DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

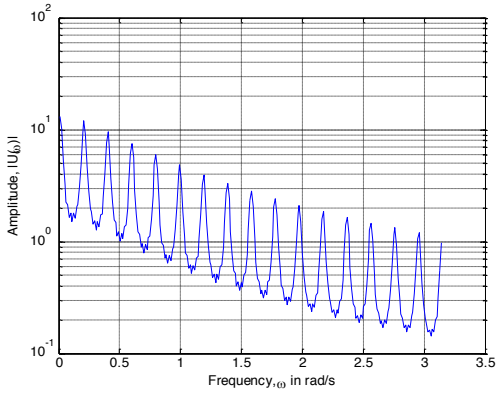
SOLUTION & MARKING SCHEME (2012 / 13)

COURSE: EIE558 YEAR: MSc SUBJECT: Speech Processing and Recognition

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
K: Knowledge A: Application E: Extrapolation	M.W. Mak			

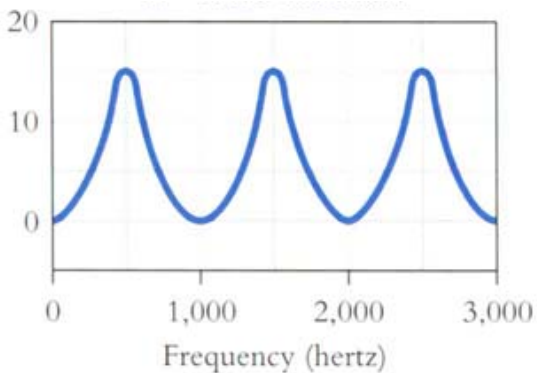
Q1

(a)



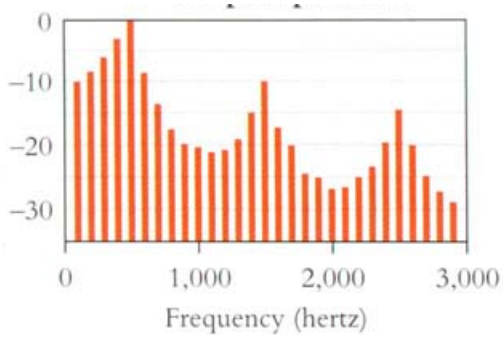
(5 marks, A)

(b)



(5 marks, K)

(c)



(5 marks, AE)

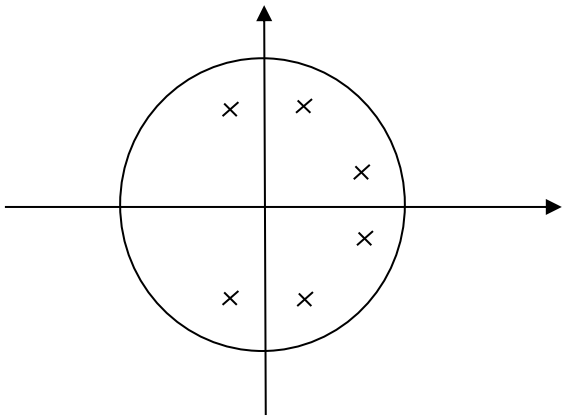
	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
K: Knowledge A: Application E: Extrapolation	M.W. Mak			

(d)

Because the frequency response of G(z) is similar to that of a low-pass filter, the glottal shaping filter introduces a spectral-tilted effect on the speech signal $|S(\omega)|$.

(5 marks, E)

(e)



(5 marks, AE)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2012 / 13)

COURSE: EIE558 YEAR: MSc SUBJECT: Speech Processing and Recognition

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
K: Knowledge A: Application E: Extrapolation	M.W. Mak			

Q2(a)

A frame size of 1 second is too long because the signal within a frame cannot be assumed stationary. For example, the frequency spectrum of the first frame will represent the frequency characteristics of signal in the unvoiced region (0 to 0.6sec) and a voiced region (0.6 to 0.9 sec). Therefore, the resulting spectrum, after applying FFT, will represent a mixture of both voiced and unvoiced speech. This will cause a smearing effect in the spectrogram.

(6 marks, KA)

Q2(b)

(i)

The purpose of smoothing is to reduce the fluctuation of each frequency component against time, which has the effect of reducing musical noise. A possible implementation is

$$\left|\widetilde{X}(\omega,m)\right|^2 = \lambda\left|\widetilde{X}(\omega,m-1)\right|^2 + (1-\lambda)\left|\breve{X}(\omega,m)\right|^2 \qquad \lambda = 0.85$$

(6 marks, A)

(ii)

The right panel corresponds to $\alpha = 1$. This is because α controls the degree of signal attenuation/suppression in the frequency domain and the degree of suppression in the right panel is less than that in the left panel. Specifically, the larger the value of α the bigger the suppression. At a particular frequency ω, if $|B(\omega)|$ is large (i.e., SNR at that frequency is small), a large α will cause a large value in the denominator of $H_s(\omega,m)$, causing large suppression. On the other hand, if α is small or there is no noise at ω (i.e., $|B(\omega)|$ is small), $H_s(\omega,m)$ approaches 1, leading to no suppression.

(6 marks, E)

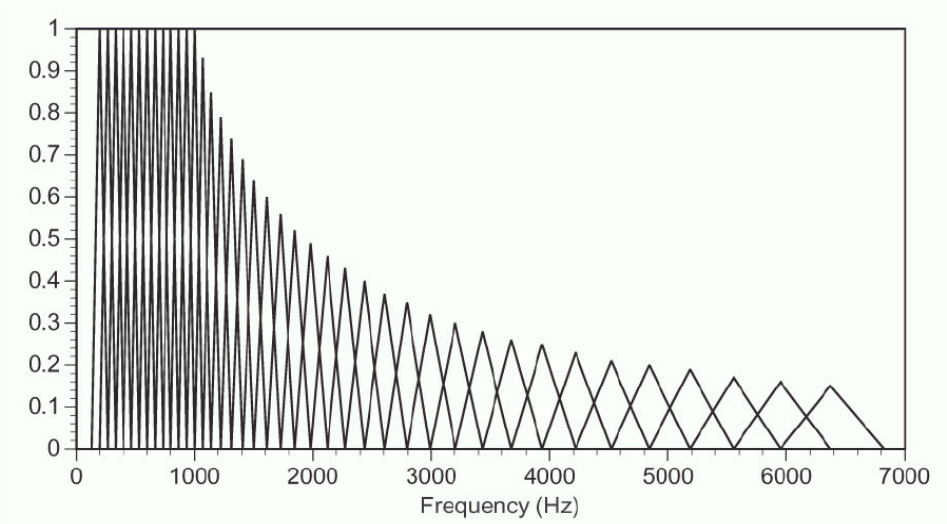
Q2(c)

(i) M = 20 ~ 26

(3 marks, A)

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
K: Knowledge A: Application E: Extrapolation	M.W. Mak			

(ii)



(4 marks, K)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2012 / 13)

COURSE: EIE558 YEAR: MSc SUBJECT: Speech Processing and Recognition

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
K: Knowledge A: Application E: Extrapolation	M.W. Mak			

Q3(a)

(i) Solid circles correspond to male speaker. This is because for the same vowel, male speakers have lower formant frequency.

(5 mark, K)

(ii) The formant frequencies of male speakers are lower than the female counterparts because the vocal tracts of male speakers are typically longer.

(5 marks, KA)

(iii)

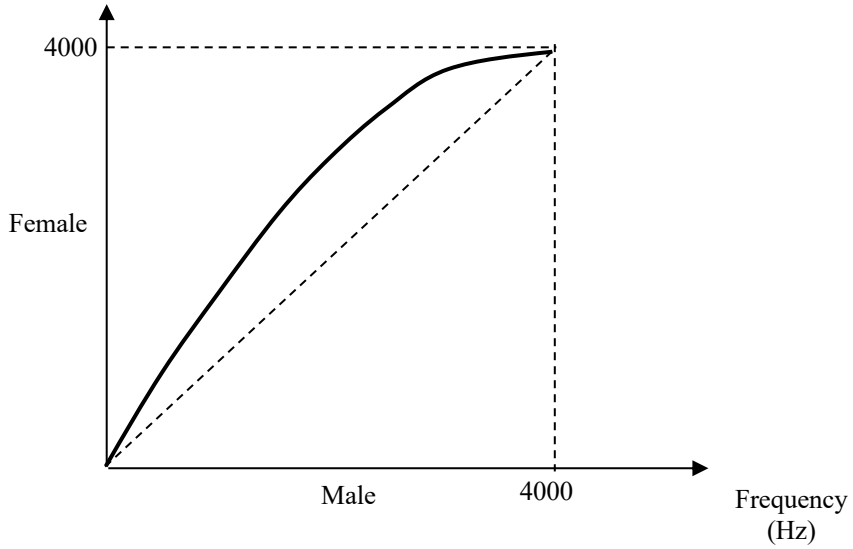


Fig. Q3(b)

(6 marks, AE)

Q3(b)

(i)

$P(z) = \frac{1}{1 - bz^{-D}}$, where b is the degree of voicing and D is the pitch period or integer multiple of pitch period.

(3 marks, K)

(ii) P(z) is responsible for producing the periodicity of voiced speech.

(3 marks, A)

(iii)

If P(z) = 1, the synthetic speech $\hat{s}(n)$ will become very noisy and sound like whisper.

(3 marks, E)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2012 / 13)

COURSE: EIE558 YEAR: MSc SUBJECT: Speech Processing and Recognition

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
K: Knowledge A: Application E: Extrapolation	M.W. Mak			

Q4(a)

(i) For long utterance, α_j will be close to 1 because n_j will be large. As a consequence, the resulting target GMM will be almost dependent on the enrollment data. For short utterance, α_j will be close to 0 because n_j will be small. As a consequence, the resulting target GMM will be almost dependent on the background GMM.

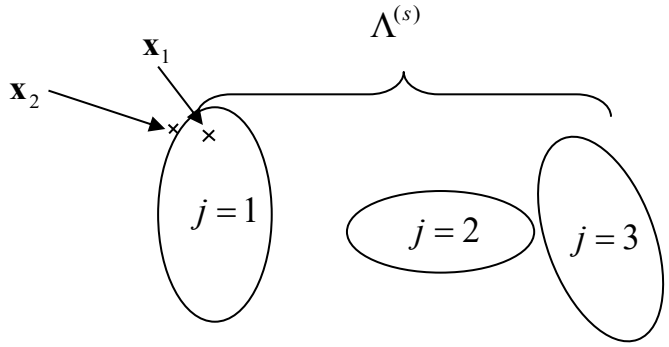
(6 marks, KA)

(ii) When $r = 0$, $\mu_j^{(s)} = E_j(X)$, which depends entirely on the enrollment data. Therefore, Eq. Q4-1 is equivalent to the maximum-likelihood solution of GMM.

(4 marks, AE)

(iii) $\gamma_1(1), \gamma_1(2), \gamma_2(1), \gamma_2(2), \gamma_3(1), \gamma_3(2)$

(3 marks, E)



(3 marks, E)

Q4(b)

(i) The total variability matrix ***T*** defines the subspace in which the GMM-supervectors can vary. For ***T*** with 400 loading factors, the subspace has 400 dimension.

(3 marks, K)

(ii) Dimension of ***T***: $(60)(1024) \times 400 = 61440 \times 400$.

(3 marks, A)

(iii) Because ***T*** is estimated by using the utterances of many speakers and these utterances may be subject to different kinds of channel distortions, the subspace defined by ***T*** represents both speaker variability as well as channel variability. WCCN and LDA are to suppress the channel variability.

(3 marks, E)