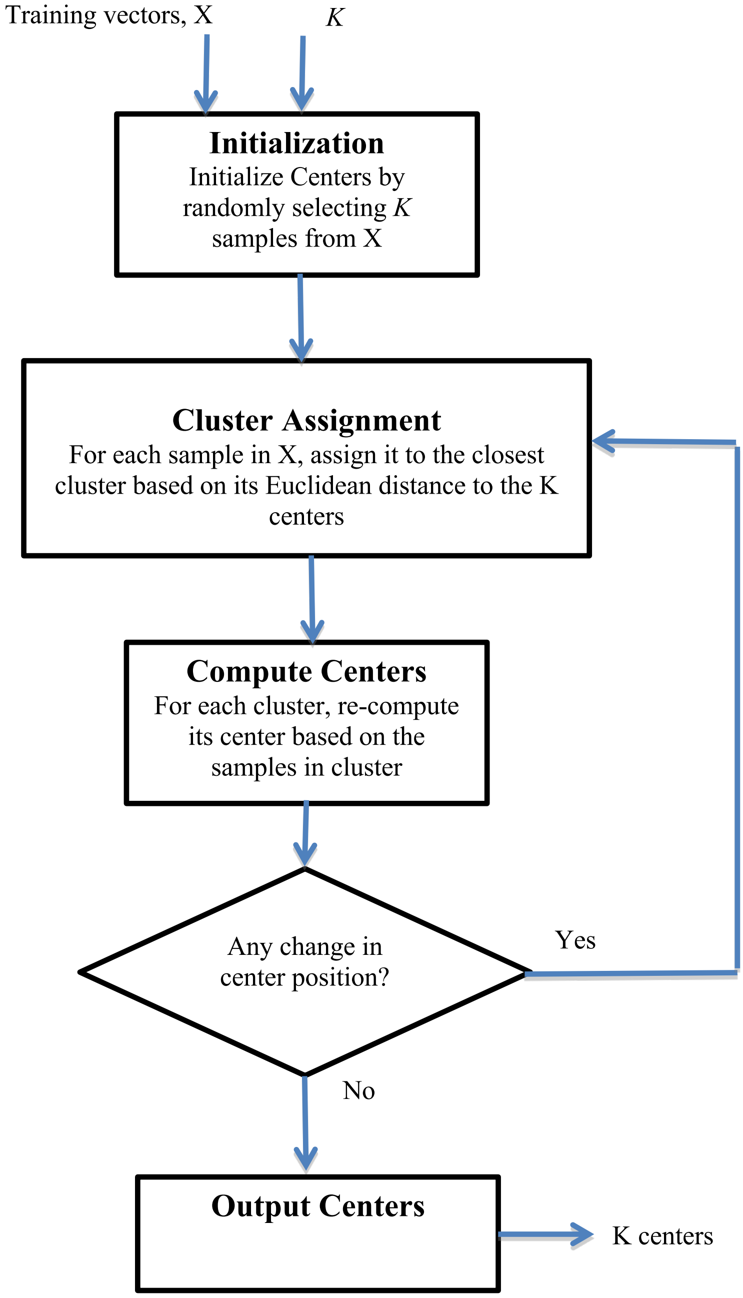
DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF ______EIE__________

SOLUTION & MARKING SCHEME (2016/ 2016)

COURSE: EIE4105 YEAR: 3/4 SUBJECT: Multimodal Human Computer Interaction Technology

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

Q1(a)



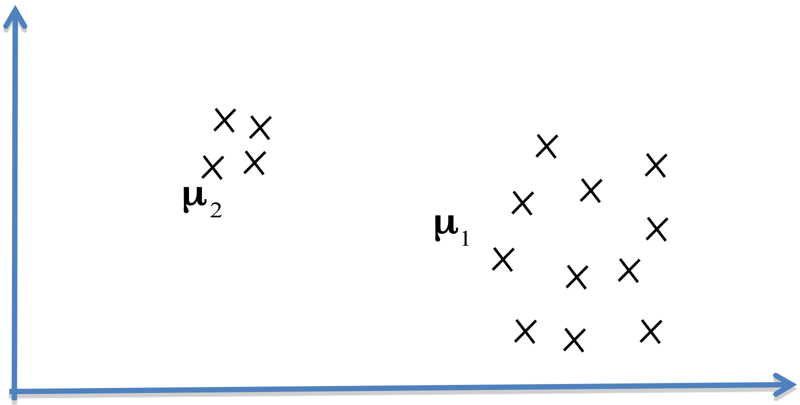
Training vectors, X
K
Initialization
Initialize Centers by randomly selecting *K* samples from X
Cluster Assignment
For each sample in X, assign it to the closest cluster based on its Euclidean distance to the K centers
Compute Centers
For each cluster, re-compute its center based on the samples in cluster
Any change in center position?
Yes
No
Output Centers
K centers

(7 marks, KA)

COURSE: EIE4105 YEAR: 3/4 SUBJECT: Multimodal Human Computer Interaction Technology

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

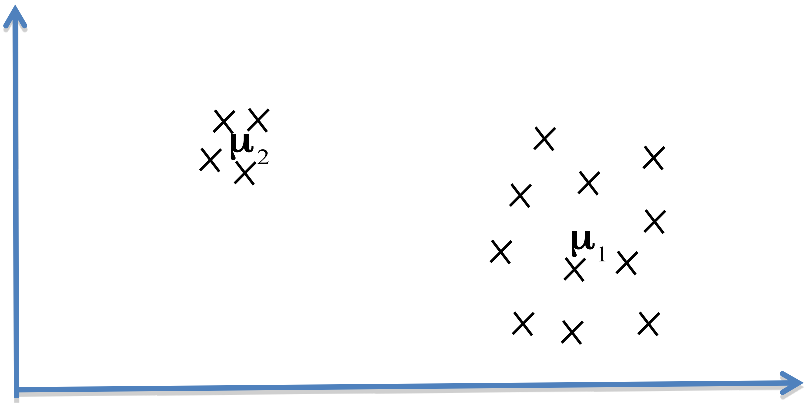
Q1(b)(i)



After 1st iteration

(5 marks, AE)

Q1(b)(ii)



After 2nd iteration

(5 marks, AE)

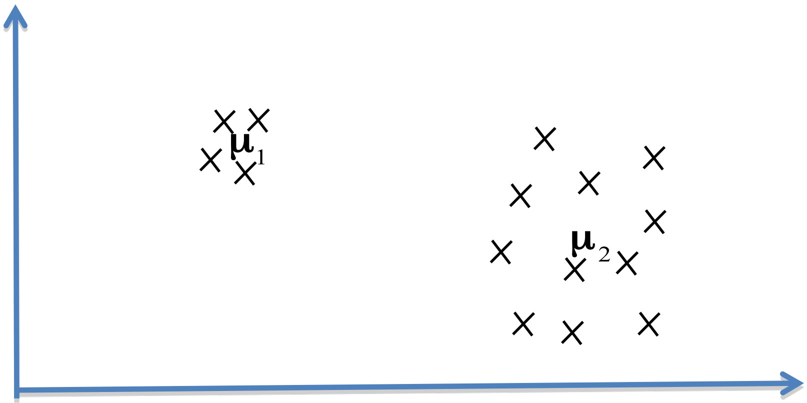
DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2016/ 2016)

COURSE: EIE4105 YEAR: 3/4 SUBJECT: Multimodal Human Computer Interaction Technology

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

Q1(c)
(i)



(2 marks, K)

(ii)

$$\lambda_1 = 4/15 = 0.267$$

(3 marks, AE)

(iii)
Determinant of Σ_2 is larger.

(3 marks, AE)

Reason (not required): The variances in both dimension in Cluster 2 are larger than that of Cluster 1

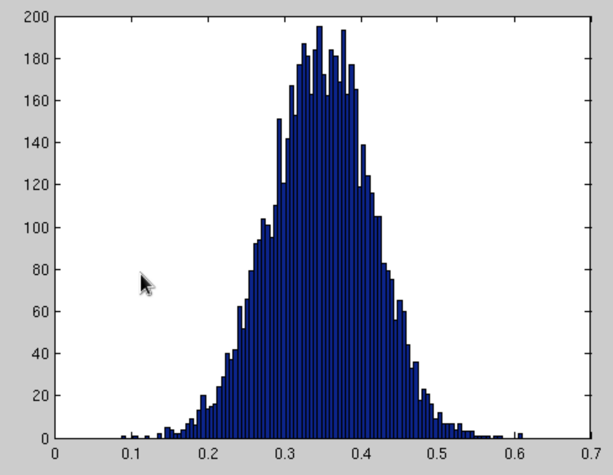
DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2016/ 2016)

COURSE: EIE4105 YEAR: 3/4 SUBJECT: Multimodal Human Computer Interaction Technology

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

Q2 (a)(i)



(4 marks, K)

Q2(a)(ii)

We first convert the training images into vectors by stacking the pixels, which result in 5000 training vectors per digit. Then, we train 10 Gaussian models (density function) with mean vectors and covariance matrices $\Lambda = \{\boldsymbol{\mu}_i, \Sigma_i\}_{i=0}^{9}$, one for each digit. A Gaussian classifier is formed by combining the outputs of these 10 Gaussian PDFs and the prior of the 10 digits.

(4 marks, KA)

Q2(a)(iii)

Given a test vector **x**, the Gaussian classifier determines the class label of **x** by the following formula:

$$l(\mathbf{x}) = \underset{i}{\operatorname{argmax}} \left\{ \begin{array}{ll} P(i \text{ is even}) p(\mathbf{x} \mid \mu_i, \boldsymbol{\Sigma}_i) & i \text{ is even} \\ P(i \text{ is odd}) p(\mathbf{x} \mid \mu_i, \boldsymbol{\Sigma}_i) & i \text{ is odd} \end{array} \right\}$$

$$= \underset{i}{\operatorname{argmax}} \left\{ \begin{array}{ll} \tfrac{2}{3} p(\mathbf{x} \mid \mu_i, \boldsymbol{\Sigma}_i) & i \text{ is even} \\ \tfrac{1}{3} p(\mathbf{x} \mid \mu_i, \boldsymbol{\Sigma}_i) & i \text{ is odd} \end{array} \right\}$$

where $p(\mathbf{x} \mid \boldsymbol{\mu}_i, \Sigma_i)$ is the likelihood of **x** given the Gaussian model of Digit i. The prior probability is obtained as follows:

$$P_e = 2P_o \quad \text{and} \quad P_e + P_o = 1$$
$$\Rightarrow 2P_o + P_o = 1$$
$$\Rightarrow P_o = \tfrac{1}{3}$$
$$\Rightarrow P_e = \tfrac{2}{3}$$

(8 marks, AE)

Q2(a)(iv)

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

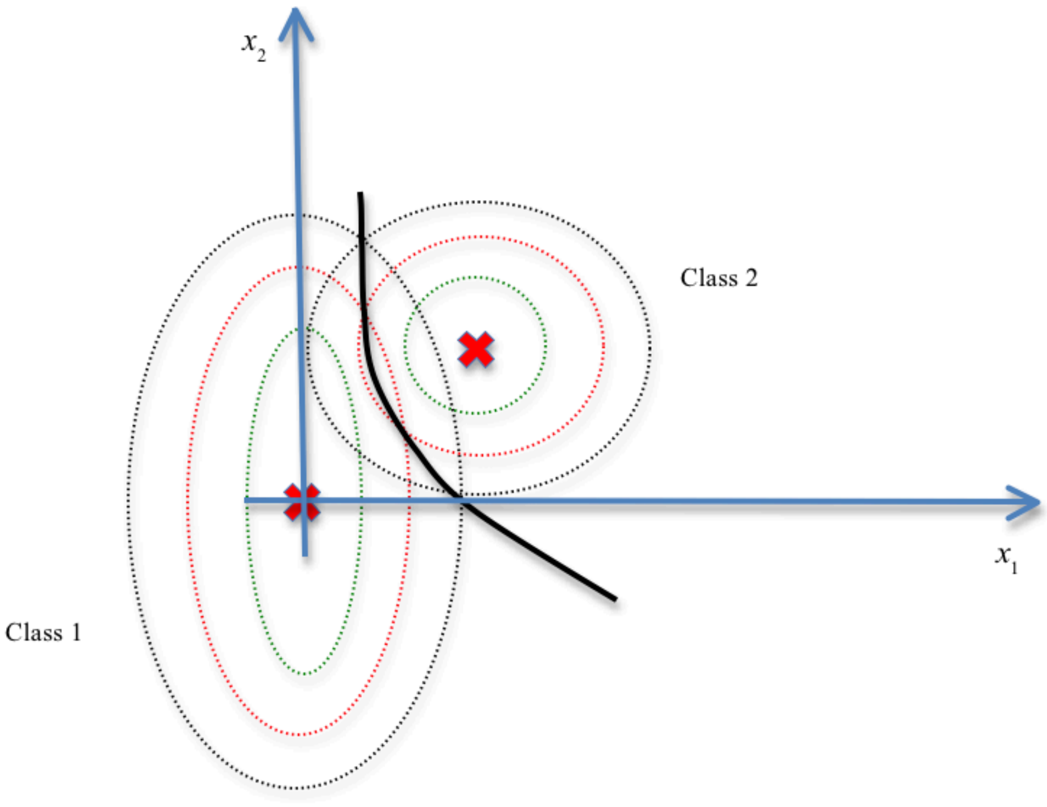
Q2(a)(iv)
No.

(2 marks, A)

Because 100 samples could not produce a covariance matrix with valid inverse.

(2 marks, AE)

Q2(b)



(5 marks, AE)

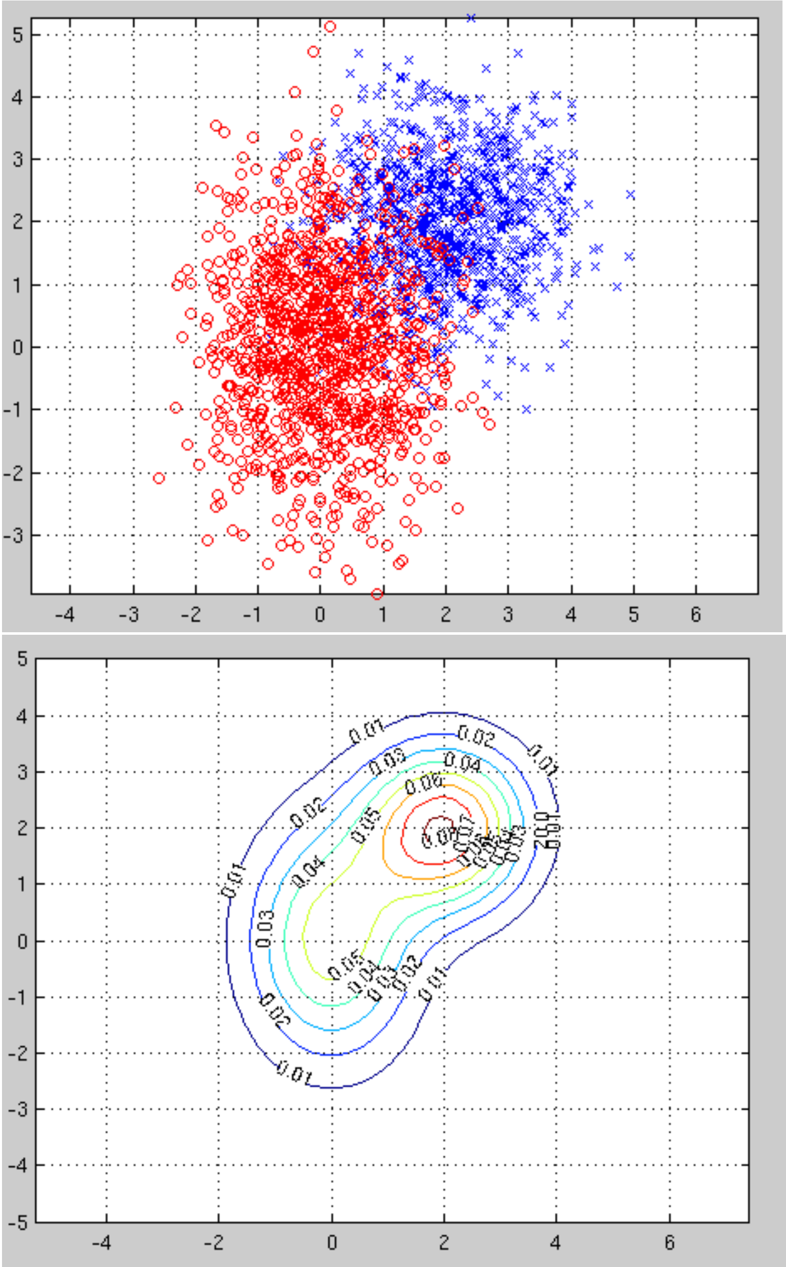
DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2016/ 2016)

COURSE: EIE4105 YEAR: 3/4 SUBJECT: Multimodal Human Computer Interaction Technology

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

Graphs produced by matlab [Students do not need to draw these figures]

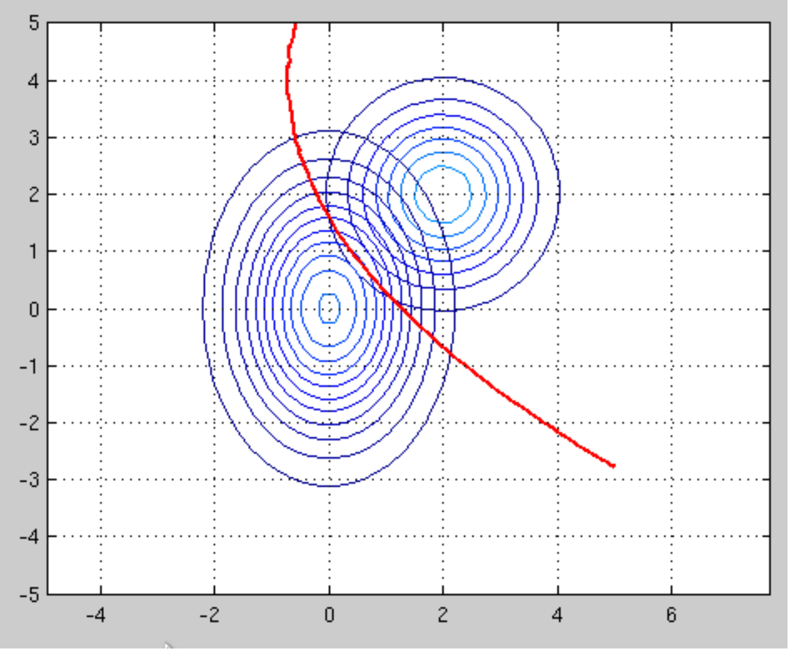


DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2016/ 2016)

COURSE: EIE4105 YEAR: 3/4 SUBJECT: Multimodal Human Computer Interaction Technology

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			



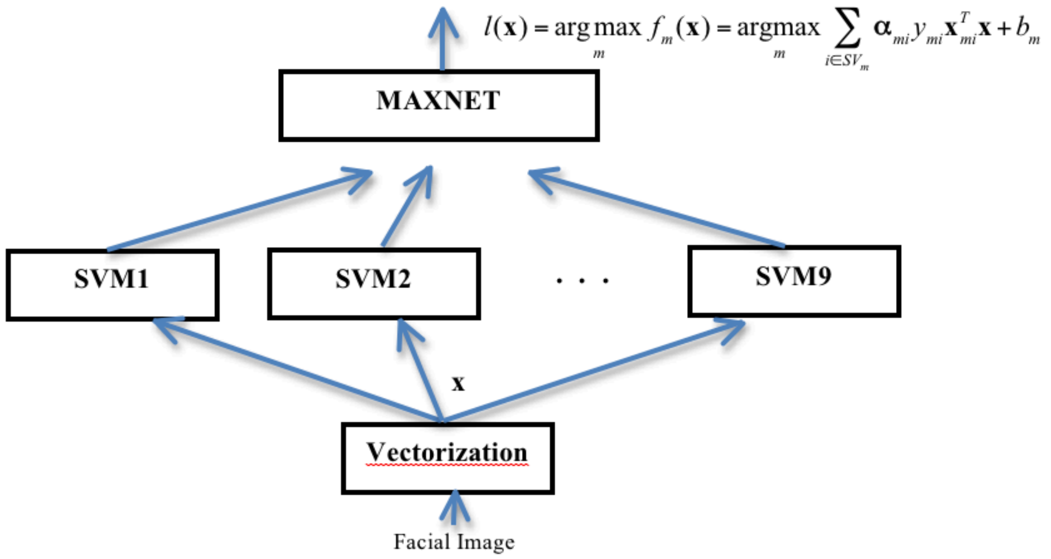
DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2016/ 2016)

COURSE: EIE4105 YEAR: 3/4 SUBJECT: Multimodal Human Computer Interaction Technology

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

Q3(a)(i)
We first convert each facial image into a vector by stacking the pixels. Then, we train 10 linear SVMs, one for each person. Each SVM is trained to classify one person against the other 9. For example, the *p*-th SVM is trained to produce a +1 if the training images from Person *p* are presented, whereas it is trained to produce a -1 if the images from the other persons are presented. After having these 10 SVMs, we combine their outputs using a MAXNET where an unknown image will be classified as Person *p* if the *p*-th SVM produces the largest output.



$$l(\mathbf{x}) = \arg\max_{m} f_m(\mathbf{x}) = \operatorname{argmax}_{m} \sum_{i\in SV_m} \boldsymbol{\alpha}_{mi} y_{mi} \mathbf{x}_{mi}^T \mathbf{x} + b_m$$

(10 marks, KA)

(a) (ii)
A linear kernel should be used.

(2 marks, K)

Because the number of training vectors per SVM is only 200, which is much smaller than the feature dimension (360 x 260 = 93600).

(3 marks, A)

(b)(i) 199

(2 marks, E)

(b)(ii) Fig. Q3(a) corresponds to the eigenfaces with the largest eigen value because it contains more facial features than Fig. Q3(b).

(4 marks, AE)

(b)(iii). Project the mean-subtracted facial vectors to the 5-dimensional space defined by the top-5 eigenfaces. Then, use SVMs or LDA to classify the 5-dim projected vectors.

(4 marks, A)

Q4(a)

(i)

The weights of deep neural network can be **initialized** by a number of RBMs stacked on top of each other. These RBMs can be trained by using contrastive divergence in a layer-wise basis. More specifically, an RBM with the same number of visible units as the number of inputs of the DNN is firstly trained. Then, another RBM with the number of visible units equals to the number of hidden units in the first RBM is trained, using the hidden nodes' outputs of the first RBM as its inputs. The second RBM is then stacked on top of the first RBM. This process is repeated until a desirable number of layers is reached.

(6 marks, KA)

(ii)

Because the error gradients with respect to the weights in the bottom layers are almost zero (the so-called vanishing gradient effect), the backpropagation algorithm is not able to train the weights in the bottom layers. RBMs provide a good initial weights for these layers so that even though the weights in the bottom layers can only have very minor adjustment, the whole DNN can still achieve its goal.

(4 marks, K)

(b)

Using Bayes rule, we can express the posterior in terms of likelihood x prior:

$$\text{DNN}_k(X) = P(s_k \mid X) = \frac{p(X \mid s_k)P(s_k)}{P(X)} \propto p(X \mid s_k)P(s_k)$$

where $\text{DNN}_k(X)$ is the output of the *k*-th output node subject to the input X and $P(s_k)$ is the prior probability of the phone state. Then,

$$p(X \mid s_k) \propto \frac{\text{DNN}_k(X)}{P(s_k)}$$

The prior $P(s_k)$ of HMM states can be obtained from the frequency of occurrences of the corresponding states in the forced alignment during DNN fine-tuning (Students are not required to know this).

(5 marks, E)

(c)(i)

The data should contains the utterances of many (several hundreds) speakers to maximize the acoustic diversity of the general population. The EM algorithm is used to trained the UBM. It can be initialized by using K-means.

(4 marks, K)

(ii)

When the enrollment is very long, $\alpha_j \rightarrow 1$ so that the GMM means depend almost on the enrollment utterance. This is reasonable because when there are many acoustic vectors, we should believe our observations.

When the utterance is very short, $\alpha_j \rightarrow 0$ so that the GMM means are almost the same as the UBM's means. This is reasonable because when there are not many acoustic vectors, we better believe the prior, i.e., the UBM means.

(6 marks)

-- END OF PAPER --