E4.14
IE4.17
S016

1. Water Shortage wɔtɜ ʃɔtədʒ

Note here that the aim is to assess overview understanding of phonetic transcription and not detailed knowledge of IPA phonetic alphabet. Hence full marks will be given when good understanding is shown even if errors are present in the transcription.

The glottal stop is a break in the air flow at the glottis. It is used instead of a t in some dialects of English, such as Cockney.

A narrow spectrogram because the frequency resolution is high - high enough to see the pitch harmonics during voicing.

The pitch of the speech can be estimated from the frequency spacing of the harmonics during voiced regions. This is just over 100 Hz and therefore we deduce it is male speech. Full marks will be given if the explanation is correct.

2. (a)

$$-\tfrac{1}{2}\frac{\partial E}{\partial a_i} = \sum_n \left(s(n) - \sum_{k=1}^{p} a_k s(n-k) \right) s(n-i)$$

$$= \sum_n s(n-i)s(n) - \sum_{k=1}^{p} a_k \sum_n s(n-i)s(n-k)$$

$$= r_{i,0} - \sum_{k=1}^{p} r_{i,k} a_k$$

Setting these partial derivatives to zero gives:

$$\sum_{k=1}^{p} r_{i,k} a_k = r_{i,0} \text{ for } i = 1, \ldots, p$$

or in matrix form $\mathbf{Ra} = \mathbf{b}$.

(b) An unstable filter can be made stable by using coefficient reversal. Form a new filter by conjugating the coefficients and putting them in reverse order.

$$H(z) = b_0 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_p z^{-p}$$

$$G(z) = b_p^* + b_{p-1}^* z^{-1} + b_{p-2}^* z^{-2} + \ldots + b_0^* z^{-p} = z^{-p} H^*\left(z^{*-1}\right)$$

G(z) has the same magnitude response as H(z) but a different phase response:

$$G(e^{j\omega}) = e^{-jp\omega} H^*(e^{j\omega})$$

$$\left|G(e^{j\omega})\right| = \left|H(e^{j\omega})\right| \qquad \text{Arg}\left(G(e^{j\omega})\right) = -\text{Arg}\left(H(e^{j\omega})\right) - p\omega$$

(c) The difference between covariance and autocorrelation LPC is in their choice of speech frame, {*F*}. For autocorrelation LPC analysis chose {*F*} to be infinite in extent but bounded by a windowing function, for covariance LPC analysis chose {*F*} to be a finite segment of speech: {*F*} = *s*(*n*) for $0 \le n \le (N\text{-}1)$. Note that in a direct implementation of the covariance method, samples s(n) for $-p+1 < n < 0$ will be required.

- Autocorrelation
  - Requires a windowed signal Þ tradeoff between spectral resolution and time resolution
  - Requires >20 ms of data
  - Has a fast algorithm because F is toeplitz
  - Guarantees a stable filter *V(z)*
- Covariance
  - No windowing required
  - Gives infinite spectral resolution
  - Requires >2 ms of data
  - Slower algorithm because F is not Toeplitz
  - Sometimes gives an unstable filter *V(z)*

(d) The 1200 Hz sinewave should have an amplitude $20\log_{10}(0.2) = -14\ \text{dB}$.

Plot W has the poorest frequency resolution and broadest spectral peaks and is therefore autocorrelation LPC with a window length of 3.5 ms. This gives a frequency resolution of about 2/3.5 kHz = 571 Hz.

Plot Z is better, but the 1200 Hz sinewave has been almost overwhelmed by the sidelobes of the 1 kHz sinwave. This is therefore autocorrelation LPC with a window length of 80 ms.

Plot Y has accurately found the three sinewaves although their amplitudes are not quite right. It has also generated a spurious peak at around 3300 Hz. This is covariance LPC with a window length of 3.5 ms (28 samples) and the white noise has resulted in the false peak.

Finally Plot X is covariance LPC with a window length of 80 ms. The sinewaves have been accurately modelled and the remaining four poles have been used to create a noise floor at –10 dB.

3(a) If we differentiate the expression for E, we get [8]

$$\tfrac{1}{2}\frac{\partial E}{\partial g_k} = \sum_{n=0}^{N-1} e(n)\frac{\partial e(n)}{\partial g_k} = -\sum_{n=0}^{N-1}\big(u(n) - g_k y_k(n)\big)y_k(n)$$

Setting this to zero gives $g_k = \dfrac{\sum_{n=0}^{N-1} u(n)y_k(n)}{\sum_{n=0}^{N-1} y_k^2(n)}$

Substituting this back in the expression for E gives

$$E_k = \sum_{n=0}^{N-1}\big(u(n) - g_k y_k(n)\big)^2 = \sum_{n=0}^{N-1} u^2(n) - g_k\left(2\sum_{n=0}^{N-1} u(n)y_k(n) - g_k\sum_{n=0}^{N-1} y_k^2(n)\right)$$

$$= \sum_{n=0}^{N-1} u^2(n) - g_k\left(2\sum_{n=0}^{N-1} u(n)y_k(n) - \sum_{n=0}^{N-1} u(n)y_k(n)\right)$$

$$= \sum_{n=0}^{N-1} u^2(n) - g_k\sum_{n=0}^{N-1} u(n)y_k(n) = \sum_{n=0}^{N-1} u^2(n) - \frac{\left(\sum_{n=0}^{N-1} u(n)y_k(n)\right)^2}{\sum_{n=0}^{N-1} y_k^2(n)}$$

(b) We can calculate the number of operations as follows:

+/–	*/÷	Calculation
n	n+1	$y_k(n) = \sum_{j=0}^{n} h(j)x_k(n-j)$
½N(N–1)	½N(N+1)	$y_k(n)$ for $n = 0, \ldots, N-1$
N–1	N	$\sum_{n=0}^{N-1} u(n)y_k(n)$
N–1	N	$\sum_{n=0}^{N-1} y_k^2(n)$
N–1	N	$\sum_{n=0}^{N-1} u^2(n)$
1	2	E_k

Hence the totals are:

+/– $\quad N - 1 + K\big(1 + 2N - 2 + \tfrac{1}{2}N(N-1)\big) = 1934395$

*/÷ $\quad N + K\big(2 + 2N + \tfrac{1}{2}N(N+1)\big) = 1998908$

c) With this new relationship we can now write

$$y_k(n) = \sum_{j=0}^{n} h(j)x_k(n-j) = h(n)x_k(0) + \sum_{j=0}^{n-1} h(j)x_{k-1}(n-j-1) = h(n)x_k(0) + y_{k-1}(n-1)$$

+/–	*/÷	Calculation
1	1	$y_k(n)$ for $k>0$
N	N	$y_k(n)$ for $n = 0, \ldots, N-1$ for $k>0$

Thus for $K-1$ codewords we have a computation saving and the computation required is now:

$$+/- \qquad 1934395 - (K-1)\left(\tfrac{1}{2}N(N-1) - N\right) = 185065$$

$$*/\div \qquad 1997884 - (K-1)\left(\tfrac{1}{2}N(N+1) - N\right) = 187174$$

4 (a) Line Spectrum Frequencies: f_i

Advantages: Robust to numerical rounding;
Stability check easy;
Can interpolate;
Vary smoothly in time;
Strongly correlated;
Related to spectral peaks (formants).
Disadvantage: Awkward to calculate

Calculate LSF's from inverse of vocal tract filter

$$A(z) = G \times V^{-1}(z) = 1 - \sum_{j=1}^{p} a_j z^{-j} = 1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_p z^{-p}$$

Form symmetric and antisymmetric polynomials:

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)} A^*(z^{*-1}) \\ &= 1 - (a_1 + a_p) z^{-1} - (a_2 + a_{p-1}) z^{-2} - \ldots - (a_p + a_1) z^{-p} + z^{-(p+1)} \\ Q(z) &= A(z) - z^{-(p+1)} A^*(z^{*-1}) \\ &= 1 - (a_1 - a_p) z^{-1} - (a_2 - a_{p-1}) z^{-2} - \ldots - (a_p - a_1) z^{-p} - z^{-(p+1)} \end{aligned}$$

If the roots of $P(z)$ are at $\exp(j2\pi f_i)$ for i=1,3,... and those of $Q(z)$ are at $\exp(j2\pi f_i)$ for i=0,2,... with $f_i + 1 > f_i \geq 0$ then the LSF frequencies are defined as $f1, f2, \ldots, fp$.

To show that the roots lie on the unit circle, write

$$P(z) = 0 \quad \Leftrightarrow \quad A(z) = -z^{-(p+1)} A^*(z^{*-1}) \quad \Leftrightarrow \quad H(z) = -1$$

$$Q(z) = 0 \quad \Leftrightarrow \quad A(z) = +z^{-(p+1)} A^*(z^{*-1}) \quad \Leftrightarrow \quad H(z) = +1$$

$$\text{where } H(z) = \frac{A(z)}{z^{-(p+1)} A^*(z^{*-1})} = z \prod_{i=1}^{p} \frac{(1 - x_i z^{-1})}{z^{-1}(1 - x_i^* z)} = z \prod_{i=1}^{p} \frac{(z - x_i)}{(1 - x_i^* z)}$$

Set P(Z) and Q(Z) to 0 to find roots, and define H – an arbitrary intermediate transfer function (here the *xi* are the roots of *A(z)=V–1(z)*).

Providing all the *xi* lie inside the unit circle, the absolute values of the terms making up *H(z)* are either all > 1 or else all < 1. Taking | | of a typical term:

$$\left| \frac{(z - x_i)}{(1 - x_i^* z)} \right| > 1 \quad \Leftrightarrow \quad |1 - x_i^* z| < |z - x_i|$$

$$\Leftrightarrow \quad (1 - x_i^* z)(1 - x_i^* z)^* < (z - x_i)(z - x_i)^*$$

$$\Leftrightarrow \quad (1 - x_i^* z)(1 - x_i z^*) < (z - x_i)(z^* - x_i^*)$$

$$\Leftrightarrow \quad 1 - x_i^* z - x_i z^* + x_i x_i^* z z^* < z z^* - x_i^* z - x_i z^* + x_i x_i^*$$

$$\Leftrightarrow \quad 1 - x_i x_i^* - z z^* + x_i x_i^* z z^* < 0$$

$$\Leftrightarrow \quad (1 - |x_i|^2)(1 - |z|^2) < 0 \quad \Leftrightarrow \quad |z| > 1 \quad \text{since each } |x_i| < 1$$

x must equal 1 for H to equal +1 or -1

(b) When different people say the same phoneme, the feature vectors should have similar values.

Different phonemes from the same or different speakers should give dissimilar values.

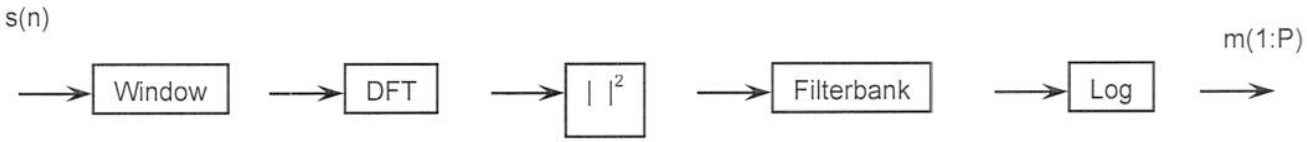
For different examples of the same phoneme, the features should be independent and uncorrelated: this allows us to multiply their probabilities.

For different examples of the same phoneme, each feature should preferably follow a probability distribution that is well described as a sum of gaussians.

The features should not be affected by the amplitude of the speech signal otherwise recognition performance would vary with your distance from the microphone.

Mel-cepstrum coefficients are substantially independent (because of the DCT) and are perceptually significant.

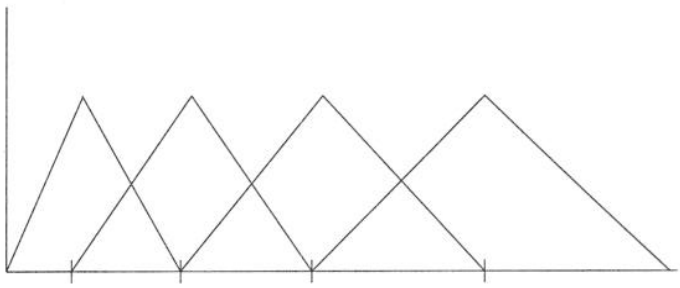
Mel-cepstrum coefficients:



For the Mel filterbank to be uniformly spaced:
min freq = 0 Hz = 0 mel.
max freq = 4000 Hz = 2146 mel

For 4 uniformly spaced filters, centres as follows:

Mel	268	804	1340	1876
Hz	188	729	1598	2999



5(a) The non-linear quantiser has quantisation levels that are optimal for a fixed-variance gaussian pdf. The quantisation levels are closer together at low signal levels: this reduces the mean square quantisation error since these levels occur with higher probability.

The adaptive scale factor normalises the input to the quantiser so that its variance remains approximately constant at the value required by the quantiser design. The scale factor is reduced each time the quantiser output gives zero and increased each time the quantiser output gives a large value. The scale factor thus converges to a value proportional to the rms value of *e(n)* and the scaled quantisation error will be proportional to the variance of *e(n)*.

The predictor uses previously transmitted information to predict the value of *s(n)*. If the prediction is a good one, the signal *e(n)* will have a smaller variance than *s(n)* and the quantisation error will be correspondingly reduced. The ratio of the variances is the prediction gain.

(b) Values of *w(n)* in the range ±½*k* will be quantised to 0 while those outside this range will be quantised to ±*k*. Therefore we must split the integration up. By symmetry, we need only consider +ve values of *w* and multiply by 2 for the –ve values.

$$\mathrm{E}\left[e^2(n)\right] = 2\left(\int_0^{½k} w^2 p(w)dw + \int_{½k}^{\infty}(w-k)^2 p(w)dw\right)$$

$$= \int_0^{½k} w^2 e^{-w}dw + \int_{½k}^{\infty}(w^2 - 2wk + k^2)e^{-w}dw$$

$$= \int_0^{\infty} w^2 e^{-w}dw - 2k\int_{½k}^{\infty} we^{-w}dw + k^2\int_{½k}^{\infty} e^{-w}dw$$

$$= 2 - 2k(1+½k)e^{-½k} + k^2 e^{-½k} = 2 - 2ke^{-½k}$$

The minimum is obtained by setting the derivative to zero:

$$-2e^{-½k} + ke^{-½k} = 0 \quad \Rightarrow \quad k = 2$$

(c) The probability that w(n) is quantised to ±k is

$$2\int_{½k}^{\infty} p(w)dw = \int_{½k}^{\infty} e^{-w}dw = e^{-½k}$$

Hence the expected value of ln(*k*(*n*+1)) – ln(*k*(*n*)) is

$$be^{-½k} - a(1-e^{-½k})$$

For *k* = 2, this is zero if $\frac{b}{a} = \frac{1-e^{-1}}{e^{-1}} = e - 1 = 1.718$

6(a) $$P(t,s) = d_s(\mathbf{x}_t)\sum_{k=1}^{S} a_{ks} P(t-1,k)$$

Every alignment going through state s at time t must go though some state, say k, at time t-1. Thus the total probability of all alignments going through state s at time t can be obtained by adding up P(t-1,k) for all k and multiplying by the probability of a transition from state k to state s. We must then multiply by $d_s(x_t)$ to include the output probability at time t.

We initialise this recursion by setting $P(1,1)=d_1(x_1)$

$$Q(t,s) = \sum_{k=1}^{S} a_{sk} d_k(\mathbf{x}_{t+1}) Q(t+1,k)$$

This is the same argument as above but working in reverse time. We need to initialise $Q(T,S)=a_{S=}$ but this does not affect the optimal alignment.

(b) We need to calculate P(3,2) ×Q(3,2)

$$P(1,1) = 0.5$$
$$P(2,1) = 0.4\times(0.9\times0.5) = 0.18$$
$$P(2,2) = 0.5\times(0.1\times0.5) = 0.025$$
$$P(3,2) = 0.8\times(0.1\times0.18 + 0.5\times0.025) = 0.0244$$

$Q(6,4)$=1 (or 0.5 if we take into account $a_{4=}$)
$Q(5,4)=0.5\times0.8\times1=0.4$
$Q(5,3)=0.8\times0.8\times1=0.64$
$Q(4,4)=0.5\times0.5\times0.4 = 0.1$
$Q(4,3)=0.2\times0.2\times0.64+0.8\times0.5\times0.4 = 0.1856$
$Q(4,2)=0.5\times0.2\times0.64 = 0.064$
$Q(3,2)=0.5\times0.6\times0.064+0.5\times0.8\times0.1856 = 0.09344$ (or 0.04672)

Hence $P(3,2)\times Q(3,2) = 0.0022799$ (or 0.00113997)