Engineering Tripos Part IIB FOURTH YEAR

MODULE 4F11: SPEECH PROCESSING

Examples Paper 1

1. Considering the example speech of figure 1, what is the fundamental frequency?

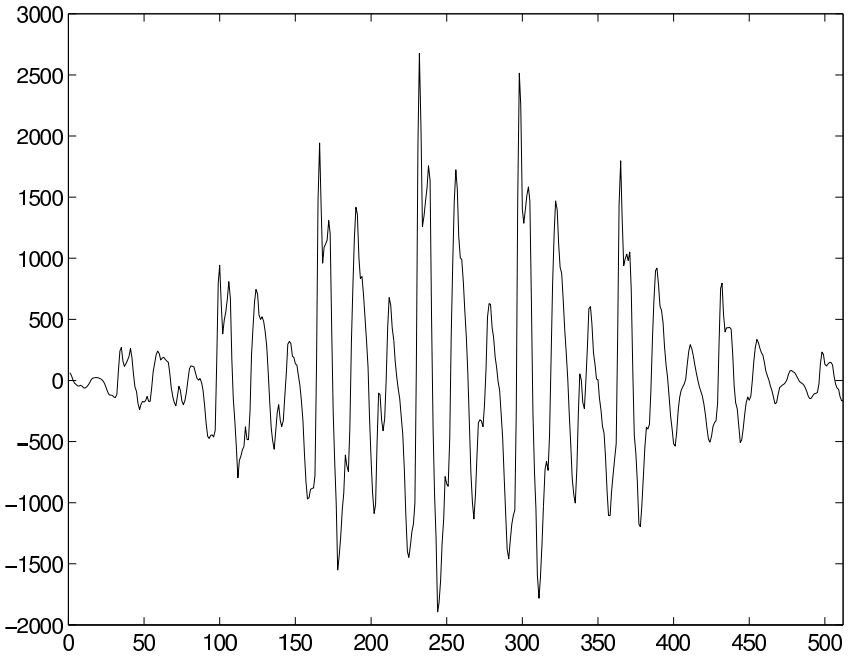

Figure 1: A segment of a vowel sampled at 16 kHz and extracted with a hamming window.

Considering a source filter model of speech production with pulse train excitation, how many cycles of the response due to the main resonance occur within a fundamental period?

What is the frequency of the most prominent formant?

2. Discuss how a spectrum estimate, based on the linear prediction analysis of speech and the frequency response of the vocal tract filter, differs from one obtained by smoothing a discrete Fourier transform of speech.

3. An HMM, with two emitting states is used to model sequences of 1-D feature vectors. The transition matrix is

	2	3	4
1	0.7	0.3	0.0
2	0.6	0.4	0.0
3	0.0	0.8	0.2

where state 1 is the non-emitting entry state and state 4 is the non-emitting exit state. The output probabilities in states 2 and 3 are Gaussian with means of 0.0 and 1.0, respectively. Both states have a variance of 0.5.

For the observation sequence $\mathbf{O} = \{0.2, 0.1, 0.1, 0.5, 0.6, 0.8, 0.7\}$ calculate

(a) the forward probability $\alpha_j(t)$ and the backward probability $\beta_j(t)$

(b) the total probability $p(\mathbf{O}|\lambda)$

(c) the most likely state sequence

(d) the probability $\hat{p}(\mathbf{O}|\lambda)$ corresponding to the most likely state sequence

4. Show that the probability $L_j(t)$ of occupying state j of an HMM, M, at time t given an observation sequence $\mathbf{O}$ is given by

$$L_j(t) = p(\mathbf{O}, x(t) = j|\lambda)/p(\mathbf{O}|\lambda)$$

and calculate this function for each emitting state of the HMM in Q3. Verify that $\sum_j L_j(t) = 1.0$ for each time t.

For the model and observation data given in Q3

(a) use the most likely state sequence to calculate the Viterbi estimate for the means by averaging the observation values aligned with each state

(b) use the occupation probabilities $L_j(t)$ to calculate the Baum-Welch estimate of the means by computing a weighted average over all observation values.

5. An HMM state has a self-transition probability of a. Write down an expression for the probability $P(\tau)$ of occupying the state for τ successive time slots. Hence, show that the expected state occupation duration is given by

$$E[\tau] = \frac{1}{1-a}$$

Sketch $P(\tau)$ and comment on the suitability of HMMs for representing the durational properties of real speech.

6. (a) Write down the equations that describe the *Viterbi algorithm* as used in a hidden Markov model based speech recogniser.

(b) Outline how the Viterbi algorithm is used for continuous speech recognition.

(c) The network in Fig. 2 allows digit sequences of any length to be recognised. An alternative network is to be designed for fixed length digit sequences.

i. Draw a network that constrains the recognised utterances to be exactly three digits long.

ii. Compare the computational complexity for the two networks assuming whole word HMMs are used for each digit. Include the effect of beam-search in your answer.

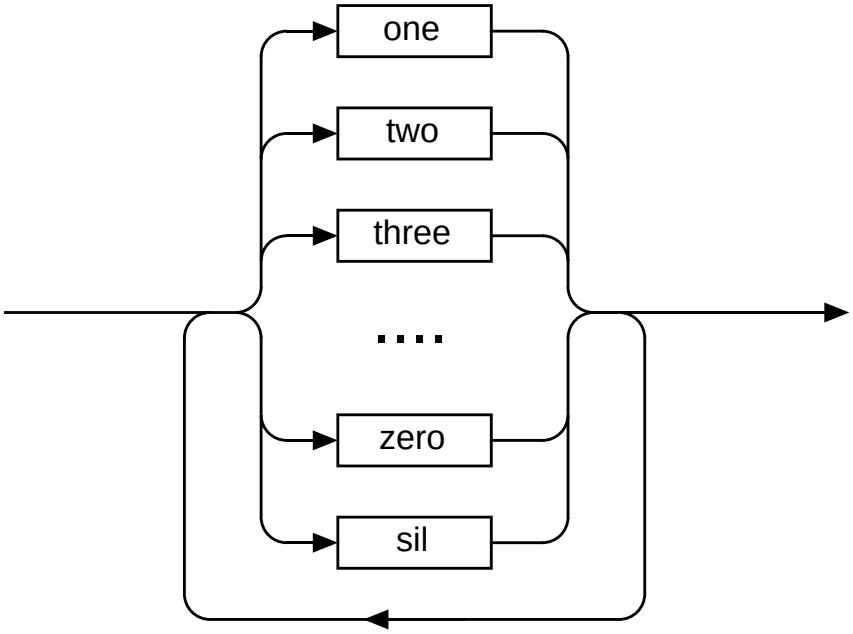

Figure 2: Digit Network

7. (a) What factors influence the choice of basic recognition unit in large vocabulary speech recognition systems?

(b) In an HMM-based recognition system using context dependent phone models, describe how the effects of limited training data can be dealt with by:

i. backing off to more general models

ii. parameter tying (sharing)

8. A large vocabulary speech recogniser based on a standard Viterbi continuous word algorithm uses an N-gram grammar.

   (a) Explain the advantages of a trigram (N=3) over a bigram (N=2) grammar

   (b) Describe how the trigram parameters can be estimated from a training corpus of 5 million words

   (c) Suggest a method by which trigram probabilities can be stored and accessed in the recogniser

9. A speaker independent speech recogniser consists of the following components

   (a) A front-end filter bank analyser consisting of 10 equally spaced identical band pass filters generating a vector of 10 channel amplitudes every 20ms.

   (b) A set of 44 context-independent phone models each of which is a 3 state left-to-right HMM with no skip transitions. Each state consists of a single Gaussian mixture with diagonal covariance matrix.

   (c) A lexicon holding pronunciations for 1000 words. Each pronunciation consists of a single sequence of phonemes.

   (d) A basic Viterbi decoder which utilises a network of word models joined in parallel and placed in a loop. Each word model consists of the sequence of phone models corresponding to the pronunciation in the lexicon.

   Suggest ways in which each component of the above recogniser could be improved. For each case, explain carefully why the modification leads to an improvement in recognition performance and note any consequence for the amount of training data needed and the computational requirements in both the training and recognition phases.

P.C.Woodland
W.J. Byrne

Lent 2014