Problem 1

1. Describe local path and global path constraints in word spotting using dynamic time warping (Hint: Refer chapter 11 of Deller's book)
2. Give a step-wise description of the
   (a) K-means algorithm
   (b) LBG algorithm, for delivering VQ codebooks in speech recognition.

Solution.

1. DTW involves the minimisation of distance between reference and test vectors, with warping of the time samples of test signal to the reference one. Let $w(n) = m$ be the path alignment (warping) function.

$$D = \sum_{n=1}^{N} \tilde{d}(T(n), R(w(n)))$$

Assuming word endings are known accurately, i.e. $w(1) = 1$ and $w(N) = M$,
We proceed with Itakura path constraints, i.e. the local path constraints are

$$0 \leq w(n) - w(n-1) \leq 2$$

$$w(n) = w(n-1) \text{ } \textit{iff } w(n-1) - w(n-2) > 0$$

Such local paths guarantee that average slope of warping function lies between $\frac{1}{2}$ and 2 and path monotonicity.

Global constraints include

$$m_L(n) \leq m \leq m_H(n)$$

where
$m_L(n) : min\{2(n-1)+1, M - \frac{1}{2}(N-n), M\}$
$m_H(n) : max\{\frac{1}{2}(n-1)+1, M - 2(N-n), 1\}$

2. (a) K-means is a clustering algorithm, that's used to group unlabelled data-points, bringing out patterns in the dataset. It works by making k clusters out of the dataset where each point belongs to only one group with similar properties.
   Steps:
   i. Assign k centroids in the vector space of the dataset.
   ii. Group all other points to these centroids' cluster on the basis of least euclidean distance.
   iii. Calculate the new mean of each cluster and assign them to be the new centroids, and repeat step 2 until centroids location converge.

(b) Linde Buzo Grey Algorithm is used to generate a codebook with minimum error from the training set.
Steps:

i. Take an image as input
ii. Decompose image into non-overlapping blocks
iii. Select N vectors arbitrarily
iv. Group and compute those vectors' centroid just like K-means algorithm, until the centroids converge.
v. These representative centroids from each group are gathered to form the codebook.

■

Problem 2

1. Derive the expression for the maximum likelihood estimates for $\hat{\mu}$ and $\hat{\sigma}^2$ for a univariate Gaussian pdf and multivariate case.

2. Read Appendix 14.A (page 752-754) of Thomas Quatieri book and list your thoughts.

Solution.
$\hat{\mu}$ and $\hat{\sigma}^2$ are gaussian pdf parameters that are found via the labelled data already at hand. For univariate Gaussian pdf,

$$\hat{\mu_i} = \bar{x_i} \quad \forall x_i \in i^{th} \text{ group}$$
$$\hat{\sigma_i} = \frac{1}{n}\sum_{j=1}^{n}(x_{i_j} - \bar{x_i})^2$$

For multivariate Gaussian MLE, μ_i is the state mean vector and equal to the Expectation of x_i of that group, and Σ_i is the state covariance matrix representing cross-correlations (off-diagnol terms) and variance (diagnol terms) of the elements of the feature vectors.

MLE is based on the classification of data-points into classes, by using a set of labelled ones and fitting in unlabelled into those classes on the basis of the algorithm. GMMs (Gaussian Mixture Models) are essentially linear combination of multiple Gaussian pdfs, so have capability to approximate an arbitrary pdf given enough number of mixture components. The modelling of pdf takes place by apt selection of means, covariance matrices and probability weights of the GMM.

■

Problem 3

Programming Assignment : (MATLAB) In this problem, you investigate the time resolution properties of the Mel-scale and sub-band filter output energy representations. You will use the speech signal speech1 10k in workspace *exl4MI.mat* and functions *make_mel_filters.m* and *make_sub_filters.m*.

1. Argue that the sub-band filters, particularly for high frequencies, are capable of greater temporal resolution of speech energy fluctuations within auditory critical bands than are the Mel-scale filters. Consider the ability of the energy functions to reflect speech transitions, periodicity, and short-time events such as plosives in different spectral regions. Assume that the analysis window duration used in the STFT of the Mel-scale filter bank configuration is 20 ms and is about equal to the length of the filters in the low 1000 Hz band of the sub-band filter bank. What constrains the temporal resolution for each filter bank?

2. Write a MATLAB routine to compute the Mel-scale filter and sub-band filter energies. In computing the Mel-scale filter energies, use a 20 ms Hamming analysis window and the 24 component Mel-scale filter bank from function make mel filters.m, assuming a 4000 Hz bandwidth. In computing the sub-band filter energies, use complex zero-phase sub-band filters from function make sub filters.m. For each filter bank, plot different low- and high frequency filter-bank energies in time for the voiced speech signal speech1_10k in workspace ex14MI.mat. For the sub-band filter bank, investigate different energy smoothing filters p[n] and discuss the resulting temporal resolution differences with respect to the Mel-scale filter analysis.

Solution.

■

MATLAB Code