The Hong Kong Polytechnic University
Department of Electronic and Information Engineering

Programme(s)	:	Postgraduate Scheme in Engineering (05002) BEng(Hons) in Electronic and Information Engineering (42070) MSc in Information Technology (61030) BSc(Hons) in Internet and Multimedia Technologies (42077)
Subject	:	Speech Processing and Recognition Subject Code : EIE558
Session	:	Semester 2, 2012/2013
Date	:	20 May 2013
Time Allowed	:	2 hours Time : 7:00 p.m. – 9:00 p.m.
Subject Lecturer(s)	:	Dr M.W. Mak

This question paper has 9 pages including this cover page.

Instructions to Candidates: This question paper contains FOUR (4) questions.

Answer ALL questions.

ALL questions carry equal marks.

DO NOT TURN OVER THIS PAGE UNTIL YOU ARE TOLD TO DO SO

Q1 Fig. Q1(a) shows a discrete-time model for speech production.

(a) Given the impulse train $e(n)$ shown in Fig. Q1(b) and the frequency response of $G(z)$ shown in Fig. Q1(c), sketch the frequency spectrum of the glottal signal $u_{\text{glottis}}(n)$, i.e., draw $\left|U_{\text{glottis}}(\omega)\right|$ against ω.

(5 marks)

(b) Assume that the sampling frequency is 8kHz and that the first three format frequencies are 500Hz, 1500Hz, and 2500Hz, respectively. Sketch the frequency response of $V(z)$.

(5 marks)

(c) Based on your answers in Q1(a) and Q1(b), sketch the frequency spectrum of $s(n)$.

(5 marks)

(d) Discuss the effect of $G(z)$ on the frequency spectrum of the output speech signal $s(n)$.

(5 marks)

(e) Assume that the sampling frequency is 8kHz and that $V(z)$ has the form

$$V(z) = \frac{1}{1 - \sum_{k=1}^{6} v_k z^{-k}}.$$

Based on your result in Q1(b), roughly draw the pole(s) of $V(z)$ on a z-plane.

(5 marks)

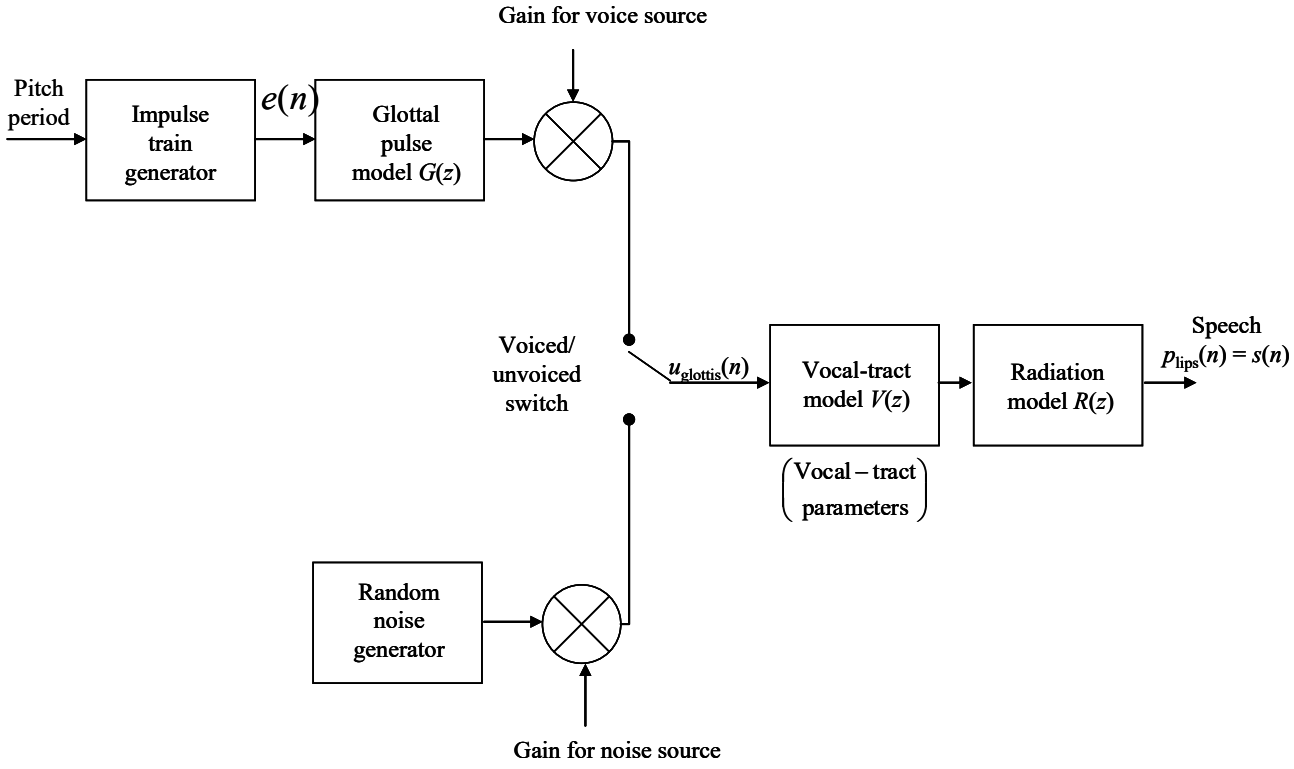


Fig. Q1(a)

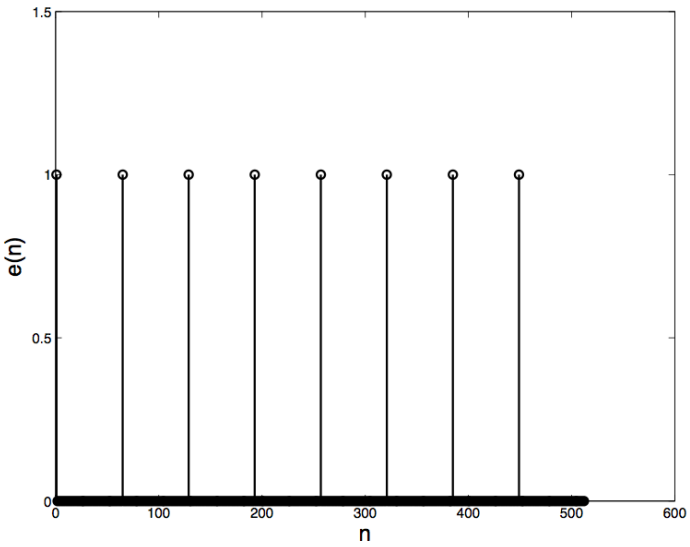


Fig. Q1(b)

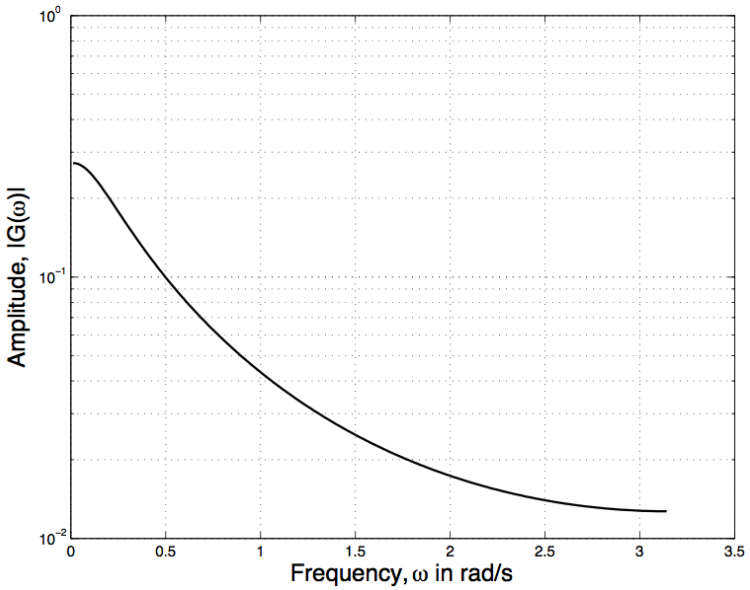


Fig. Q1(c)

Q2(a) Fig. Q2(a) shows the waveform of the word "seven". Discuss and explain the problems that may arise if the spectrogram of the signal is obtained by using overlapping frames with frame size equal to 1 second.

(6 marks)

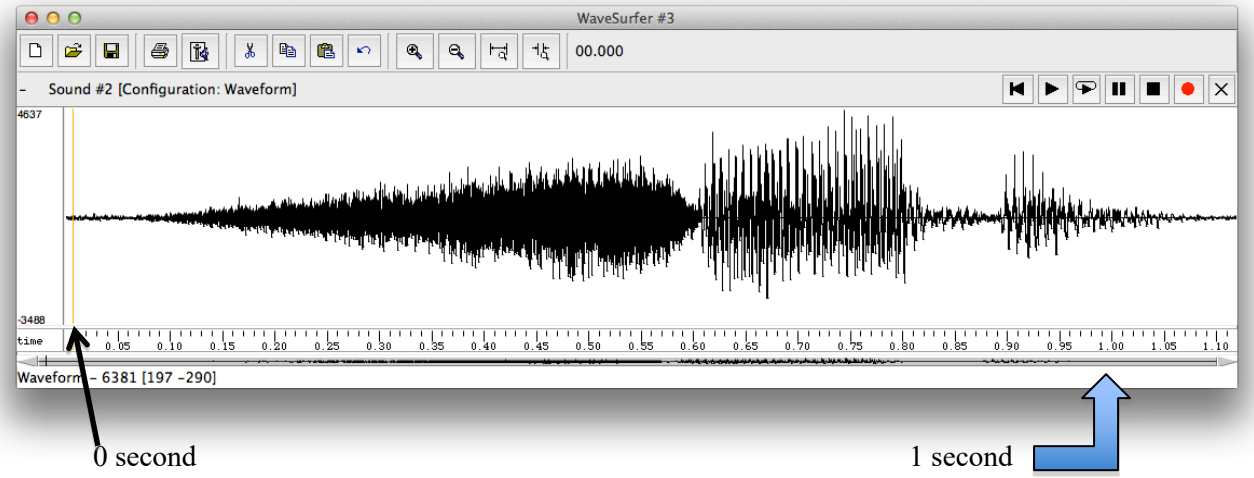


Fig. Q2(a)

(b) Fig. Q2(b) illustrates the operation of a Wiener filter with frequency response given by

$$H_s(\omega,m)=\frac{\left|\tilde{X}(\omega,m)\right|^2}{\left|\tilde{X}(\omega,m)\right|^2+\alpha\left|B(\omega)\right|^2}, \qquad \text{(Eq. Q2-1)}$$

where m is the frame index.

(i) Discuss the purpose of the smoothing operation in Fig. Q2(b). Suggest a way to implement the smoothing.

(6 marks)

(ii) Fig. Q2(c) shows the spectrogram of a noisy speech signal. Fig. Q2(d) shows the spectrograms of enhanced speech signals $\hat{x}(n,m)$ produced by using two different values of α (1 and 2) in Eq. Q2-1. Identify which of the enhanced speech signals (left panel or right panel) of Fig. Q2(d) corresponds to $\alpha=1$. Briefly explain your answer.

(6 marks)

(c) Given a frame of speech $s(n)$, its mel-frequency cepstral coefficients (MFCC) are given by

$$\text{MFCC}_n=c_n=\sum_{m=1}^{M}\cos\left[n\left(m-\frac{1}{2}\right)\frac{\pi}{M}\right]X(m), \qquad 0\le n\le P \qquad \text{(Eq. Q2-2)}$$

where

$$X(m)=\ln\left[\sum_{k=0}^{N-1}\left|S(k)\right|^2 H_m(k)\right],$$

where $|S(k)|$ is the frequency spectrum of $s(n)$ and $H_m(k)$'s are filters.

(i) Suggest a typical value for M in Eq. Q2-2.

(3 marks)

(ii) Sketch the filters $H_m(k)$ for $m=1,\ldots,M$.

(4 marks)

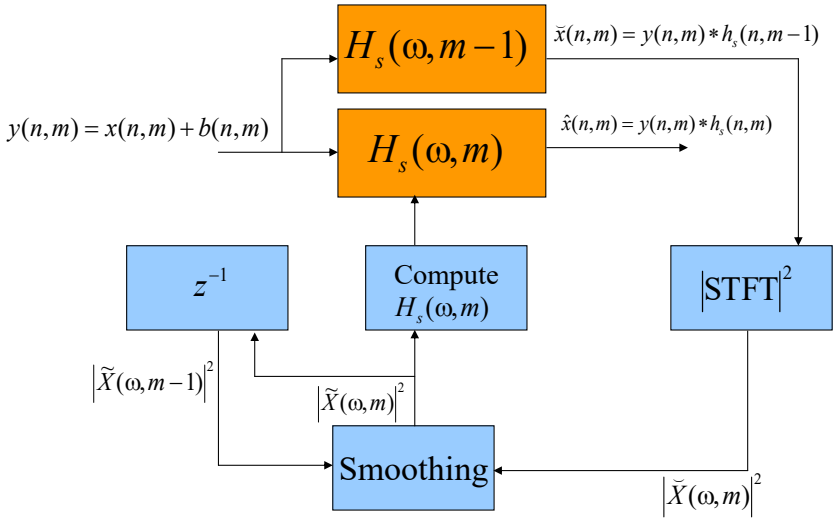


Fig. Q2(b)

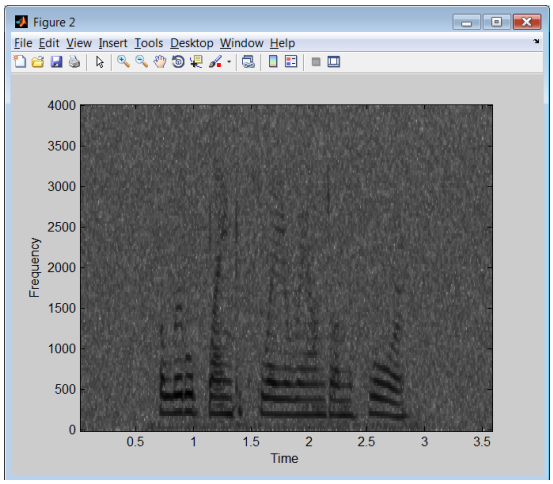
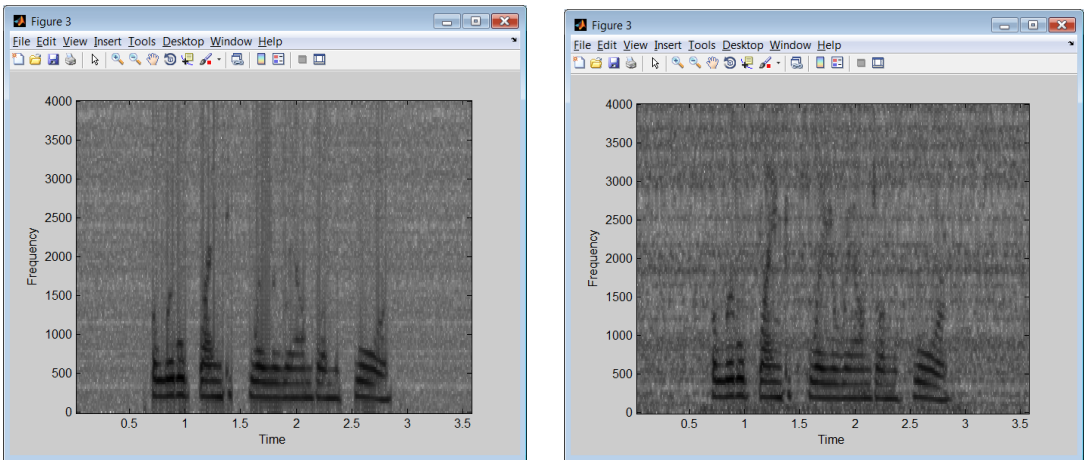


Fig. Q2(c)



Left Panel Right Panel

Fig. Q2(d)

Q3(a) Fig. Q3(a) plots the first formant frequency (F_1) against the second formant frequency (F_2) of eight vowels produced by 10 male and 10 female speakers. Each marker (hollow circle or solid circle) represents the mean of F_1 and F_2 of one vowel, averaged over 10 male or 10 female speakers.

(i) Identify which set of markers (solid circles or hollow circles) correspond to male speakers. Briefly explain your answer.

(5 marks)

(ii) Explain the causes of the difference between the formant frequencies of male and female speakers.

(5 marks)

(iii) Vocal-tract length normalization (VTLN) is a possible approach to reducing the discrepancy between the speech spectra produced by male and female speakers. An important step in VTLN is to find a warping function to normalize the vocal tract length of speakers. Assuming that a speech recognizer has been trained based on the speech of female speakers, complete Fig. Q3(b) so that the recognizer can also recognize the speech of male speakers. Put your answer in your answer book.

(6 marks)

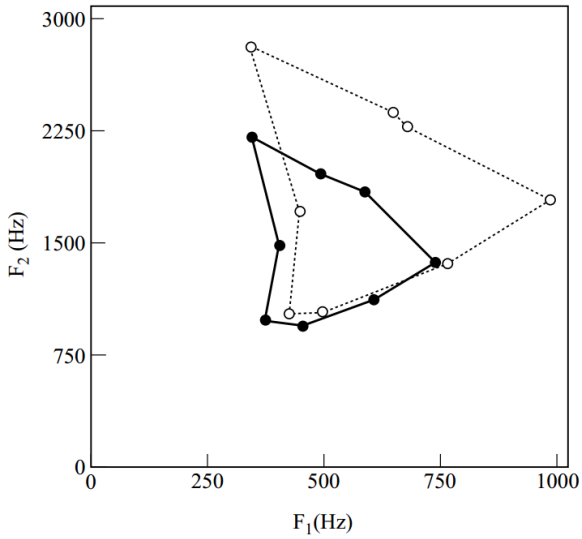


Fig. Q3(a)

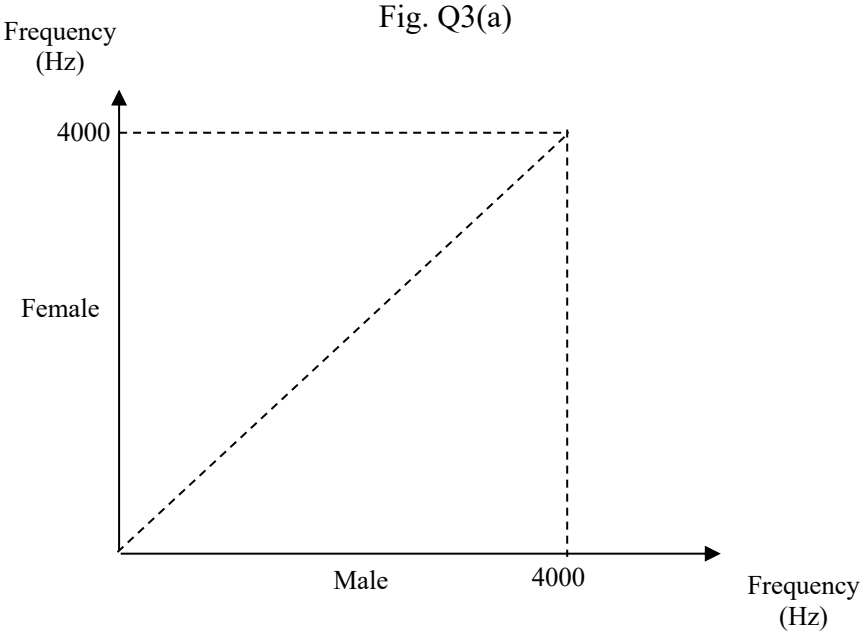


Fig. Q3(b)

Q3(b) Fig. Q3(c) shows the structure of a code-excited linear prediction (CELP) encoder.

(i) Suggest a transfer function for $P(z)$. Define the parameters in your transfer function. (3 marks)

(ii) What is the purpose of $P(z)$? (3 marks)

(iii) What will be the consequence on the output speech $\hat{s}(n)$ if $P(z)=1$? (3 marks)

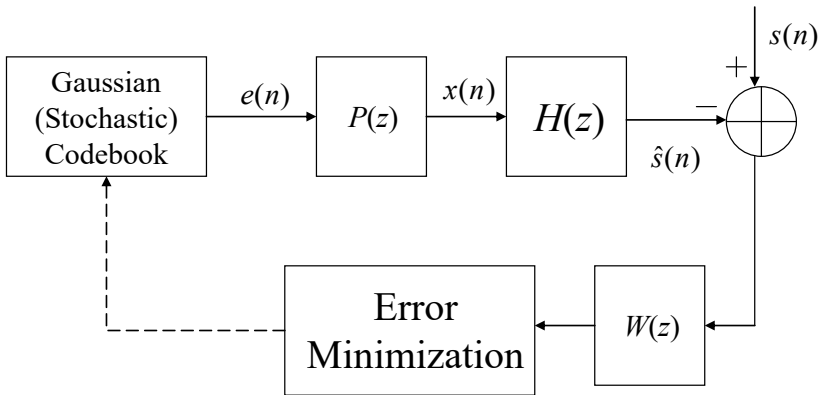


Fig. Q3(c)

Q4(a) GMM-UBM is a classical approach to speaker verification. Given a background GMM $\Lambda^{(b)} = \{\lambda_j^{(b)}, \mu_j^{(b)}, \Sigma_j^{(b)}\}_{j=1}^{M}$ and an enrollment utterance with MFCC sequence $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T\}$, a target-speaker GMM $\Lambda^{(s)} = \{\lambda_j^{(s)}, \mu_j^{(s)}, \Sigma_j^{(s)}\}_{j=1}^{M}$ can be constructed by means of the maximum a posteriori (MAP) adaptation as follows:

$$\mu_j^{(s)} = \alpha_j E_j(X) + (1-\alpha_j)\mu_j^{(b)}; \qquad \Sigma_j^{(s)} = \Sigma_j^{(b)}; \qquad \lambda_j^{(s)} = \lambda_j^{(b)}; \quad j = 1, \ldots, M$$

$$\alpha_j = \frac{n_j}{n_j + r}; \qquad r = 16$$

$$E_j(X) = \frac{1}{n_j}\sum_{t=1}^{T} \Pr(j \mid \mathbf{x}_t)\mathbf{x}_t = \frac{1}{n_j}\sum_{t=1}^{T} \gamma_j(t)\mathbf{x}_t$$

$$n_j = \sum_{t=1}^{T} \Pr(j \mid \mathbf{x}_t) = \sum_{t=1}^{T} \gamma_j(t) \qquad \text{(Eq. Q4-1)}$$

$$\gamma_j(t) = \Pr(j \mid \mathbf{x}_t) = \frac{\lambda_j^{(b)} p(\mathbf{x}_t \mid \mu_j^{(b)}, \Sigma_j^{(b)})}{\sum_{k=1}^{M} \lambda_k^{(b)} p(\mathbf{x}_t \mid \mu_k^{(b)}, \Sigma_k^{(b)})}$$

$$p(\mathbf{x}_t \mid \mu_j^{(b)}, \Sigma_j^{(b)}) = \frac{1}{\sqrt{(2\pi)^D \left|\Sigma_j^{(b)}\right|}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}_t - \mu_j^{(b)}\right)^{\mathrm{T}} \Sigma_j^{(b)-1}\left(\mathbf{x}_t - \mu_j^{(b)}\right)\right\}$$

(i) Discuss the effect of long utterances (T very large) and short utterances (T very small) on the value of α_j and $\Lambda^{(s)}$.

(6 marks)

(ii) When $r = 0$, the MAP solution in Eq. Q4-1 reduces to a maximum likelihood solution. Briefly explain this statement.

(4 marks)

(iii) Fig. Q4(a) shows the Gaussian distributions of a background model $\Lambda^{(b)} = \{\lambda_j^{(b)}, \mu_j^{(b)}, \Sigma_j^{(b)}\}_{j=1}^{M}$ with $M = 3$. The cross markers ($\times$) represent the MFCC vectors $\mathbf{x}_t$ of the enrollment utterance, where $t = 1, 2$. Arrange $\gamma_j(t)$ in Eq. Q4-1 in descending order for $j = 1, 2, 3$ and $t = 1, 2$. Hence, sketch the Gaussian distributions of the adapted model $\Lambda^{(s)}$. Your diagram should also include the two markers ($\times$).

(6 marks)

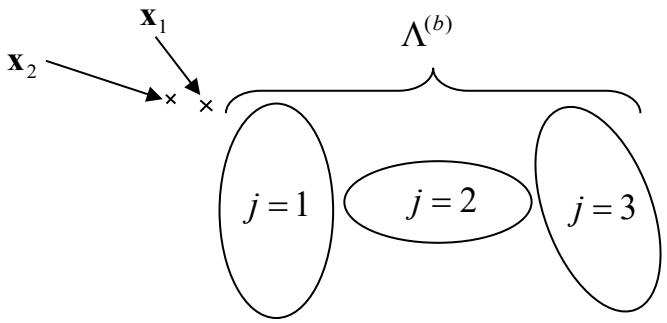


Fig. Q4(a)

Q4(b) Fig. Q4(b) shows an i-vector based speaker verification system.

(i) Discuss the purpose of the total variability matrix ***T***.

(3 marks)

(ii) For a system with 60-dimension acoustic vectors, a 1024-mixture UBM, and 400 loading factors, calculate the dimension of matrix ***T***.

(3 marks)

(iii) What is the purpose of WCCN and LDA projections in the system?

(3 marks)

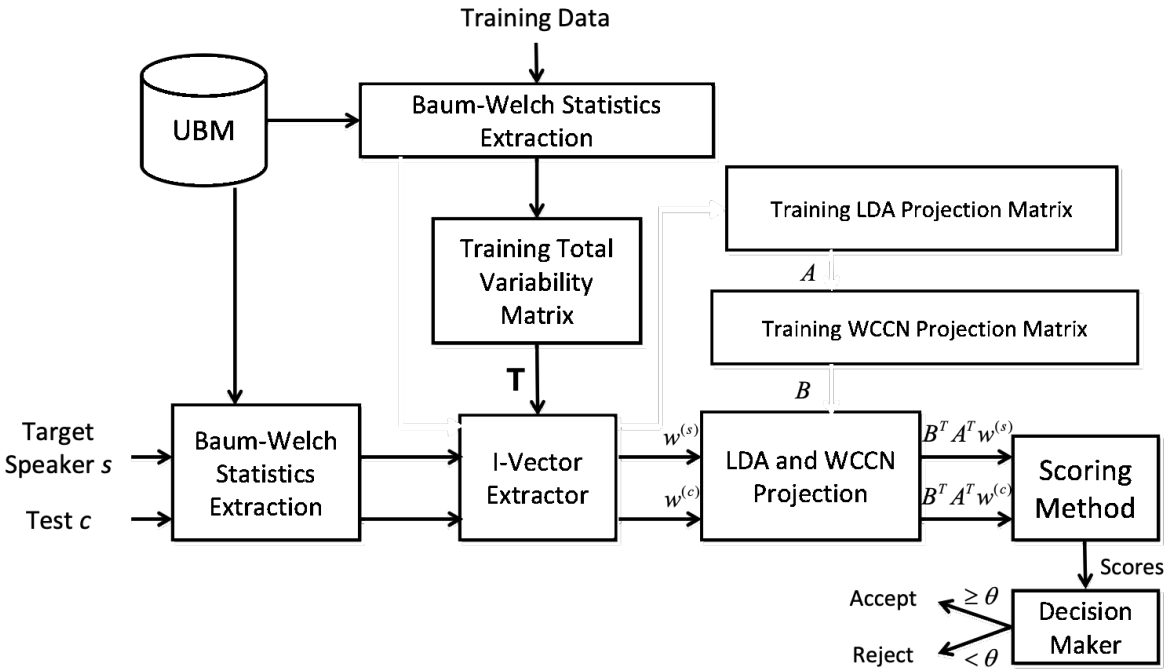


Fig. Q4(b)

-- END --