GSLT 2006

Exercises in speech and speaker recognition

1. Compute the four lowest cepstral coefficients ($C_0 – C_3$) for the following vowel and fricative frequency spectra produced by a 16 channel mel scale filter bank:
[a:] S_j = {75, 78, 86, 91, 82, 79, 83, 78, 70, 72, 73, 71, 74, 71, 66, 52} dB
[s] S_j = {31, 33, 35, 34, 32, 33, 42, 49, 47, 50, 52, 61, 62, 66, 75, 74} dB
The following formula is used: $C_j = \sqrt{\frac{2}{N}} \sum_{i=1}^{N} S_i \cos(j\pi(i - 0.5/N)$
Try to comment on the general difference in these coefficients for the phoneme category distinctions voiced/unvoiced, vowel/voiced consonant, front/back vowel.

2. In a certain recognition system the continuous HMM-models can be based on either monophones or triphones. The number of defined phones is 50. The acoustic input vector consists of energy + 12 cepstral coefficients plus their first and second time derivatives. Each model has three states as defined in the figure. Transition probabilities are stored in a reduced matrix, i.e., elements with zero probability do not occupy memory storage. The probability distribution of the acoustic vector is modelled by an 8-component Gaussian mixture. Each component is specified by a weight value and average and variance values for each acoustic vector element. Each parameter is stored with 4 bytes. Tying is performed at the state level, i.e., certain states share the same acoustic vector probability distribution. The tying rate is 5% for monophones and 20% for triphones. How much computer memory is occupied by complete (all possible units) monophone and triphone libraries?

3. The probability density function of the frame-level observation vector in HMM- and GMM-based speech and speaker recognition systems is modeled as a mixture of Gaussian distributions according to the following formula:
$$b_j(y) = \sum_{m=1}^{M} c_{jm} b_{jm}(y); \qquad \sum_{m=1}^{M} c_{jm} = 1$$
where the emission probability of each mixture component b_{jm} is
$b_{jm}(\mathbf{y}) = N(\mathbf{y}; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})$.

a. What is the likelihood value for one frame of an utterance when matched against a GMM model consisting of three mixture components? For computational simplicity, the input vector is 2-dimensional. The covariance matrix is diagonal.
Frame vector $\mathbf{y}$ = {7,12}
Mixture components:
b_{j1} = {4,3}, {9,21}; b_{j2} = {0,2}, {6,5}; b_{j3} = {11,30}, {18,73}
Component weights: c_1 = 0.4; c_2 = 0.5, c_3 = 0.1

b. Estimate the order of magnitude of the total likelihood of one utterance of 100 frames under the (non-realistic) assumptions that they all have the same value as above and are statistically independent. What is a computer's value range of a 32-bit floating point number? What is the problem and what is normally done to avoid it?
c. How would the likelihood values change if the same distribution was modeled by, e.g. , 30 components instead of 3? Increase or decrease by a factor of 10, or remain unchanged?
d. In what direction will the likelihood value change by increasing the vector size?

4. A spellcheck program might use the following simple dynamic programming (DP) algorithm in order to find the corresponding correctly spelled word in the lexicon instead of the incorrectly spelled word. Which of the two words from the lexicon would be chosen by the algorithm to replace the incorrectly spelled input word? What would the corresponding distances be and what are the spelling errors (deletions, insertions and substitutions as interpreted by the algorithm) in the typed string?
Typed string: "PARRALELL"
Lexicon word A: "PARROT"
Lexicon word B: "PARALLAX"
Lexicon word C: "PARALLEL"

DP algorithm:
Local distance between two characters: d[i,j] = 0 if A[i] = B[j]; else = 1.
Global distance (accumulated): D[i,j] = Min(D[i-1,j], D[i-1,j-1], D[i,j-1]) +d[i,j]
Initialisation: D[0,0] = d[0,0]

5. The topology of a discrete HMM model is normally described by a transition matrix and an observation probability matrix. The transition matrix defines the probability of transitions between the states. The row number and the column number specifies the previous and the following state numbers, respectively, and the value at this position is the probability of a transition between the two states between two time observations (samples).
In the observation probability matrix, the row number defines the state and the column number the index of the observation symbol. The value at each coordinate is the probability of observing the corresponding symbol.
A speech signal consisting of connected digits is described by one acoustic variable which is quantised to eight discrete values, ranging from 1 through 8.
Our Markov model has as transition matrix (rows: previous state nbr, columns: next state nbr)

$$\mathbf{A} = \begin{bmatrix} 0.5 & 0.5 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.5 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.5 & 0.4 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.5 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.5 & 0.5 \\ 0.2 & 0.0 & 0.0 & 0.4 & 0.0 & 0.4 \end{bmatrix}$$

and the densities of the observations for each state are described by (rows: acoustic variable value, columns: state nbr)

$$\mathbf{B} = \begin{bmatrix} 0.5 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.5 & 0.5 & 0.2 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.3 & 0.2 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.2 & 0.4 & 0.2 & 0.0 & 0.2 \\ 0.0 & 0.0 & 0.2 & 0.6 & 0.0 & 0.0 \\ 0.3 & 0.0 & 0.0 & 0.2 & 0.3 & 0.2 \\ 0.3 & 0.0 & 0.0 & 0.0 & 0.7 & 0.3 \\ 0.3 & 0.0 & 0.0 & 0.0 & 0.0 & 0.5 \end{bmatrix}$$

Its initial probability vector is
π = (0.3 0.1 0.1 0.3 0.1 0.1)

The observation sequence is
O = 5 5 5 6 6 7 7 5 5 6 6 7 1 1 3 3 4 4

(a) Draw the HMM state diagram (states and transition arcs) corresponding to the transition matrix. Write the corresponding transition probabilities next to all arcs.
(b) How many digits do you think the model describes, and why. Which states belong to each respective digit? Give each digit a label. A, B, etc.
(c) By looking at the model specification and the observation sequence, try to predict the optimal digit sequence.

6. In a speech-based system for time table information retrieval, a person spoke the following question: ”I want to go to Falsterbo on Sunday morning between nine and eleven o’clock.” The system recognised it as: ”I want a goal to Farsta bro Sunday morning at uhhh nine and uhhh seven o’clock”. Use the the DP algorithm given in Exercise 4 to count the word errors, classified into the categories insertion, substitution and deletion. Also compute a word accuracy value of the recognised text.