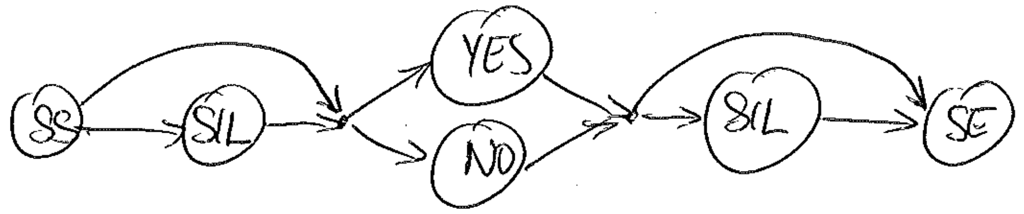
27/2/12

EEEM034 Spring 2011/12 examination solutions

1(a)(i)



8 nodes
2 yes/no all
2 sil skips

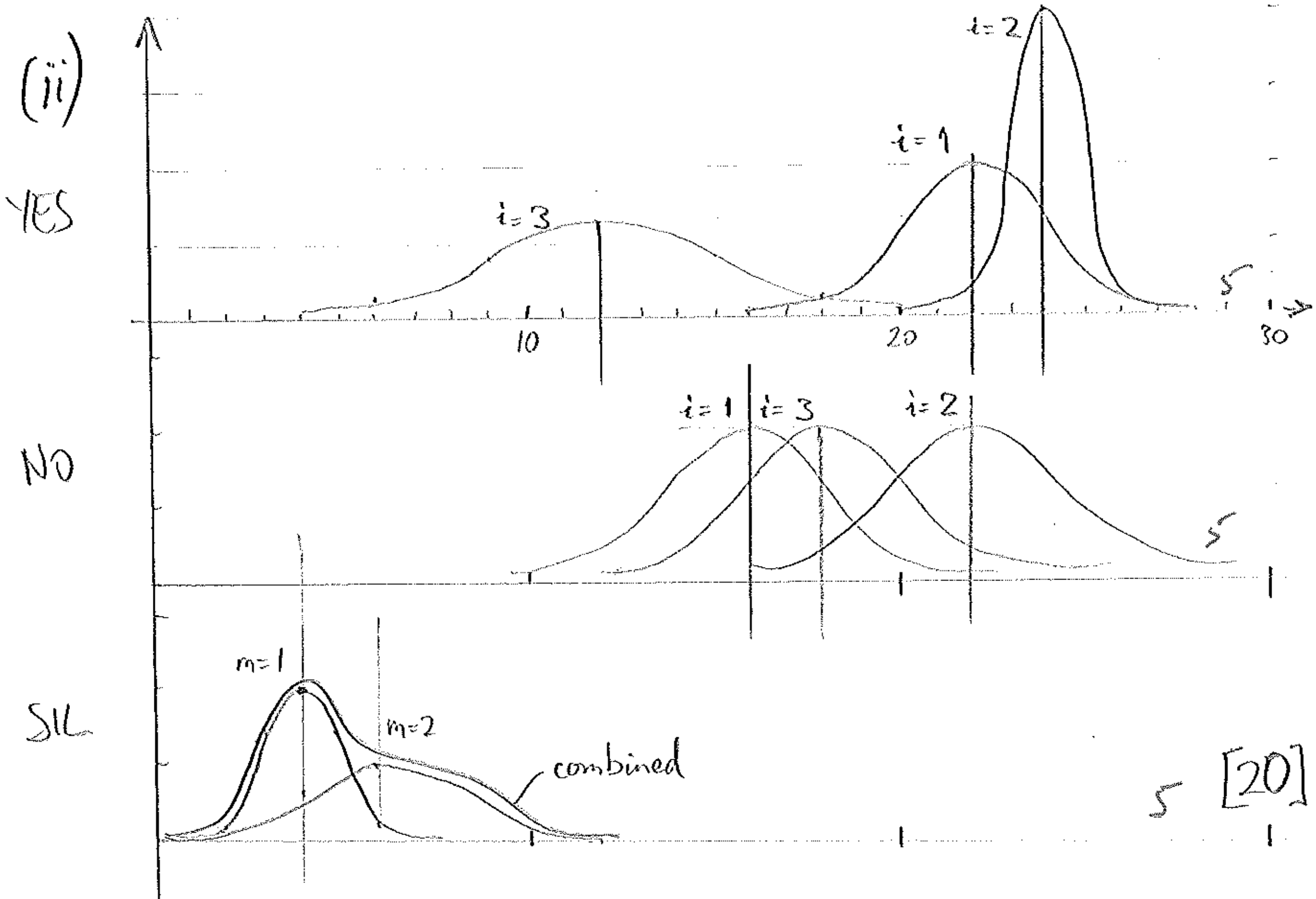
(ii) They provide optional silence at the beginning and end of the utterance, allowing flexibility of the end points. 5

[15]

(b)(i) A mixture can model multiple modes in the data distribution, allow a trade off between number of model parameters and training data, and represent arbitrary pdfs.

5

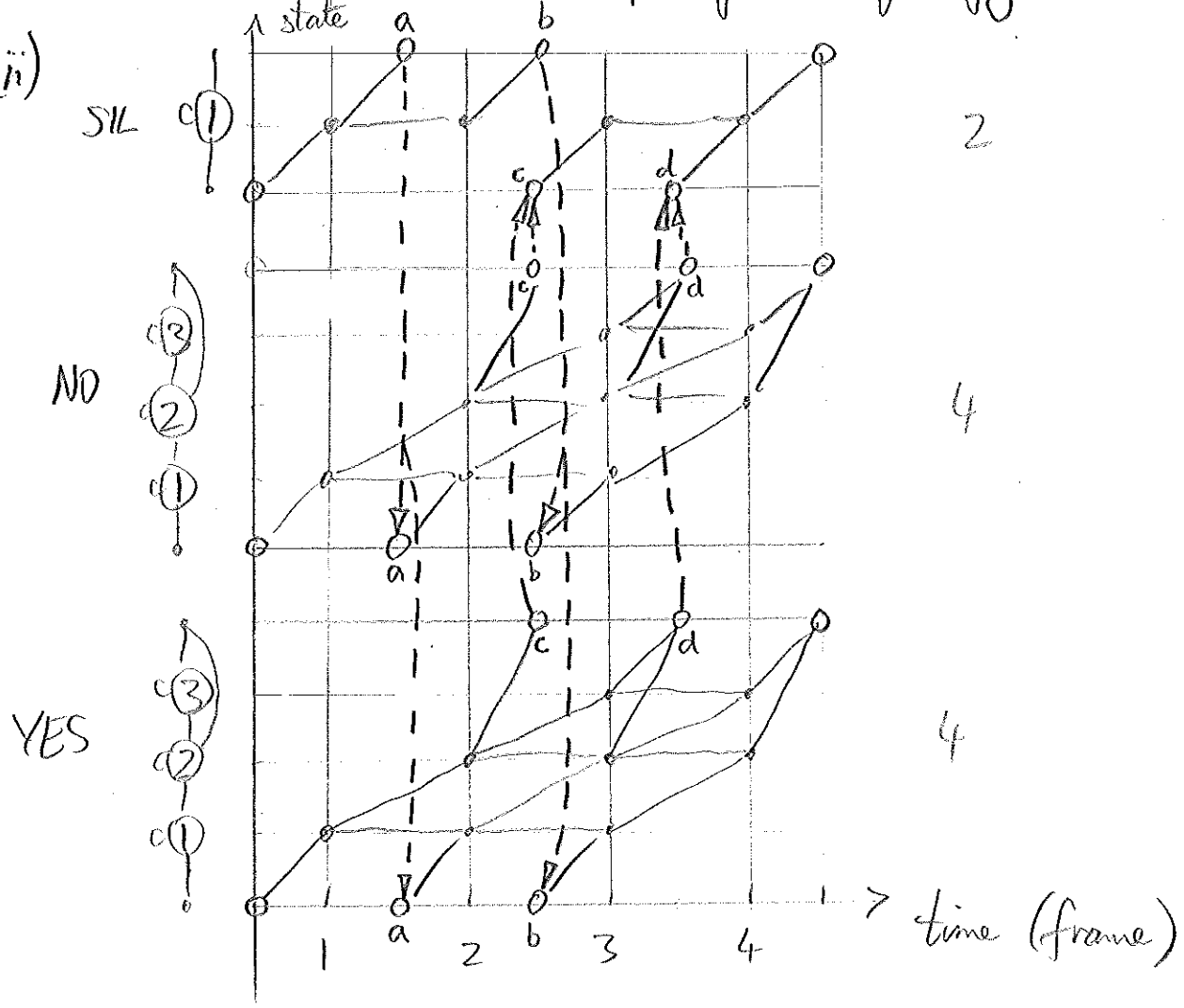
(ii)



[20]

1(c)(i) The YES model has a "left-right" topology. 5

(ii)



2

4

4

The grammar defines links between the models:

a: from SIL to YES and NO
b: from SIL to YES and NO
} 4

c: from YES and NO to SIL
d: from YES and NO to SIL
} 4

(18)

(iii)

$$\delta_t(j) = \max_i \left(\delta_{t-1}(i)\, a_{ij} \right) \cdot b_j(o_t)$$ 8 (15)

where
- i predecessor state
- j current state
- t current time
- δ maximum cumulative likelihood
- a_{ij} state transition probability
- b_j output probability density
- o_t observation 7

1(c)(iv)

By inspection of the pdf sketches in part (b)(ii), we see the states with the maximum likelihood are:

{ YES_1 and NO_2 , YES_1 and NO_2 , YES_2 , YES_3 } 6

Thus, the legal transitions dictate that the YES model 3
fits best with state sequence $X = \{1, 1, 2, 3\}$. 3 (12)

[50]

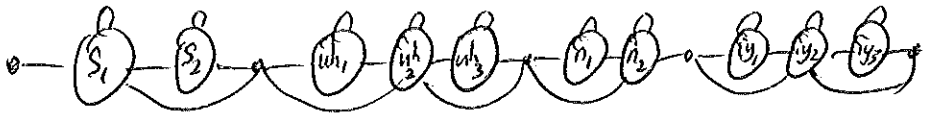
(d) Answers can include any of the following:

- use of a language model
- inclusion of a wildcard model (to catch OOV utterances and extraneous noises)
- greater number of mixture components
- increased frame rate (e.g., to 10ms)
- greater number of states per model
- inclusion of extra state in silence model to deal with transient background noise
- use of longer feature vectors (e.g., logE + 12 MFCCs)
- use of delta and acceleration features
- noise robust feature extraction methods
- speaker adaptation (using earlier dialogues)

8 ea.

[15]

2(a)



[20]

(b)(i) middle: voicing harmonics, strong F2 and F3, low F1 and high F2

(ii) bottom: noise-like spectrum, high-frequency energy [20]

(c) LP concentrates on fitting peaks, whereas cepstrum gives equal importance to peaks and valleys in the spectral envelope (cf. poles and zeros). [10]

(d)(i) impulse train with appropriate period, f_0 { [80, 200 Hz] ♂ / [150, 350 Hz] ♀

(ii) (pink or) white noise [15]

(e)(i) 500 ms ± 200 ms

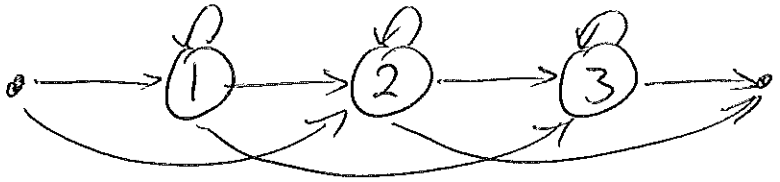
(ii) ~50 ms or ~5 frames [10]

(f) It would produce a sequence of stationary segments, with abrupt transitions, and fail to capture realistic coarticulation from one speech sound to the next. [10]

(g) Interpolate or smooth the spectral envelope from one frame to the next. Use an explicit model of the rate of change (eg. using delta features) or a dynamical model.
or bookwork: VT as acoustical resonator → all-pole system. [15]

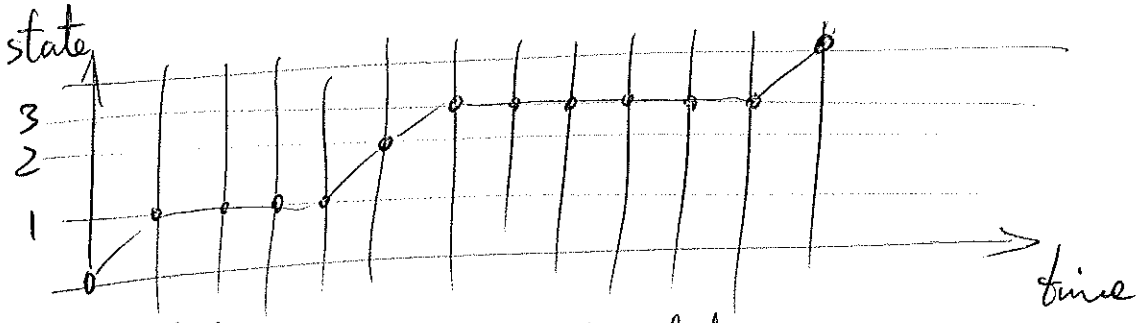
3(a) To determine optimal parameter settings for the models (in a max. likelihood sense), which then act as templates of the recorded speech patterns for recognition. [10]

(b)



[15]

(c)(i)



Entry probability is all given to state 1:

$$\pi^{Vit} = [1 \quad 0 \quad 0]$$

State 1: 3 self loops and 1 transit to state 2

$$a_{1j} = [3/4 \quad 1/4 \quad 0]$$

State 2: 1 transit to state 3

$$a_{2j} = [0 \quad 0 \quad 1]$$

State 3: 5 self-loops and 1 ~~transit~~ exit

$$a_{3j} = [0 \quad 0 \quad 5/6]$$

Exit probability is $\eta = [0 \quad 0 \quad 1/6]^T$

$$A = \left[\begin{array}{c|ccc|c} 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 3/4 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 5/6 & 1/6 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

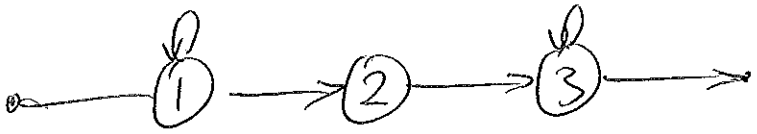
3(c)(ii) Table of observation counts per state:

State	⊕	⊖	⊗	⊘	⊙
1	1	1	1	1	0
2	0	0	1	0	0
3	1	0	0	3	2

$$B^{vit} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1/6 & 0 & 0 & 1/2 & 1/3 \end{bmatrix}$$

[45]

(d) (i)



(ii) All skips have gone, as has state 2's self loop. If no other training examples are used, the legal transitions are more limited, which could lead to poor matches against test utterances. [15]

(e) (i) Viterbi training only considers the best path alignment, i.e. a hard assignment of observations to states at each time frame. B-W uses full forward and backward probs. to give soft assignment via occupation likelihood $\gamma_t(i)$ and transition likelihood $\xi_t(i,j)$.

(ii) By including alternative paths through the trellis, some weight is given to less likely, yet possible transitions and observation assignments. So, BW yields fewer zero probability values, which maintains flexibility in the models and hence allows their generalisation. [15]

Module: EEEM.034 SPEAKER AND SPEECH RECOGNITION
Year: 2011/2012
Examiner: J Kittler
Special Requirements: None
Exam shared with: None
SOLUTION No.: 4

4. (a) Likelihood of a model given a set of outcomes is given by the probability (density) of observing these outcomes given the model. In decision making we can estimate the likelihood for every hypothesis and select the hypothesis that is most likely. In speaker verification the two hypotheses that are compared, given a speech utterance, are whether the claimed identity of a speaker is true or not. The comparison is measured in terms of *likelihood ratio*, involving *client* distribution and *world population* distribution.

[10%]

(b)
- *World population* is a set of all speakers.
- *Impostors* are individuals impersonating a particular client. Potentially, they could include all individuals in the world population, minus the particular client. In practice, this set will be much smaller.
- The main difference between these two groups is that we never build a model for impostors, but we do build a model for the world population.

[10%]

(c) Score normalisation is required to facilitate
- fusion of several multimodal or intramodal biometric experts by an untrained fusion rule
- the use of a fixed threshold $t = 0.5$

[10%]

(d) Given a distribution of scores, with the mean μ_c and standard deviation σ_c of client scores and the mean μ_i and standard deviation σ_i of imposter scores, the score normalisation can be achieved in a number of ways:
- compressing the range to interval $[0, 1]$

$$\hat{s} = \frac{s - s_{min}}{s_{max} - s_{min}}$$

- converting to posterior probabilities

$$\hat{s} = P(s|c) = \frac{p(s|c)}{p(s|c) + p(s|i)}$$

- mapping to designated means

$$\hat{s} = \frac{s - \mu_i}{\mu_c - \mu_i} \quad \mu_c \geq \mu_i$$

[10%]

(e) For the case when the mean of client scores is lower than the mean of imposter scores we have

$$\hat{s} = \frac{s - \mu_c}{\mu_i - \mu_c} \quad \mu_i \geq \mu_c$$

[10%]

(f) i. After the normalisation, the client and imposter distributions will be

$$p(s|client) = \begin{cases} (s+1) & s\epsilon[-1,0] \\ -(s-1) & s\epsilon[0,1] \\ 0 & elsewhere \end{cases}$$

and

$$p(s|i) \begin{cases} 2 & s\epsilon[0.75, 1.25] \\ 0 & elsewhere \end{cases}$$

[15%]

ii. At decision threshold $t = 0.5$ all imposter will be correctly rejected but all client claims with score greater than 0.5 will be falsely rejected. The false rejection rate will be 0.125%.

[15%]

iii. The minimum error threshold separating the scores of clients and impostors is defined by the point t where the aposteriori probabilities for the two classes are equal. Since the prior class probabilities are equal this point will be defined by the condition

$$p(s|client) = p(s|i)$$

Clearly the minimum error decision threshold is $t = 0.75$.

[10%]

iv. The error for the decision threshold derived in d(iii) is given by the area under the tail of the triangular distribution defined by the cutoff point t. The corresponding value is $\frac{1}{32}$.

[10%]