EE4-14 SOLUTIONS

1. a) i) For a discrete-time signal $x(n)$, write down the formula for its complex cepstrum $\hat{x}(n)$. [2]

ii) Discuss the main consequences of performing signal processing in the complex cepstral domain and explain why this is potentially attractive in speech signal processing. [3]

Solution

$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left\{X(e^{j\omega})\right\} e^{j\omega n} d\omega.$

Because of the intrinsic use of the log function in the definition of the cepstrum, $z-$transforms that are multiplied become additive in the cepstral domain. This offers the potential to separate, for example, the voice excitation and the vocal tract transfer function. Other good descriptions based around homomorphic signal processing will obtain full credit.

b) Consider a system represented by the rational z-transform

$$X(z) = \frac{A\prod_{k=1}^{M_a}\left(1-a_k z^{-1}\right)\prod_{k=1}^{M_b}\left(1-b_k^{-1} z^{-1}\right)}{\prod_{k=1}^{N}\left(1-c_k z^{-1}\right)}$$

in which $|a_k|, |b_k|, |c_k| < 1$.

i) Identify the terms of this expression corresponding to

- zeros inside the unit circle in z,
- zeros outside the unit circle in z,
- poles inside the unit circle in z. [3]

Solution

The zeros inside the unit circle are associated with the terms $\prod_{k=1}^{M_a}\left(1-a_k z^{-1}\right)$. The zeros outside the unit circle are associated with the terms $\prod_{k=1}^{M_b}\left(1-b_k^{-1} z^{-1}\right)$. The poles are associated with the terms $\prod_{k=1}^{N}\left(1-c_k z^{-1}\right)$.

ii) Hence factorize $X(z)$ into 3 factors:

- a delay factor involving $z^{-\lambda}$,
- a maximum phase factor $X_{\max}(z)$
- a minimum phase factor $X_{\min}(z)$

and write expressions for λ, $X_{\max}(z)$ and $X_{\min}(z)$. [3]

Solution

$X(z) = X_{\min}(z).z^{-\lambda}.X_{\max}(z)$ for which

$$\lambda = M_b,$$

$$X_{\min}(z) = \frac{A\prod_{k=1}^{M_a}(1-a_k z^{-1})}{\prod_{k=1}^{N}(1-c_k z^{-1})},$$

$$X_{\max}(z) = \prod_{k=1}^{M_b}(-b_k^{-1})\prod_{k=1}^{M_b}(1-b_k z).$$

iii) Given that $\hat{X}(z)$ is defined as the complex logarithm of $X(z)$, and adopting a suitable definition for the complex logarithm, write an expression for $\hat{X}(z)$ and hence write expressions for $\hat{x}(n)$ for the cases

- $n = 0$
- $n < 0$
- $n > 0$.

You may use the identity $\log(1-\alpha) = -\sum_{n=1}^{\infty}\frac{\alpha^n}{n}$ for $|\alpha| < 1$.

[6]

Solution

$$\hat{X}(z) = \log|A| + \sum_{k=1}^{M_b}\log|b_k^{-1}| + \log\left(z^{-M_b}\right)$$
$$+\sum_{k=1}^{M_a}\log\left(1-a_k z^{-1}\right) + \sum_{k=1}^{M_b}\log(1-b_k z) - \sum_{k=1}^{N}\log\left(1-c_k z^{-1}\right).$$

$$\hat{x}(n) = \begin{cases} \log|A| + \sum_{k=1}^{M_b}\log|b_k^{-1}| & n = 0 \\ \sum_{k=1}^{M_b}\frac{b_k^{-n}}{n} & n < 0 \\ \sum_{k=1}^{N}\frac{c_k^n}{n} - \sum_{k=1}^{M_a}\frac{a_k^n}{n} & n > 0. \end{cases}$$

iv) Find the complex cepstrum $\hat{x}(n)$ of

$$x(n) = a^n u(n)$$

in which $u(n)$ is the unit step function and $|a| < 1$. [3]

Solution:

Given $x(n)$ we can find the z-transform

$$X(z) = \sum_{n=0}^{\infty} a^n z^{-1} = \frac{1}{1 - az^{-1}}, \quad |z| > |a|.$$

The log of the z-transform can then be written

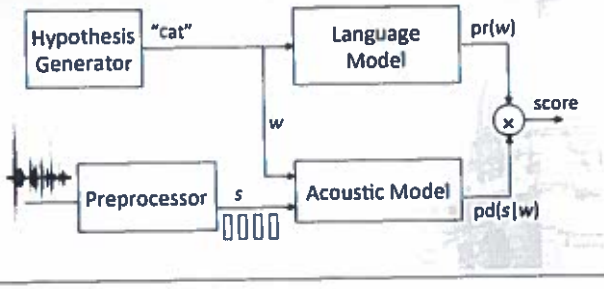
$$\begin{aligned} \hat{X}(z) &= \log(X(z)) \\ &= -\log\left(1 - az^{-1}\right) = \sum_{n=1}^{\infty} \left(\frac{a^n}{n}\right) z^{-n}. \end{aligned}$$

Then by comparing terms in z^{-n} we see that

$$\hat{x}(n) = \frac{a^n}{n} u(n-1).$$

2. a) i) Draw a block diagram that illustrates in overview the main components of an automatic speech recognition system. [2]

Solution:



ii) Explain what is meant by unigram, bigram and trigram language models. For a recognizer with a vocabulary of 30,000 words, state the number of probabilities that must be estimated for unigram, bigram and trigram language models. [3]

Solution:
Language models capture statistical information on the likely sequences of words such that for a word sequence w containing words w_i

$$pr(w) = \prod_i pr(w_i|w_{i-N+1}, \ldots w_{i-1})$$

where N defines the depth of the language model.
For a unigram model $N = 1$ so that $pr(w_i)$ is independent of any previous words. For a bigram model $pr(w) = pr(w_i|w_{i-1})$.
For a trigram model $pr(w) = pr(w|w_{i-2}, w_{i-1})$.
For a 30,000 word vocabulary, unigram: 3×10^4, bigram: 9×10^8, trigram: 27×10^{12}

iii) Consider a sequence of words denoted w and a speech signal denoted s. The conditional probability of the word sequence given the speech signal is $pr(w|s)$. Express this probability in terms of a language model and an acoustic model of speech production by employing Bayes' theorem. [3]

Solution:
The language model gives the prior probability of a word $pr(w)$. The probability that a word sequence w would generate the speech signal s is given by the acoustic model. Bayes' theoreom then gives $pr(w|s) = \dfrac{pd(s|w)ds \times pr(w)}{pd(s)ds}$.

b) In a particular application, the speech utterances are restricted to contain only the words 'red', 'blue' and 'green'. All utterances contain at least one of these words. Table 1 gives the frequencies $N(i,j)$ of each possible pair of successive words for which word i is followed by word j in a representative sample of utterances, i.e. $N(i,j)$ is the number of times word j follows word i. 'Start' and 'end' represent the start and end of an utterance.

The unigram probability for word j is denoted $p_u(j)$.

Bigram probabilities $p_b(i,j)$ for all relevant i,j are given by

$$p_b(i,j) = \begin{cases} \dfrac{N(i,j) - d(i)}{N(i)} & \text{for } N(i,j) > 2 \\ b(i)p_u(j) & \text{for } N(i,j) \leq 2 \end{cases}$$

where $N(i)$ is the number of times word i occurs.

The term $d(i)$ is defined as

$$d(i) = \begin{cases} 0 & \text{if } N(i,j) > 2 \text{ for all valid next words } j \\ 0.5 & \text{if } N(i,j) \leq 2 \text{ for all valid next words } j \end{cases}$$

and the term $b(i)$ is chosen so that

$$\sum_j p_b(i,j) = 1.$$

		Word j			
		'red'	'blue'	'green'	'end'
Word i	'start'	10	10	30	0
	'red'	1	0	60	20
	'blue'	3	50	2	10
	'green'	67	5	10	20

Table 1 Word sequence frequencies

i) Calculate the unigram probabilities for each of the words 'red', 'blue', 'green' and 'end'. [3]

Solution

The total in each of column gives $N(j)$. This gives 81, 65, 102 and 50 for the words 'red', 'blue', 'green' and 'end' respectively. Therefore the unigram probabilites are given by dividing by the total number of words = 298, to be 0.272, 0.218, 0.342, 0.168 for the words 'red', 'blue', 'green' and 'end' respectively.

ii) Find the values $d(i)$ for all words i. [2]

Solution

Word	$d(i)$
start	0
red	0.5
blue	0.5
green	0

iii) Hence find the bigram probabilities $p_b(i,j)$. [5]

Solution

Begin by calculating the $b(i)$ term for the infrequent words red and blue

$$b(red) = \frac{1+0+0.5+0.5}{81} \times \frac{1}{0.272+0.218}$$
$$= 0.0504$$
$$b(blue) = \frac{0.5+0.5+2+0.5}{65} \times \frac{1}{0.342}$$
$$= 0.157.$$

This then gives the bigram probabilities as shown in the following table.

		Word j						
		'red'	'blue'	'green'	'end'	N	d	b
Word i	'start'	0.2	0.2	0.6	0	50	0	
	'red'	0.014	0.011	0.735	0.241	81	0.5	0.050
	'blue'	0.038	0.762	0.054	0.146	65	0.5	0.157
	'green'	0.657	0.049	0.098	0.196	102	0	

iv) Briefly explain why the term $d(i)$ is needed for bigram probability calculation and summarize its effect on the bigram probabilities.

[2]

Solution

The need for discounting arises because bigram probabilities cannot be estimated accurately for sequences that occur only very rarely or not at all. For these cases, the probabilities are therefore assumed to be proportional to the corresponding unigram probabilities. These rare bigrams are likely to be under-represented in the training data. We therefore apply the discounting factor $d(i)$. The overall effect of $d(i)$ is to reduce the probability of common bigrams and increase the propability of rare bigrams.

3.

a) Speaker recognition aims to determine the most likely identity of an unknown talker from an example of their speech by comparing *features* of the unknown talker's speech to *features* previously extracted from the speech of a set of talkers with known identity. The identity of the unknown talker is determined as the closest matching known talker.

Describe the processing steps employed to extract appropriate *features* from speech. Discuss the desired properties of such *features*. Give examples of appropriate choices for such *features* for the task of speaker recognition. Include any relevant diagrams to clarify your description.

[6]

Solution

The processing steps should be clearly set out and supported by relevant diagrams.

1. Frame-based processing to segment the speech signal into (overlapping) frames. The use of a window function should be discussed.

2. In each frame, a set of features is detemined from the speech signal in that frame after windowing.

3. The properties of the features should be discriminative between speakers and substantially independent of the phonetic content of the speech.

4. Any examples of features which are related to the talker are acceptable, including features of the pitch, the voice excitation and the vocal tract parameters. Features describing the vocal tract transfer function are specifically not relevant.

b) At a particular time, the *features* from part (a) are stored in a column vector $\mathbf{z} = (z_1, z_2, \ldots, z_N)^T$ where the elements of $\mathbf{z}$ are assumed independent identically distributed Gaussian random variables having zero mean and unit variance with a probability density function

$$p(z_i) = (2\pi)^{-1/2} \exp\left(-\frac{z_i^2}{2}\right)$$

and ^{T} indicates matrix transpose.

i) Derive an expression for the probability density function of the vector $\mathbf{z}$ and show that the expected value $E\left(\mathbf{z}\mathbf{z}^T\right)$ is equal to the identity matrix. [4]

Solution:
Since the components of $\mathbf{z}$ are independent, their joint p.d.f. is the product of the individual density functions:

$$\begin{aligned} p(\mathbf{z}) &= \prod_{i=1}^{N} (2\pi)^{-1/2} \exp\left(-\frac{z_i^2}{2}\right) \\ &= (2\pi)^{-N/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{N} z_i^2\right) \\ &= (2\pi)^{-N/2} \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right). \end{aligned}$$

Element (i,j) of $E\left(\mathbf{z}\mathbf{z}^T\right)$ is written $E(z_i z_j)$. If $i = j$ then $E(z_i z_j) = 1$ since z_i has unit variance. If $i \neq j$, since z_i and z_j are independent, $E(z_i z_j) = E(z_i)E(z_j) = 0$ since the z_i have zero mean. Hence $E(\mathbf{z}\mathbf{z}^T)$ is the identity matrix.

ii) Consider a non-singular $N \times N$ matrix $\mathbf{A}$. If $\mathbf{x} = \mathbf{A}\mathbf{z}$, obtain an expression for the covariance matrix $\mathbf{C} = E\left(\mathbf{x}\mathbf{x}^T\right)$. [4]

Solution:

$$\mathbf{C} = E\left(\mathbf{x}\mathbf{x}^T\right) = E\left(\mathbf{A}\mathbf{z}\mathbf{z}^T\mathbf{A}^T\right) = \mathbf{A}E\left(\mathbf{z}\mathbf{z}^T\right)\mathbf{A}^T = \mathbf{A}\mathbf{A}^T.$$

iii) Derive an expression in terms of $\mathbf{C}$ for the natural log of the probability density function of the vector $\mathbf{x}$, where $\mathbf{x}$ is defined in part (ii) above. [6]

Solution:
Exploiting the formula given, the p.d.f. of $\mathbf{x}$ is:

$$\begin{aligned} &= (2\pi)^{-N/2} |\mathbf{A}|^{-1} \exp\left(-\frac{1}{2}(\mathbf{A}_1\mathbf{x})^T \mathbf{A}^{-1}\mathbf{x}\right) \\ &= (2\pi)^{-N/2} |\mathbf{A}|^{-1} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{A}^{-T}\mathbf{A}^{-1}\mathbf{x}\right) \\ &= (2\pi)^{-N/2} |\mathbf{A}|^{-1} \exp\left(-\frac{1}{2}\mathbf{x}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{x}\right) \\ &= (2\pi)^{-N/2} |\mathbf{A}|^{-1} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}\right). \end{aligned}$$

Hence the log pdf is $-\frac{1}{2}N\log(2\pi) + \log|\mathbf{C}| + \mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}$.

Note: You may assume that if the probability density function of $\mathbf{z}$ is $f(\mathbf{z})$, then the probability density function of $\mathbf{x} = \mathbf{A}\mathbf{z}$ is given by $f\left(\mathbf{A}^{-1}\mathbf{x}\right) \times |\mathbf{A}|^{-1}$.

4. a) Speech enhancement can be performed in the STFT domain using the formula

$$Z(l,k) = H(l,k)Y(l,k)$$

in which l is the time-frame index, k is the frequency index, $Y(l,k)$ is the noisy speech signal and $H(l,k) = \sqrt{1 - \frac{\hat{\phi}_v(l,k)}{|Y(l,k)|^2}}$.

i) State what $\hat{\phi}_v$ represents and explain two alternative methods for finding $\hat{\phi}_v$ in a speech enhancement system. Discuss the relative merits of each approach. [5]

Solution
The parameter $\hat{\phi}_v$ is the estimated noise variance. It can estimated as the average power in the signal during speech absence (pause) but that would require a Voice Activity Detection (VAD) which, if the decision is incorrect, leads to a poor estimate. It can also be estimated using the minimum statistics approach. This approach doesn't require a VAD but there is a tradeoff between reacting to changes in noise level and smoothing the estimate. The estimate tends to lag behind in time the true value for increasing noise. Minimum statistics-based approaches also rely on there being short pauses in any time-frequency bin of the STFT of the speech where energy decays to the noise floor.

ii) What additional parameter(s) does model-based spectral enhancement depend on and how might it/they be estimated? [2]

Solution
Additional parameters that are relevant are the speech level or *a priori* SNR. These would lead to the so-called 'decision direct approach.'

iii) The values of $Y(l,k)$ over a particular interval are given in Table 1. Assuming $\hat{\phi}_v = 0.16$ and is constant, on which time-frequency indices does the method fail and why? How can this be overcome?

[6]

	$l=1$	$l=2$
$k=1$	0.332-0.009i	-0.277+1.695i
$k=2$	0.668+0.482i	-0.226+0.580i
$k=3$	-0.103-0.342i	0.254+0.359i
$k=4$	2.069+1.284i	-2.912+0.738i
$k=5$	-0.009+1.260i	-0.583-0.120i

Table 1 STFT analysis of noisy speech $Y(l,k)$

Solution

$(l,k) = (1,1)$and $(l,k) = (1,3)$.

In these cells $|Y(l,k)| < 0.4$, so $|Y(l,k)|^2 < 0.16$ and $\dfrac{\hat{\phi}_v(l,k)}{|Y(l,k)|^2} > 1$ so that $1 - \dfrac{\hat{\phi}_v(l,k)}{|Y(l,k)|^2} < 0$.

The situations gives rise to the problem is known as over-subtraction and can be avoided by imposing a minimum value on the gain. Absolute values are given for reference as follows (not required necessarily in the answer).

	$l=1$	$l=2$
$k=1$	0.3325	1.7178
$k=2$	0.8242	0.6226
$k=3$	0.3573	0.4399
$k=4$	2.4353	3.0038
$k=5$	1.2600	0.5951

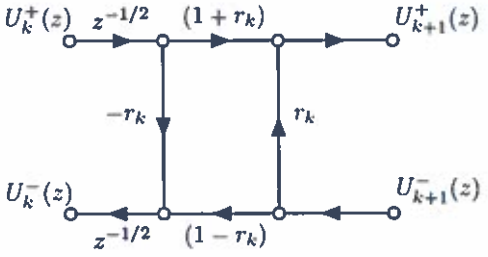


Figure 4.1

b) Consider the lossless tube model of the vocal tract for which a partial sketch diagram is shown in Fig. 4.1.

i) Draw a complete labelled diagram of a lossless tube model of order 4 and describe the model and its relationship to the human speech production system. [4]

Solution

Either a signal flow diagram or a schematic diagram are fully acceptable. The relationship between the physical acoustic tube of the vocal tract and the lossless tube model must be clearly stated. Marks will be deducted for missing or inaccurate labels. In particular the volume flow and reflection coefficients must be label throughout.

ii) For the partial model shown in Fig. 4.1, find expressions for $U_k^+(z)$ and $U_k^-(z)$ in terms of $U_{k+1}^+(z)$ and $U_{k+1}^-(z)$. [3]

Solution

From analysis of the signal flow graph, we have

$$U_{k+1}^{+}(z) = (1+r_k)z^{-1/2}U_k^{+}(z) + r_k U_{k+1}^{-}(z)$$
$$U_k^{-}(z) = -r_k z^{-1}U_k^{+}(z) + (1-r_k)z^{-1/2}U_{k+1}^{-}(z).$$

These equations can be solved as required to give

$$U_k^{+}(z) = \frac{z^{1/2}}{1+r_k}U_{k+1}^{+}(z) - \frac{r_k z^{1/2}}{1+r_k}U_{k+1}^{-}(z)$$
$$U_k^{-}(z) = \frac{-r_k z^{1/2}}{1+r_k}U_{k+1}^{+}(z) + \frac{z^{-1/2}}{1+r_k}U_{k+1}^{-}(z).$$