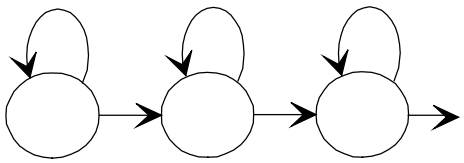
GSLT 2002

Exercises in speech and speaker recognition

1. Compute the three lowest cepstral coefficients 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: $C_n = \sum_{i=1}^{I} S_i \cos(n\pi(i-0.5)/I)$
   Comment on the general difference in these coefficients for the phoneme category distinctions voiced/unvoiced, vowel/voiced consonant, front/middle/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 example in the figure below. 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. 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. In the example below, which of the three 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? Compare the result with that of

a linear comparison (character by character).

Lexicon word A: “ALERT”
Lexicon word B: “ALLERGY”
Lexicon word C: “ALLEGORY”
Input string: “ALERGY

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]

(This exercise is analogous to the problem in some speech recognition systems where a recognised, incorrect phoneme string is compared with the correct phonetic transcriptions of the words in a lexicon.)

4. 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 3 to count the word errors, classified into the categories insertion, substitution and deletion. Also compute a word accuracy value of the recognised text.

5. A telephone banking service installed speaker verification in order to reduce the cost of illegal transactions. At the installation the accept/reject threshold was set based on the rule of minimisation of the overall cost of incorrect client/impostor decisions. The threshold was determined based on the following estimates:
- probability of impostors trying to access the service: 0.0001,
- average cost of illegal transactions during an impostor session: 1000 Euro
- average cost of rejecting a true client:10 Euro
At one later occasion, a person makes an acess attempt. The probability that the claimed client would pronounce this verification utterance is measured to be three times as high as that of anybody else speaking the utterance.
Will this person be accepted or rejected?

6. *(Non-obligatory)*
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 coordinate is the probability that a transition between the two states occurs between two time observations (frames).
In the observation probability matrix, the row number and the column number define the acoustic symbol index and the state index, respectively. The value at each coordinate is the probability of observing the corresponding symbol in this

state.
In a certain recognition system, a speech signal consisting of digits is described by one acoustic variable which is quantised to eight discrete values, ranging from 1 through 8. The 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.
(d)* Find the word sequence with the maximum likelihood using the forward algorithm and provide its likelihood.
(d)* Find the optimal state sequence and the associated word sequence. Determine the likelihood of the sequence.