E4.14
SO16
ISE4.17

IMPERIAL COLLEGE LONDON

DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
EXAMINATIONS 2009

MSc and EEE/ISE PART IV: MEng and ACGI

SPEECH PROCESSING

Corrected Copy

Thursday, 30 April 2:30 pm

Time allowed: 3:00 hours

There are FOUR questions on this paper.

Answer ALL questions.

All questions carry equal marks

Any special instructions for invigilators and information for candidates are on page 1.

Examiners responsible First Marker(s) : P.A. Naylor
Second Marker(s) : P.L. Dragotti

SPEECH PROCESSING

1. a) In the context of speech synthesis,

   i) describe, compare and contrast concatenative synthesis and formant synthesis; [3]

   ii) write a brief explanation showing why it is desirable in concatenative synthesis to be able to control the pitch and duration of speech segments without affecting their formant frequencies. [3]

   b) i) Write down the equation showing the formulation of the output signal of the PSOLA algorithm and define all the terms used. [2]

   ii) A speech signal consists of a diphthong [aɪ] as in the word 'buy' with duration 1 s and a sampling rate of 8 kHz. The 8000 samples are numbered 0 to 7999. The pitch of the diphthong corresponds to a frequency of 100 Hz, and is constant, with pitch marks (i.e. the energy maximum within each pitch cycle) located at sample numbers 40, 120, 200, ..., 7960. Explain the process by which the duration of the speech signal can be increased to 1.2 s without affecting the pitch or formant structure of the speech signal. Include as much detail as possible of the process and refer, using the sample indices, to the samples that are employed. [6]

   iii) It is desired to change the pitch to 124 Hz and increase all formant frequencies by 10% while maintaining the original signal duration of 1 s. Describe the sequence of operations required and state the output sample frequency required.

   Write an expression and evaluate the output sample $y(300)$ in terms of the input samples $s(n)$ for $n = 0, 1, 2, \ldots, 7999$. [6]

2. Features are computed from frames of a speech signal where the frame rate is 100 frames per second and the frame duration is 20 ms. The speech signal is sampled at a sampling frequency of 8 kHz.

a) State any properties of such features that would make them suitable for speech coding in a telecommunications network.

In this context, state the advantages and disadvantages of (i) prediction coefficients, (ii) reflection coefficients and (iii) predictor polynomial roots. [5]

b) State the advantages and disadvantages of Line Spectral Frequencies (LSF) in the context of speech coding for telecommunications.

Given that the vocal tract filter is determined by Linear Predictive Coding (LPC) to have a transfer function $V(z)$, describe in detail how the LSFs would then be obtained from $V(z)$.

Include in your answer a clear mathematical definition of LSFs. [8]

c) Find the LSFs corresponding to

$$V(z) = \frac{1}{1 - 0.75z^{-1} + 0.65z^{-2}}.$$

Draw a labelled plot in the z-plane showing the LSFs and give any relevant comments on the plot. [7]

3. Parts a) and b) of this question refer to the complex cepstrum whereas parts c) and d) refer to the mel-cepstrum.

a) Given that $V(e^{j\omega})$ is the frequency response of a vocal tract transfer function, write down an expression for the corresponding **complex cepstral** coefficients. [3]

b) By considering the z-transform relationship

$$C(z) = \log(V(z)) = \log\left(\frac{G}{A(z)}\right)$$

where the terms have their usual meaning and the predictor $A(z)$ is of the form

$$A(z) = \prod_{k=1}^{p}\left(1 - x_k z^{-1}\right),$$

i) derive the complex cepstral coefficients from the roots of $A(z)$;

ii) show that the complex cepstral coefficients are real;

iii) show that the complex cepstral coefficients form a right-sided sequence. [6]

c) Describe in detail how **mel-cepstrum** coefficients are computed from a speech signal, $s(n)$, and sketch a block diagram showing the steps of the computation. [4]

d) A speech signal, $s(n)$, is observed at the output of a channel, the frequency response of which is $H(j\omega)$.

i) Describe the effect of the channel at each important step in your block diagram and, hence, show how the mel-cepstral coefficients, c_k, are affected by the channel. [4]

ii) Explain why the performance of speech recognition systems operating over the telephone can often be improved by the process known as cepstral mean subtraction. [3]

4. Consider a hidden Markov model (HMM) having S states with a_{ij} representing the transition probability from state i to state j. Next consider a speech utterance consisting of T frames, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ which is compared with the HMM. The output probability density of frame t of the speech signal in state i of the HMM is denoted by $d_i(\mathbf{x}_t)$. The HMM is initialized such that frame 1 of the speech signal is always in state 1 of the HMM.

a) If frame t of the speech signal is in state s of the HMM, write down an expression for the probability density associated with the segment of the alignment path between frame $(t-1)$ in state $(s-1)$ to frame t in state s. [3]

b) Let $B(t,s)$ be defined as the highest probability density that the model generates frames $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$. For the case when $t > 1$, explain fully how $B(t,s)$ can be expressed in terms of $B(t-1,i)$ for $i = 1, 2, \ldots, S$. State the value of $B(1,i)$. [4]

c) Consider the inequality $B(t,s) < k \times \max(B(t,r))$. Explain how this inequality can be used to reduce the computational complexity of the speech recognizer and outline the criteria that should be used in choosing the factor k. [6]

d) A 6-frame utterance is compared with a 4-state Hidden Markov model. Table 1 shows the output probability density of each frame in each state of the model and Figure 4.1 shows the state diagram of the model including the transition probabilities. Determine the value of $B(6,4)$ and the state sequence to which it corresponds given that the computation complexity reduction technique of part c) is employed with $k = 0.2$. At each appropriate step, indicate precisely the effect of pruning. Perform all your calculations to at least six decimal places. Draw and label the resulting alignment lattice. [7]

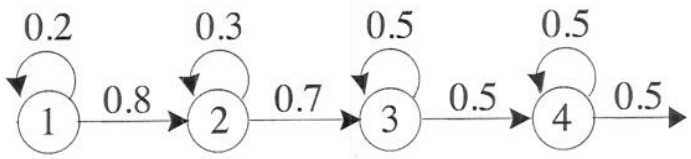

Figure 4.1

	Frame $\mathbf{x}_1$	Frame $\mathbf{x}_2$	Frame $\mathbf{x}_3$	Frame $\mathbf{x}_4$	Frame $\mathbf{x}_5$	Frame $\mathbf{x}_6$
State 1	0.5	0.2	0.5	0.5	0.5	0.5
State 2	0.5	0.7	0.3	0.1	0.5	0.5
State 3	0.5	0.5	0.1	0.6	0.6	0.5
State 4	0.5	0.5	0.5	0.1	0.4	0.4

Table 1

E4.14
Ex 4.17
S016

SPEECH PROCESSING Solutions 2009

1. a) i) Concatenative speech synthesis is usually based on joining together of pre-recorded diphone units (or words or phrases) and employs duration and pitch control to adjust prosody. Formant synthesis employs the source-filter model of speech production by which a mix of noise and period excitation is input to a set of 2nd order resonant filters, connected either in series or cascade. The benefit of using real recordings in concatentative synthesis is the inherent naturalness of the sound. Many synthesizer control parameters are require for formant synthesis. Although it has the potential to achieve high quality synthesis, formant systhesis performance is limited by the quality of the rules governing generation of its parameters.

ii) Natural-sounding speech requires appropriate variations of stress, rhythm and pitch. Since the recording of diphone (or word) units is already a long process, it is not possible to record all possible variations. Thus it is desirable that the synthesiser itself be able to adjust these parameters according to rules derived from observations of natural speech.

b) i)

$$y(r) = \frac{s(m+r-i)w((r-i)/k) + s(n+r-j)w((r-j)/k)}{\sqrt{w^2((r-i)/k) + w^2((r-j)/k)}}$$

$k = \min(j-i, n-m)$ and $w(x) = \frac{1}{2}(1+\cos(\pi x))$

ii) The process here involves (i) windowing at each pitch mark with a raised cosine window such as the Hamming window of length twice the minimum pitch period at that mark, (ii) generating a new set of pitch marks such that every 5th pitch mark is repeated so that the overall length is increased by 20%, (iii) copying windowed speech segments from each old pitch mark to each new pitch mark - the repeated pitch mark taking its segment from the nearest neighbour in the set of original pitch marks or, most commonly, the preceding pitch mark, (iv) performing an overlap-add operation to synthesize the output according to the formula previously given. An illustration with appropriate annotations will be acceptable as a fully satisfactory answer.

iii) We can determine the required transformations by choosing the sample rate to get the right formants, reducing the cycle length to get the right pitch and finally replicating cycles to get the right duration:

	Pitch	Formants	Duration
Sampling Freq.	x1.1	x1.1	x0.909
Remove 9/80 samples from each cycle	x1.127	x1	x0.8875
Repeat every 4th cycle	x1	x1	x1.25
Total	x1.24	x1.1	x1

$$y(300) = \frac{s(253+47)w(47/71) + s(253+24)w(24/71)}{\sqrt{w^2((47/71) + w^2((24/71)}}$$

2. a) Desirable properties include being able to interpolate features across, for example, a missing frame; to be robust to quantization; easy check for stability of the vocal tract filter.

Predictor coefficients: cannot interpolate; not robust to quantization; hard to check for stability.

Reflection coefficients: OK for interpolation; not robust to quantization near unity; easy to check for stability.

Predictor polynomial roots: cannot interpolate; not robust to quantization; easy to check for stability.

b) LSFs are expensive to compute but very good for quantization in transmission systems. They are obtained by first forming symmetric and anti-symmetric polynomials

$$P(z) = A(z) + z^{-(p+1)}A^*z^{*-1}$$

and

$$Q(z) = A(z) - z^{-(p+1)}A^*z^{*-1}.$$

Then if the roots of $P(z)$ are at

$$e^{2\pi jf_i}, \quad i = 1,\ 3,\ \ldots$$

and the roots of $Q(z)$ are at

$$e^{2\pi jf_i}, \quad i = 0,\ 2,\ \ldots$$

then the LSFs are defined as $f_1,\ f_2,\ \ldots$.

c) Now for the case of $A(z) = 1 - 0.75z^{-1} + 0.65z^{-2}$
we have $a_1 = 0.75$ and $a_2 = -0.65$ so

$$P(z) = 1 - 0.1z^{-1} - 0.1z^{-2} + z^{-3}$$

and

$$Q(z) = 1 - 1.4z^{-1} + 1.4z^{-2} - z^{-3}$$

with roots given by

$$P(z) = (z+1)(z^2 - 1.1z + 1) = 0$$

giving $z = -1$ and $z = 0.55 \pm j0.8352$

$$Q(z) = (z-1)(z^2 - 0.4z + 1) = 0$$

giving $z = +1$ and $z = 0.2 \pm j0.98$.

The plot in Figure 2.1 of the roots of $P(z)$ and $Q(z)$ show that they are interleaved and on the unit circle.

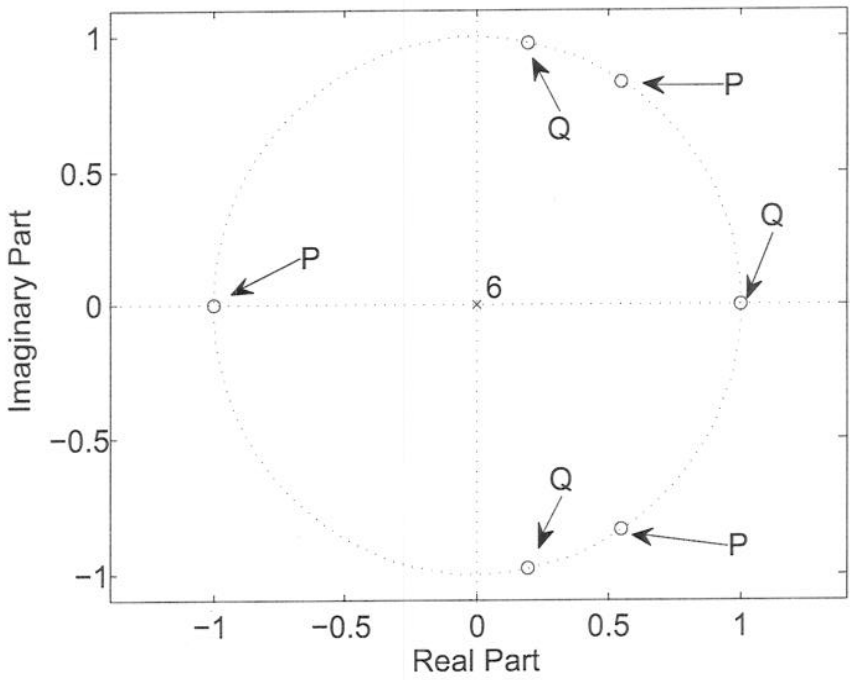

Figure 2.1

3. a) $c_n = \frac{1}{2\pi}\int_{\omega=-\pi}^{\pi} \log\left(V(e^{j\omega})\right) e^{j\omega n} d\omega$

b)

$$\begin{aligned} C(z) &= \log(V(z)) \\ &= \log(G) - \log(A(z)). \end{aligned}$$

Using Taylor's series

$$\begin{aligned} C(z) &= \log(G) - \log(A(z)) \\ &= \log(G) - \sum_{k=1}^{p} \log(1 - x_k z^{-1}) \\ &= \log(G) + \sum_{k=1}^{p} \sum_{n=1}^{\infty} \frac{x_k^n}{n} z^{-n}. \end{aligned}$$

By collecting terms in z^{-n} we find

$$c_n = \begin{cases} 0 & n < 0 \\ \log(G) & n = 0 \\ \sum_{k-1}^{p} \frac{x_k^n}{n} & n > 0 \end{cases}$$

It is necessary to note that since $A(z)$ is the prediction filter formed from the inverse of $V(z)$ and since $V(z)$ is an all pole filter whose poles must lie inside the unit circle for stability, $|x_k|$ must all be < 1, thus the c_n decrease exponentially with n.

c) The computation of the mel cepstrum is an FFT based procedure as illustrated in Figure 3.1.

Segment

Divides the incoming speech into overlapping frames of 10 to 30 ms duration. The duration needs to be long enough to give adequate spectral resolution but short enough to detect speech sounds of short duration. The DFT calculation is more efficient if the segment length is an exact power of 2.

Window

The segment is multiplied by a window function in the time domain (omitting this step is equivalent to choosing a rectangular window). The speech spectrum is convolved with that of the window function. The choice of window function is a compromise between the width of the central lobe (and hence the spectral resolution) and the height of the sidelobes (which cause artefacts in the spectrum).

DFT and $|\cdot|^2$

These steps calculate the energy spectrum of the windowed segment. Filterbank The filterbank smooths the spectrum by taking a weighted average of several adjacent frequency bins. The widths of the filters vary according to the mel scale with narrow filters at low frequencies and wider ones at high frequencies.

log We take the log of the filterbank outputs because the coefficients corresponding to a particular speech sound then follow an approximately gaussian probability density function.

DCT We take the discrete cosine transform of the log energies in order to reduce the correlations between coefficients. This allows us to model the coefficient distributions as independent gaussians or gaussian mixtures.

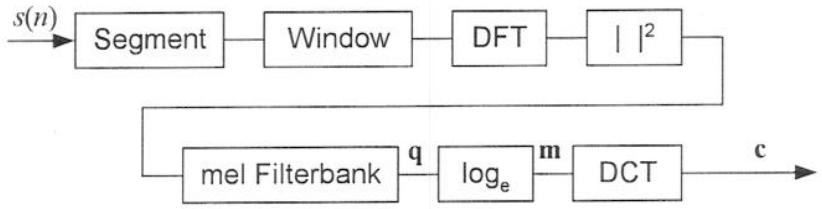

Figure 3.1

d) The speech spectrum is multiplied by the channel frequency response and this product is then convolved with the window spectrum. Providing the channel response is smooth the elements of the **q** vector will be multiplied by a factor that is approximately equal to $|H|^2$ evaluated at the peak frequencies of the corresponding filters. It follows that the log of these factors will be added to the elements of **m** and the DCT of the log added to **c**.

e) The frequency response of telephone microphones is very variable, but within a given telephone call, the response is **constant**. The time-average of **c** will be the cepstrum of the microphone/channel combination added to the average cepstrum of the speech. Subtracting this average from the **c** parameter vectors will remove the effects of the microphone and channel.

The speech recognition system must be trained on speech that has undergone the same processing, i.e. the cepstral mean must also be subtracted during training.

At some frequencies, the channel response may be very low. Any additive noise at these frequencies that is not affected by the channel will worsen the SNR and hence increase the variability of the **c** coefficients. It may be necessary to compensate for this by increasing the variances within the speech model.

4. a) The key point here is to show the product of the transition probability and the output probability.

Segment prob. density = $a_{s-1,s} \times d_s(\mathbf{x}_t)$

b) The best path for $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$ with frame t in state s must have frame $t-1$ in one of the earlier states, state i. Since the sub-path $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1}$ must also be optimum, we must have

$B(t,s) = B(t-1,i) \times a_{is} \times d_s(\mathbf{x}_t$

Since $B(t,s)$ represents the probability density of the best path, we have

$B(t,s) = \max_{i=1,2,\ldots,S} B(t-1,i) \times a_{is} \times d_s(\mathbf{x}_t$.

Since we require frame 1 to be in state 1, $B(1,s) = d_1(\mathbf{x}_1)$ if $s = 1$ and 0 otherwise.

c) The maximization in the above expression involves a great deal of calculation if S is large. We can reduce this by eliminating from consideration, values of i for which $B(t-1,i)$ is small. To achieve this, after calculating $B(t,s)$ for all s, we delete (or prune) all those less than $k \times \max_s B(t,s)$ for some factor $k < 1$.

The choice of k is a compromise. If k is too large, the computational saving will be slight since few states will be pruned. If k is too small then it is possible that the pruning will delete one of the states that in fact lies along the optimum path.

d)

$B(1,1) = 0.5$

$B(2,1) = 0.5 \times 0.2 \times 0.2 = 0.02$

$B(2,2) = 0.5 \times 0.8 \times 0.7 = \mathbf{0.28}$

$B(3,2) = 0.28 \times 0.3 \times 0.3 = \mathbf{0.0252}$

$B(3,3) = 0.28 \times 0.7 \times 0.1 = 0.0196$

$B(4,2) = 0.0252 \times 0.3 \times 0.1 = 0.000756$

$B(4,3) = \max(0.0252 \times 0.7,\ 0.0196 \times 0.5) \times 0.6 = \mathbf{0.010584}$

$B(4,4) = 0.0196 \times 0.5 \times 0.1 = 0.00098$

$bw = 0.0028224$ hence prune B(4,4) and B(4.2)

$B(5,3) = 0.010584.5 \times 0.6 = \mathbf{0.003175}$

$B(6,4) = 0.003175 \times 0.5 \times 0.4 = \mathbf{0.00063504}$

The alignment path is therefore [1,2,2,3,3,4]. The lattice of Fig. 4.1 should clearly show which segments have the highest probability and where segments have been pruned.

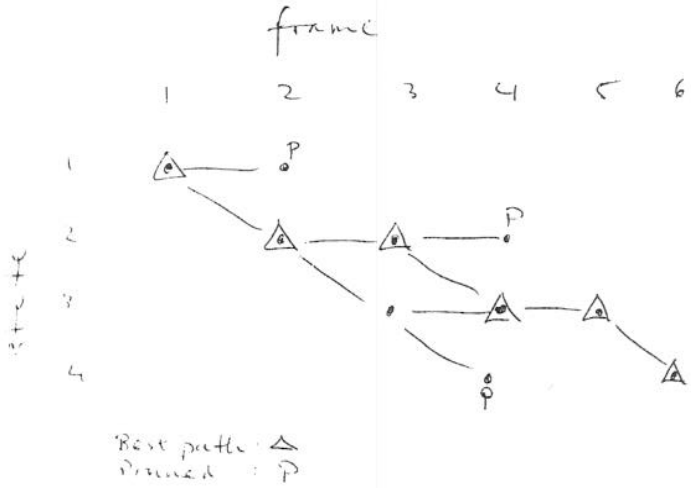

Figure 4.1