech recognition. Cepstral difference features are defined by

(iv)

$$\begin{aligned}\partial c(m) &= c(m+1) - c(m-1) \\ \partial\partial c(m) &= \partial c(m+1) - \partial c(m-1)\end{aligned}$$

where $\partial c(m)$ and $\partial\partial c(m)$ are the first order and the second order time-difference of cepstral features.

[2 marks]

(b)

(i)

Dynamic Time Warping (DTW) is a method for aligning two sequences. Speech is a time-varying process in which the duration of a word and its subwords varies randomly. Hence a method is required to find the best time-alignment between a sequence of vector features representing a spoken word and the model candidates. For isolated-word recognition the time-alignment method used is dynamic-time warping (DTW). The best matching template is the one with the lowest distance path aligning the input pattern to the template.

[1 mark]

(ii) With the aid an appropriate sketch and the relevant equations explain how DTW works.

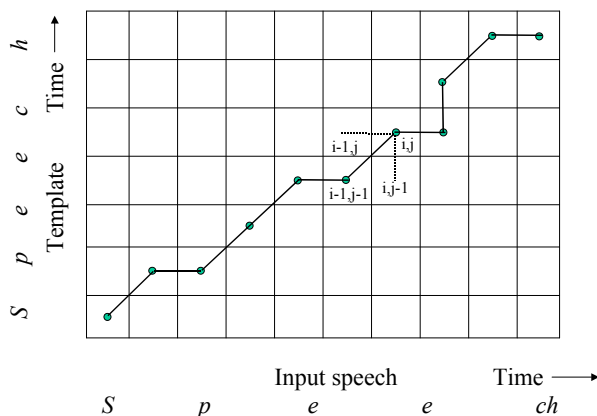


Illustration of DTW of input speech and a template model.

For illustration of DTW consider a point (i, j) in the time-time matrix (where i indexes the input pattern frame, and j the template frame), then previous point must have been $(i-1, j-1)$, $(i-1, j)$ or $(i, j-1)$. The key idea in dynamic programming is that at point (i, j) we continue with the lowest accumulated distance path from $(i-1, j-1)$, $(i-1, j)$ or $(i, j-1)$. If $D(i, j)$ is the global distance up to (i, j) and the local distance at (i, j) is $d(i, j)$ then we have the recursive relation

$$D(i, j) = \min[D(i-1, j), D(i, j-1), D(i-1, j-1)] + d(i, j) \quad (13.38)$$

Given that $D(1,1) = d(1,1)$, we have the basis for an efficient recursive algorithm for computing $D(i, j)$. The final global distance $D(n, N)$ gives us the overall matching score of the template with the input. The input word is then recognized as the word corresponding to the template with the lowest matching score.

[4 marks]

(iii) Two applications of DTW: Speech recognition, DNA sequence matching

[1 mark]

(c)

(i)

The dialling method should be developed with the objective of optimising the requirements for high accuracy, simplicity of use and minimum of; memory, computational and power requirement. The system stores an acoustic template for each name along with its text form and telephone number. The text for each name and telephone number are input manually. The recognition system is an isolated-word recognition that makes use of the LPC-Cepstrum parameters and the voice activity detector (VAD) of the mobile phone system.

For name dialling task the vocabulary is simple and consists of N names

$$\text{Names} = \text{sil} \langle \text{Name1} \mid \text{Name2} \mid \text{Name3} \mid \dots \text{Name } N \rangle \text{sil}$$

where N is typically 8, and sil indicates silence/noise i.e. no speech activity. Names, numbers and acoustic templates for each name are stored as shown.


Name	Number	Acoustic Template
Luke	07777206036	
William	05678226226
Colin
Kate

Illustration of stored codebook of name, number and feature template for name-dialling.

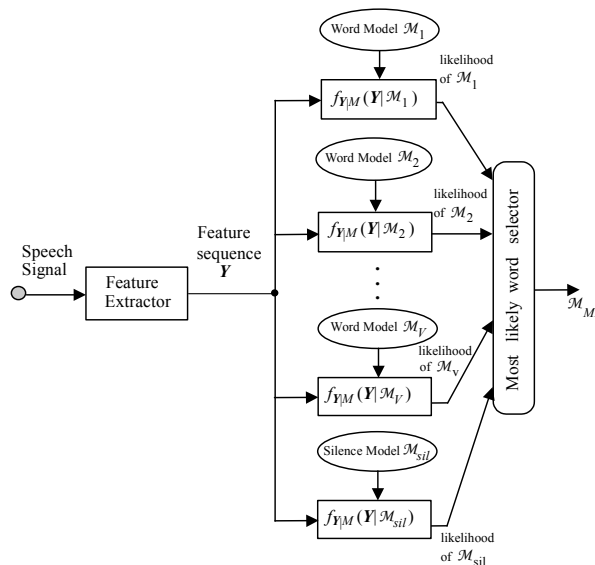


Illustration of a speech recognition system.

[4 marks]

(ii) Cepstral feature vector size 10 plus delta cepstrum = 20.

(iii)

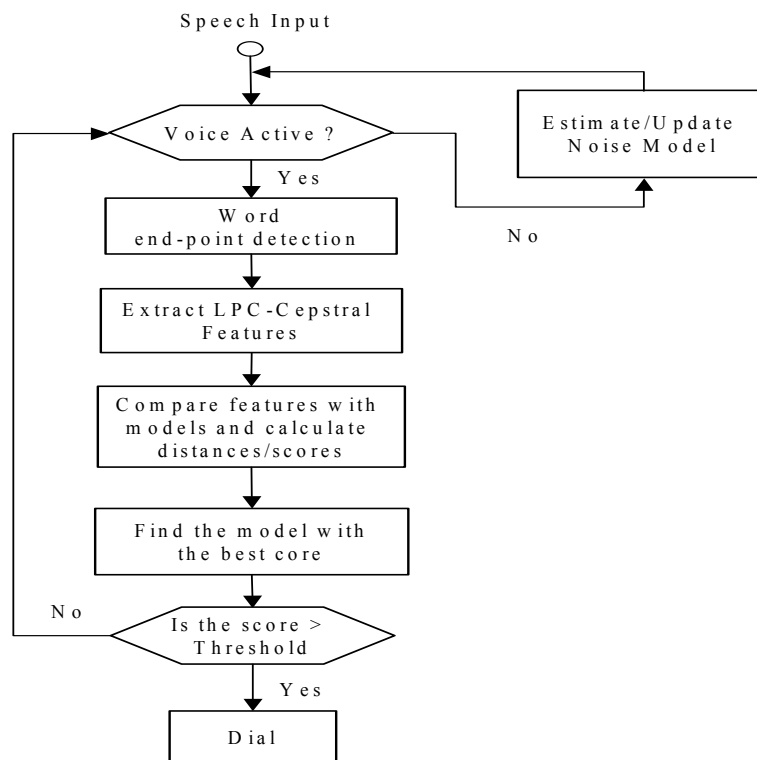


Illustration of voice-activated dialling.

Each name is modeled (trained) using one spoken example of the name. The distance metric used for the selection of the nearest name in the memory $\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k\}$ to the spoken input name \mathbf{X} can be the mean squared error or alternatively the mean absolute value of error. The labelling of the input feature matrix \mathbf{X} is achieved as

$$\text{Label}(\mathbf{X}) = \arg \min_k \{ |\mathbf{X} - \mathcal{M}_k|^2 \}$$

$k=1, \dots, N$
[3 marks]

Q4

(a)

(i) The Z transfer function of a finite impulse response filter H(z) is given by

$$H(z) = 1 - 2.5z^{-1} + 5.25z^{-2} - 2.5z^{-3} + z^{-4}$$

[1 mark]

(ii)

$$H(z) = (1 - 0.5z^{-1} + 0.25z^{-2})(1 - 2z^{-1} + 4z^{-2})$$

$$H(\omega) = 1 - 2.5e^{-j\omega} + 5.25e^{-j2\omega} - 2.5e^{-j3\omega} + e^{-j4\omega}$$

$$H(\omega) = e^{-j2\omega} (e^{j2\omega} - 2.5e^{j\omega} + 5.25 - 2.5e^{-j\omega} + e^{-j2\omega})$$

$$H(\omega) = e^{-j2\omega} (2 \cos(2\omega) - 5 \cos(\omega) + 5.25)$$

hence the phase is a linear function of frequency.

[3 marks]

(iii)

$$H(z) = (1 - 0.5z^{-1} + 0.25z^{-2})(1 - 2z^{-1} + 4z^{-2})$$

Zeros at radii of 0.5 and 2 angles of +60 and -60.

The constraint for a linear FI filter is that the zeros must be reciprocally mirrored w.r.t the unit circle.

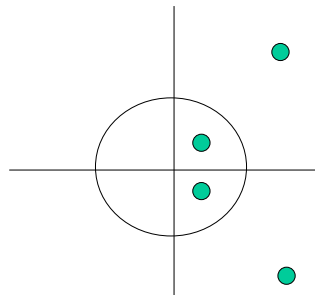


Figure 4.1

(b)

To boost the low frequency part of an audio signal spectrum. The required frequency response of a typical Bass booster is shown in the figure. Using the window design technique and the inverse DFT design a digital bass filter with the frequency response shown in figure(4.2).

$$H(f) = \begin{cases} 1 + 2.1623f / 400 & |f| \leq 400 \\ 1 & |f| > 400 \end{cases}$$

The inverse Fourier transform of a rectangular frequency response is given by

$$h(m) = 2F_c \text{sinc}(2\pi F_c m)$$

The Bass response is the sum of a rectangular response and two symmetric ramp functions, furthermore a ramp is the integral of a rectangular pulse and integration in frequency w.r.t the variable f is equivalent to multiplication by $j2\pi m$ with m being the time variable.

$$\begin{aligned}
 h(m) &= 2F_c \operatorname{sinc}(2\pi F_c m) + j2\pi m e^{j\pi F_c} 2F_c \operatorname{sinc}(2\pi F_c m) + j2\pi m e^{-j\pi F_c} 2F_c \operatorname{sinc}(2\pi F_c m) \\
 &= 2F_c \operatorname{sinc}(2\pi F_c m) + j4\pi m \cos(2\pi F_c m) F_c \operatorname{sinc}(2\pi F_c m)
 \end{aligned}$$

[7 marks]

The DFT is used to obtain the impulse response corresponding to the specified frequency response. The inverse frequency response has in theory infinite duration. A truncated version of the impulse response and its frequency spectrum are shown in figures (3.21-a) and (3.21-b). Note that the frequency response of the filter obtained through IDFT is a good approximation of the specified frequency response.

The filter can be made causal by a shift of the time domain axis. [2 marks]

The filter response tends to the ideal case as the filter length increases.

[1 mark]

Gain at 400 = $20 \cdot \log(1 + 2.1623 \cdot 400/400) = 10 \text{ dB}$

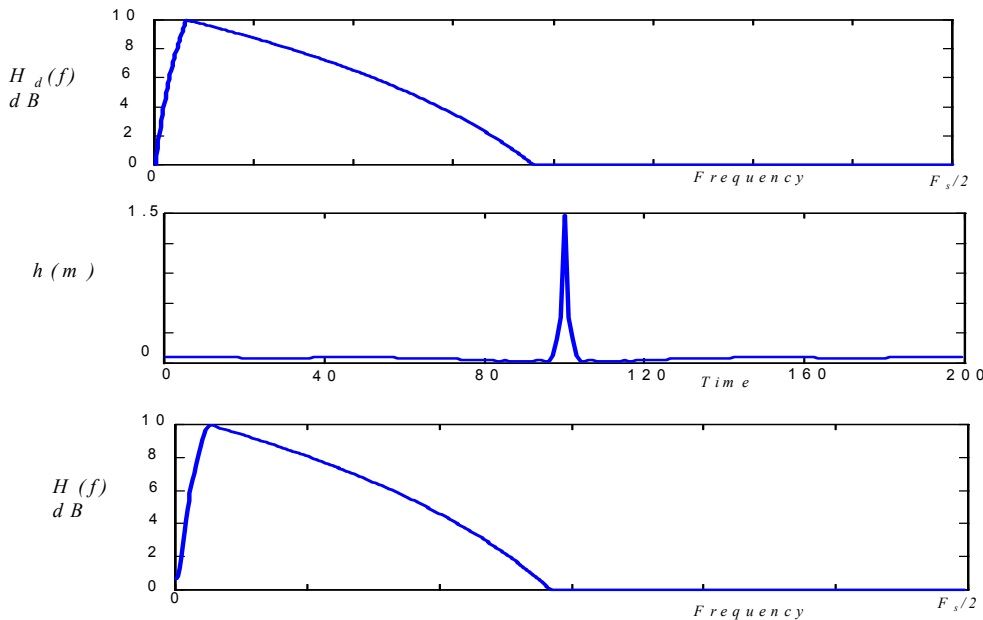


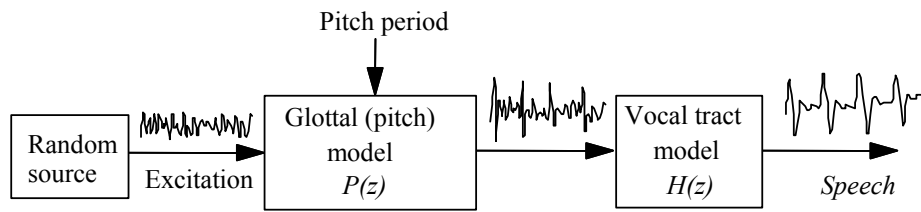
Figure 4.2

(a) The specified frequency response of a digital Bass filter, (b) Truncated inverse DFT of the frequency response specified in (a), and (c) the frequency response of the truncated impulse response.

[3 marks]

Q5

(a)



A source-filter model of speech production.

Source models the input from lung (exhaled air).

Pitch filter models the glottal cords' vibrations.

A linear predictor filter models the vocal tract resonance vibrations.

[3 marks]

(b)

(i) GSM sampling rate = 8 kHz, bit per sample =1 bit, bps=8000.

(ii) Linear speech 13 bit/sample coded speech 1 bit per sample compression ratio 13:1.

(iii) Three benefits of speech compression for mobile phone applications:

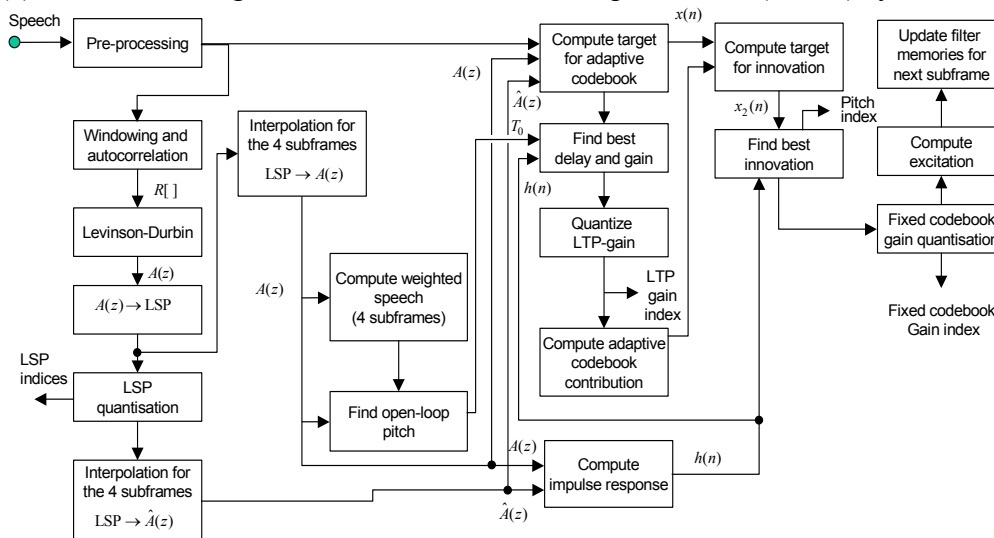
Bandwidth saving,

Power saving,

Improved immunity to noise through adding some of the removed redundancy in the form of error control bits.

[3 marks]

(c) The block diagram of a code excited linear prediction (CELP) system



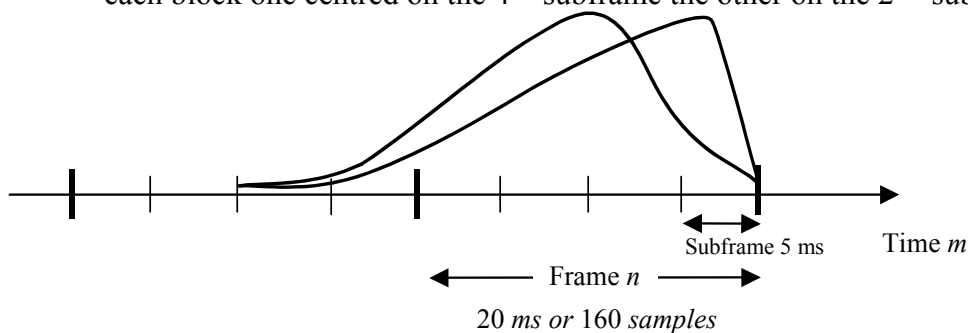
Block diagram of a code excited linear prediction (CELP) coder.

[4 marks]

The principle outline of the operation of CELP coders is as follows:

- (1) The signal is analysed at regular time-intervals to obtain the parameters of a 10th order linear prediction model. The LP parameters are transformed into line spectral pairs and quantised.
- (2) The input to LP filter is a weighted mixture of two inputs: a noise-like excitation from a fixed codebook and a periodic (pitch) excitation from an adaptive codebook. The optimum excitation is chosen using an analysis-by-synthesis search method with the objective of minimising a perceptually weighted difference between the original and the synthesised speech.
- (3) The reconstructed speech is passed through an adaptive post-filter.

- (i) Signal segmentation windowing operation: Speech is segmented into blocks of 20 ms (160 samples). Each segment is then divided into 4 subframes. Two windows are placed on each block one centred on the 4th subframe the other on the 2nd subframe.



[2 marks]

- (ii) Open-loop pitch estimation is performed every 10 ms, i.e. twice per frame, from the autocorrelation function

$$r(k) = \sum_{m=0}^{79} x(m)x(m-k)$$

At each stage the three maxima of the autocorrelation function are obtained in the following ranges: $k=18-35$, $k=36-71$ and $k=72:143$. The retained maxima are normalised by $\sqrt{\sum_m x^2(m-t_i)}$, where t_i are the maxima. The normalised maxima and corresponding delays are denoted as (M_i, t_i) $i=1,2,3$. The following method is then used to select the pitch from the three candidates

```

 $T_{op} = t_1$ 
 $M(T_{op}) = M_1$ 
if  $M_2 > 0.85M(T_{op})$ 
   $M(T_{op}) = M_2$ 
   $T_{op} = t_2$ 
end
if  $M_3 > 0.85M(T_{op})$ 
   $M(T_{op}) = M_3$ 
   $T_{op} = t_3$ 
end

```

The above procedure is designed to avoid choosing pitch multiples.

[2 marks]

- (iii) Linear predictor model order is 10 and this can model upto 5 formants, deemed sufficient for speech.

Pitch effects is modelled by a first order linear predictor.

[2 marks]

- (iv) Excitation estimation method:

Adaptive codebook search: The adaptive codebook search is performed every 5 ms for each speech subframe to obtain the pitch parameters i.e. the pitch delay and the pitch gain. The pitch values are optimised in a closed-loop pitch analysis performed around the open-loop pitch estimates by minimising the difference between the input speech and the synthesised speech. The pitch delay is coded with 9 bits in the first and third subframes and the relative delay of the other frames is coded with 6 bits.

Algebraic codebook structure and search

The algebraic code structure is based on interleaved single-pulse permutation (ISPP) method. Each codebook vector of size 40 samples contains 10 non-zero pulses with amplitudes ± 1 . Each subframe of 40 samples is subdivided into five tracks, where each track contains two pulses as shown in table 2. Each two pulse positions in an eight-positions track is coded with 3 bits (a total of 6 bits/track) and the sign of the first pulse in each track is coded with one bit (a total of 5 bits per subframe). The sign of the second pulse is the opposite of the first pulse if its position is smaller than the first pulse but it has the same sign as the first pulse otherwise. This gives a total of 35 bits for each 40-samples subframe which are Grey coded for robustness. The algebraic codebook is searched for the best vector by minimising the difference between the input speech and the synthesised speech.

[4 marks]

Q6

- (a)

- (i) Three applications for LP model, speech/video coding, data forecasting, spectral estimation

[1 Mark]

(ii)
$$x(m) = \sum_{k=1}^P a_k x(m-k) + e(m)$$

[1 Mark]

(iii)
$$H(z) = \frac{G}{1 - \sum_{k=1}^P a_k z^{-k}} = \frac{G}{\prod_{k=1}^P (1 - r_k z^{-k})}$$

[2 Mark]

(iv)
$$e(m) = x(m) - \hat{x}(m) = x(m) - \sum_{k=1}^P a_k x(m-k)$$

[1 Mark]

- (v) The “best” predictor coefficients are obtained by minimising a mean square error criterion defined as

$$\begin{aligned} \mathbb{E}[e^2(m)] &= \mathbb{E}\left[\left(x(m) - \sum_{k=1}^P a_k x(m-k)\right)^2\right] \\ &= \mathbb{E}[x^2(m)] - 2 \sum_{k=1}^P a_k \mathbb{E}[x(m)x(m-k)] + \sum_{k=1}^P a_k \sum_{j=1}^P a_j \mathbb{E}[x(m-k)x(m-j)] \\ &= r_{xx}(0) - 2\mathbf{r}_{xx}^T \mathbf{a} + \mathbf{a}^T \mathbf{R}_{xx} \mathbf{a} \end{aligned}$$

The gradient of the mean square prediction error with respect to the predictor coefficient vector \mathbf{a} is given by

$$\frac{\partial}{\partial \mathbf{a}} \mathbb{E}[e^2(m)] = -2\mathbf{r}_{xx}^T + 2\mathbf{a}^T \mathbf{R}_{xx}$$

The least mean square error solution, is given by

$$\mathbf{R}_{xx} \mathbf{a} = \mathbf{r}_{xx}$$

or

$$\mathbf{a} = \mathbf{R}_{xx}^{-1} \mathbf{r}_{xx}$$

[5 Mark]

(b)

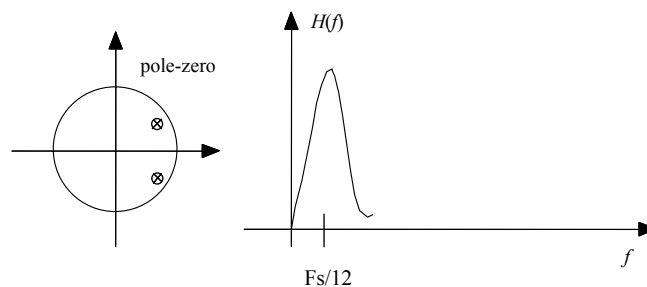
- (i) $a_1=1.645, a_2=-0.9025,$
 $r=0.95, \phi=30^\circ$

[6 Marks]

- (ii) Use the model to write an expression for the frequency response of the process and sketch the spectrum of this predictor.

$$H(z) = \frac{1}{1 - 1.645z^{-1} + 0.9025z^{-2}}$$

[2 Marks]



[2 Marks]

Q7

(a)

The decimated signal $x_d(m)$ can be expressed as

$$x_d(m) = x(Im)$$

The spectrum of the zero-inserted signal is related to the spectrum of the original discrete-time signal by

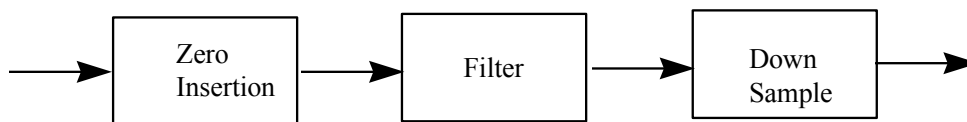
$$\begin{aligned} X_d(f) &= \sum_{m=-\infty}^{\infty} x_d(m) e^{-j2\pi f m} \\ &= \frac{1}{I} \sum_{m=-\infty}^{\infty} x(Im) e^{-j2\pi \frac{f}{I} m I} \\ &= \frac{1}{I} X(f/I) \end{aligned}$$

[4 Marks]

Hence the decimated signal is expanded and aliasing can happen.

(b) With the aid of a block diagram, describe the outline of a system for re-sampling a HiFi digital audio signal $x(m)$ originally sampled at a rate of 44 kHz to a new sampling rate of 10 kHz. What is the quantitative effect of this resampling on spectrum of the signal.

Figure shows the block diagram of a re-sampling system.



The system is composed of an up-sampling sub-block, this inserts 4 zeros in-between every two samples taking the sampling rate up by 5 times to 220 kHz. This is followed by a lowpass filter to replace the zeros with interpolated values. The cutoff frequency of the low pass filter should be $\pi/22$. The final sub block is a 1 to 22 down sampler.

The effect on the spectrum is that the frequency content will be limited to 5 kHz.

[4 Marks]

(c) Using the window design technique and the inverse Fourier transform, design a lowpass and a highpass digital finite impulse response (FIR) filters to split a total bandwidth of 20 kHz into 2 equal bandwidth sub-bands. Write the impulse response of each sub-band filter. State how this filter can be made causal.

- (d) Explain how you can use down sampling and further applications of the high pass and low pass filters to split the signal into four bands.

Assume the sampling frequency is 44 kHz. The cutoff frequency for lowpass and highpass filter is 10 kHz or in normalised form $10/44$ or $5\pi/22$. Using the window design technique the FIR impulse response is obtained from the inverse Fourier transform as

$$h_d(m) = \int_{-5/22}^{5/22} 1.0 e^{j2m\pi f} df$$

we obtain the FIR filter response as

$$h_1(m) = w(m) \times \frac{5}{11} \operatorname{sinc}\left(\frac{5}{11} \pi(m - M/2)\right) \quad 0 \leq m \leq M$$

Note for causality the filter impulse response is windowed and delayed.

[3 marks]

The impulse response of this filter is obtained via the inverse Fourier integral as

$$\begin{aligned} h_d(m) &= \int_{-1/2}^{-F_c} 1.0 e^{+j2m\pi f} df + \int_{F_c}^{1/2} 1.0 e^{+j2m\pi f} df = \frac{e^{+j2m\pi f}}{+j2m\pi} \Big|_{-1/2}^{-F_c} + \frac{e^{+j2m\pi f}}{+j2m\pi} \Big|_{F_c}^{1/2} \\ &= \frac{\sin \pi m}{m\pi} - \frac{\sin 2\pi F_c m}{m\pi} \end{aligned}$$

The impulse response $h_d(m)$ is non-causal (it is nonzero for $m < 0$) and infinite in duration. To obtain an FIR filter of order M we multiply $h_d(k)$ by a rectangular window sequence of length $M+1$ samples. To introduce causality ($h(k) = 0$ for $k < 0$) truncated shift $h(k)$ by $M/2$ samples

$$h(m) = \frac{\sin \pi (m - M/2)}{(m - M/2)\pi} - \frac{\sin 2\pi \frac{5}{22} (m - M/2)}{(m - M/2)\pi} \quad 0 \leq m \leq M$$

[3 marks]

- (c) The output of each filter can be down sampled by a factor of 2. If the filters are applied again to down-sampled signal then the original band would be split into 4 bands. In this way using a combination of 2 filters and down samplers we can progressively split the signal into a number of bands.

[4 Marks]

Q8

(a) The discrete Fourier transform (DFT) is given by

$$X(k) = \sum_{m=0}^{N-1} x(m) e^{-j\frac{2\pi}{N}mk} \quad k = 0, \dots, N-1$$

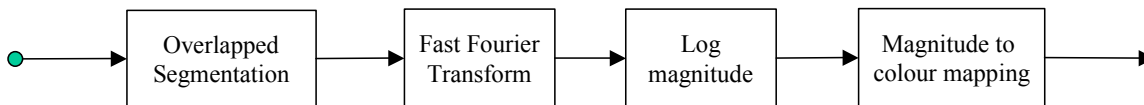
The DFT is used in spectrogram to plot the time-variation of the spectrum of a signal in spectrogram.

(i)

$$\begin{aligned} X(k + n * N) &= \sum_{m=0}^{N-1} x(m) e^{-j\frac{2\pi}{N}(k+n*N)m} = \sum_{m=0}^{N-1} x(m) e^{-j\frac{2\pi}{N}km} e^{-j2\pi n*m} \\ &= \sum_{m=0}^{N-1} x(m) e^{-j\frac{2\pi}{N}km} = X(k) \end{aligned}$$

[5 marks]

(ii) The main signal processing steps in the design of a spectrogram.



[3 Marks]

(iii) Music signal is sampled at 44100 Hz. Assuming that music is relatively stationary for about 50 milliseconds, the best choice of window length is 2205 samples.

This corresponds to a frequency resolution of $F_s/N=44100/2205=20$ Hz.

[3 Marks]

(iv) Speech is stationary about 20-30 ms.

Window length at 8 kHz is $(20/1000) \times 8000 = 160$ to $(30/1000) \times 8000 = 240$ samples.

Frequency resolution = $1/30 = 33.3$ Hz.

[3 Marks]

- (v) The spectrum of a short length signal can be interpolated to obtain a smoother spectrum. Interpolation of the frequency spectrum $X(k)$ is achieved by *zero-padding* of the time domain signal $x(m)$. Consider a signal of length N samples $[x(0), \dots, x(N-1)]$. Increase the signal length from N to $2N$ samples by padding N zeros to obtain the padded sequence $[x(0), \dots, x(N-1), 0, \dots, 0]$. The DFT of the padded signal is given by

$$\begin{aligned} X(k) &= \sum_{m=0}^{2N-1} x(m) e^{-j\frac{2\pi}{2N}mk} \\ &= \sum_{m=0}^{N-1} x(m) e^{-j\frac{\pi}{N}mk} \end{aligned} \quad k = 0, \dots, 2N-1$$

The spectrum of the zero-padded signal, is composed of $2N$ spectral samples; N of which, $[X(0), X(2), X(4), X(6), \dots, X(2N-2)]$ are the same as those that would be obtained from a DFT of the original N samples, and the other N samples $[X(1), X(3), X(5), X(7), \dots, X(2N-1)]$ are interpolated spectral lines that result from zero-padding. Note that zero padding does not increase the spectral resolution, it merely has an *interpolating or smoothing* effect in the frequency domain

[4 marks]

- (vi) Actual resolution is $441000/2000=220.5$ Hz, apparent resolution after zero padding is 10.025 Hz.

[2 Marks]