Solutions

1. a) A model of speech production is illustrated in Figure 1.1. Explain the meaning and give the definition of each of the signals, parameters and operations in this model. Discuss the rate at which the parameters of the model should be updated in a speech coding application. [4]

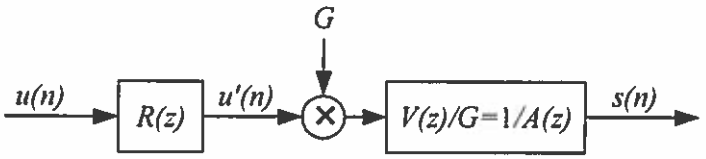


Figure 1.1 A model of speech production.

Solution:
$R(z)$ is the lip radiation model, normally represented by a first order numerical difference operator to give a first order highpass filter characteristic.
$V(z)$ is the vocal tract filter, represented by an allpole filter with order typically in the range 10-20.
$A(z)$ is the prediction filter.
G is a gain.
$u(n)$ is the glottal flow and it's first derivative is $u'(n)$.
The rate of updating is determined by the epoch over which the speech is quasi-stationary. Updating once every 20 ms is typical. (Longer periods between updates reduces data rate requirements in a speech coder but reduces quality. Shorter rates induce higher levels of estimation variance in the parameter estimates.)

b) Consider a speech signal $s(n)$ over a frame of samples $\{F\}$ and the speech covariance matrix $\mathbf{\Phi}$ with elements ϕ_{ij}. Consider also the prediction of $s(n)$ using LPC with predictor coefficients denoted a_i, for $i = 0, 1, 2, \ldots$.

i) Show how $\mathbf{\Phi}$ and the elements ϕ_{ij} are formed in terms of $s(n)$. [2]

ii) Formulate an expression for the prediction error in terms of the prediction coefficients a_i. [2]

iii) Write down an expression for the sum squared prediction error over the frame of speech samples $\{F\}$. [1]

iv) Derive an expression in terms of $s(n)$ for the predictor coefficients that minimize the sum squared prediction error and show that this expression can be written in termsa of ϕ_{ij}. [5]

v) How is $\{F\}$ chosen in the case of autocorrelation LPC? State any consequences of this choice on the computation of a_i. [2]

vi) Let the normalized power spectrum of the prediction error signal

$e(n)$ be defined

$$P_E(e^{j\omega}) = \frac{|E(e^{j\omega})|^2}{Q_E}$$
$$Q_E = \frac{1}{2\pi}\int_{\omega=0}^{2\pi} |E(e^{j\omega})|^2 d\omega$$

where $E(z)$ is the z-transform of $e(n)$.

Next let the spectral roughness be defined

$$R_E = \frac{1}{2\pi}\int_{\omega=0}^{2\pi} \left(P_E(e^{j\omega}) - 1 - \log\left(P_E(e^{j\omega})\right)\right) d\omega.$$

By considering $E(z) = A(z)S(z)$, show that minimizing Q_E is equivalent to minimizing the spectral roughness of the prediction error. [4]

Solution

The $(i,j)^{th}$ element of the covariance matrix is given by $\phi_{ij} = \sum_{n\in\{F\}} s(n-i)s(n-j)$ from which the prediction error is obtained as

$$e(n) = s(n) - \sum_{j=1}^{p} a_j s(n-j) = s(n) - a_1 s(n-1) - a_2 s(n-2) - \ldots - a_p s(n-p).$$

The sum squared prediction error is given by

$$Q_E = \sum_{n\in\{F\}} e^2(n).$$

The predictor coefficients that minimize the sum squared prediction error are obtain by writing

$$\frac{\partial Q_E}{\partial a_i} = \sum_{n\in\{F\}} \frac{\partial\left(e^2(n)\right)}{\partial a_i} = \sum_{n\in\{F\}} 2e(n)\frac{\partial e(n)}{\partial a_i} = -\sum_{n\in\{F\}} 2e(n)s(n-i)$$

$$\sum_{n\in\{F\}} e(n)s(n-i) = 0 \quad \text{for} \quad i = 1,\ldots,p$$

$$\Rightarrow \sum_{n\in\{F\}} \left(s(n)s(n-i) - \sum_{j=1}^{p} a_j s(n-j)s(n-i) \right) = 0 \quad \text{for} \quad i = 1,\ldots,p$$

$$\Rightarrow \sum_{j=1}^{p} a_j \sum_{n\in\{F\}} s(n-j)s(n-i) = \sum_{n\in\{F\}} s(n)s(n-i)$$

$$\Rightarrow \sum_{j=1}^{p} \phi_{ij}a_j = \phi_{i0} \quad \textit{where} \quad \phi_{ij} = \sum_{n\in\{F\}} s(n-i)s(n-j)$$

In matrix form:

$\boldsymbol{\Phi}\mathbf{a} = \mathbf{c}$ which implies that $\mathbf{a} = \boldsymbol{\Phi}^{-1}\mathbf{c}$. The existance of $\boldsymbol{\Phi}^{-1}$ is required.

For autocorrelation LPC, the frame is chosen over all non-zero values after applying a (typically) tapered window. A note of interest here is that if the window $w(m)$ is non-zero over the interval $0 \le m \le L-1$, then the order p prediction error is non-zero over an additional p samples $0 \le m \le L-1+p$.

The spectral roughness

$$R_E = \frac{1}{2\pi}\int_{\omega=0}^{2\pi}\left(P_E(e^{j\omega}) - 1 - \log\left(P_E(e^{j\omega})\right)\right)d\omega$$

can be written

$$R_E = \frac{1}{2\pi}\int_{\omega=0}^{2\pi} -\log\left(P_E(e^{j\omega})\right)d\omega$$

since the term involving $\int P_E(e^{j\omega})d\omega = 1$. Therefore

$$R_E = \log(Q_E) - \frac{1}{2\pi}\int_{\omega=0}^{2\pi}\log\left(|E(e^{j\omega})|\right)d\omega$$

since

$$P_E(e^{j\omega}) = \frac{|E\left(e^{j\omega}\right)|^2}{Q_E}.$$

Using $E(z) = A(z)S(z)$ we can now write

$$\log\left(|E(e^{j\omega})|^2\right) = \log\left(|S(e^{j\omega})|^2\right) + \log\left(|A(e^{j\omega})|^2\right).$$

Substituting in the expression for spectral roughness gives

$$\begin{aligned} R_E &= \log(Q_E) - \frac{1}{2\pi}\int_{\omega=0}^{2\pi}\log\left(|E(e^{j\omega})|^2\right)d\omega \\ &= \log(Q_E) - \frac{1}{2\pi}\int_{\omega=0}^{2\pi}\log\left(|S(e^{j\omega})|^2\right)d\omega - \frac{1}{2\pi}\int_{\omega=0}^{2\pi}\log\left(|A(e^{j\omega})|^2\right)d\omega \end{aligned}$$

Since the term involving $A(e^{j\omega}) = 0$ when all the roots of $A(z)$ lie within the unit circle, then the expression simplifies to

$$R_E = \log(Q_E) - \frac{1}{2\pi}\int_{\omega=0}^{2\pi}\log\left(|S(e^{j\omega})|^2\right)d\omega.$$

The final step of the development is to note that $S(e^{j\omega})$ is independent of $A(z)$, so that choosing $A(z)$ to minimize Q_E, as is the case in the LPC analysis above, will also minimize R_E.

2. a) Consider a signal being quantized using a quantization process employing a set of quantization bins. Each bin covers an amplitude range w and any particular amplitude is contained in exactly one bin. Now consider an input signal such that the distribution of signal amplitude in any bin is uniform over the amplitude range of the bin.

i) Briefly explain what is meant by 'one least significant bit' in this context. [1]

ii) For a quantization bin spanning the range $-w/2$ to $+w/2$, within which all input values are quantized to zero, derive an expression for the RMS quantization error in this bin in terms of the bin width. [3]

Solution:
In this context, one LSB can be used to refer to the amplitude difference between one quantization bin and the next. This interpretation of the context is essential to obtain the mark.
The mean square error is given by

$$\int_{-w/2}^{+w/2} x^2 \frac{dx}{w} = \left[\frac{x^3}{3w}\right]_{-w/2}^{+w/2} = \frac{w^2}{12}$$

so that the RMS error is $= 0.289w$

b) i) Consider a speech signal $s(n)$ at time index n with probability density function $p(s)$. Further consider a nonuniform quantizer such that the input signal amplitudes in the range $[a_{i-1}, a_i]$ are quantized to output amplitude values s_i. Find an appropriate expression for the quantized amplitudes s_i in terms of a_i, a_{i-1}, and $p(s)$ such that the quantization error is minimum in the mean square. [4]

ii) Give an example of a probability density function $p(s)$ for which nonuniform quantization would be preferable compared to uniform quantization and explain your reasoning. [2]

Solution

Quantization error $q(s) = s_i - s$.

The mean square quantization error is given by $E = \int_{-\infty}^{\infty} p(s)q^2(s)ds = \sum_{i=1}^{N} \int_{a_{i-1}}^{a_i} p(s)(s_i - s)^2 ds$.

By differentiation w.r.t. the s_i we obtain $\frac{\partial E}{\partial s_i} = \int_{a_{i-1}}^{a_i} -2p(s)(s - s_i)ds = 2s_i \int_{a_{i-1}}^{a_i} p(s)ds - 2\int_{a_{i-1}}^{a_i} sp(s)ds$.

The minimum error is found by equating the above expression to zero to give

$$s_i = \frac{\int_{a_{i-1}}^{a_i} sp(s)ds}{\int_{a_{i-1}}^{a_i} p(s)ds}.$$

In an non-uniform quantization, bin centres are typical chosen to match the (non-uniform) pdf of the input signal. Speech has non-uniform pdf (super-Gaussian) and therefore lower quantization error can be obtained using non-uniform quantization. Examples of pdfs sometimes chosen to represent speech include the Laplacian distribution.

iii) A speech signal $s(n)$ is represented using PCM with a precision of 16 bits per sample. It is now intended to encode $s(n)$ using μ-law encoding in which **s**, **e** and **m** denote the sign, exponent and mantissa bits respectively, and the quantization scheme has bin centres at

$$\pm\{(m + 16.5)2^e - 16.5\}.$$

For some particular $n = n_1$, it is found that $s(n_1) = -1793$. Determine the bit values used to μ-law encode $s(n_1)$ and state the amplitude of the error introduced by μ-law coding of this sample.

[5]

Solution

The exponent necessary to attain this value is 6. For $e = 6$, the range of bin centre values is between -1039.5 and -1999.5. The nearest bin centre occurs with a mantissa of 12 to give the μ-law quantized value of -1807.5. The bit pattern is therefore

$$\{seeemmmm\} = \{11101100\}.$$

Hence the encoding error is 14.5.

c) The quantizer labelled Q in Figure 2.1 is a uniform 5-level quantizer with outputs such that

$$w(n) \in \{-2, -1, 0, 1, 2\}.$$

The block labelled 'Update k' modifies the input value of k at time index

n according to the following rule:

$$k(n+1) = \begin{cases} 3k(n)) & \text{when} \quad w(n) = \pm 2 \\ 1.1k(n) & \text{when} \quad w(n) = \pm 1 \\ 0.9k(n) & \text{when} \quad w(n) = 0. \end{cases}$$

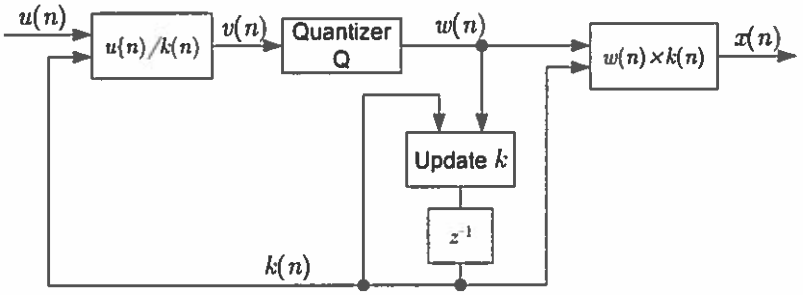


Figure 2.1 Quantizer

An input signal $u(n) = \{1,2,2,8,10,10\}$ is applied. Determine the resulting values of $k(n)$ and the quantization error $(u(n) - x(n))$ for $n = 0,1,\ldots,5$, given that k is initialized to 1. [5]

Solution

n	0	1	2	3	4	5
$u(n)$	1	2	2	8	10	10
k	1.00	1.10	3.30	3.63	10.89	11.98
$v = u/k$	1.00	1.82	0.61	2.20	0.92	084
w	1	2	1	2	1	1
$x = w * k$	1.00	2.20	3.30	7.26	10.89	11.98
new k	1.10	3.30	3.63	10.89	11.98	13.18
$q_e(n)$	0.00	-0.20	-1.30	0.74	-0.89	-1.98

3. a) Draw and label a general block diagram of a single channel noise reduction system for speech enhancement. [5]

Solution:
The block diagram is of the form shown in Fig. 3.1 and must be fully labelled to receive full marks.

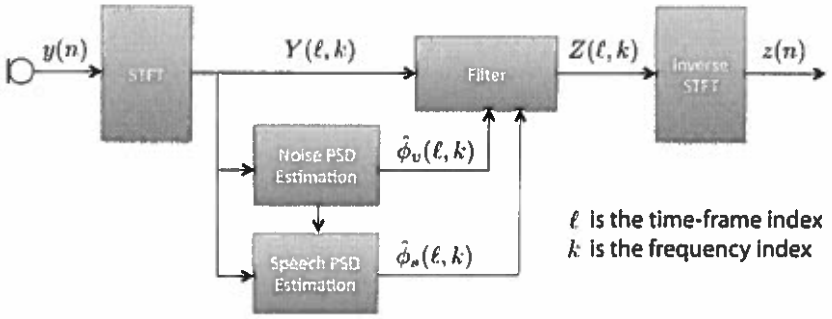


Figure 3.1 General Speech Enhancement System

b) Consider a specific case of a frequency domain-based single channel noise reduction scheme employing power spectrum subtraction. The noisy signal in time frame l and at frequency bin k is denoted $Z(l,k)$.

i) Describe the operation of this noise reduction scheme and explain what information is needed by the scheme, in addition to the noisy signal. [2]

ii) Show that the noise reduction scheme can be viewed as a filtering operation such that the output, $Z_o(l,k)$, after processing by the scheme can be written

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

and then determine the expression for $H(l,k)$ in this case. [5]

Solution
The description follows directly from the general description but employs power spectral estimation and noise power spectral variance estimation. Additional credit will be given for depth of understanding shown, particularly concerning methods for estimation of the noise variance, $\hat{\phi}_v(l,k)$.
Power spectrum subtraction is written

$$A^2_{Zo}(l,k) = A^2_Z(l,k) - \hat{\phi}_v(l,k).$$

This can be formulated as a filtering operation

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

when

$$H(l,k) = \left(1 - \frac{\hat{\phi}_v}{|Z^2(l,k)|}\right)^{1/2}.$$

c) Now consider a system with two microphones, separated physically by a small distance. Each of the two microphones receives the same speech signal $s(n)$ corrupted by independent additive noise sources $e_1(n)$ and $e_2(n)$ respectively. The signals at the two microphones $x_1(n)$ and $x_2(n)$ are given by

$$x_1(n) = s(n) + e_1(n)$$
$$x_2(n) = s(n) + e_2(n).$$

With the expectation operator denoted $E[\cdot]$, the noise sources have the following properties in terms of their means and variances:

$$E[e_1(n)] = E[e_2(n)] = 0$$
$$E\left[e_1^2(n)\right] = \sigma_1^2$$
$$E\left[e_2^2(n)\right] = \sigma_2^2.$$

Now consider a weighted sum of the microphone signals

$$z(n) = ax_1(n) + (1-a)x_2(n)$$

where a is a scalar constant.

i) If $z(n)$ is written as

$$z(n) = s(n) + e_3(n),$$

find $e_3(n)$ in terms of $e_1(n)$ and $e_2(n)$. [2]

ii) Next, find the optimal value of a that minimizes the variance of $e_3(n)$. [4]

iii) Finally, determine the minimum variance of $e_3(n)$ in terms of σ_1^2 and σ_2^2. [2]

Solution

The weighted sum of the noise sources is simply:

$$e_3(n) = ae_1(n) + (1-a)e_2(n).$$

Since $E[e_3(n)] = 0$, the variance of the combined noise is found as

$$\begin{aligned}\sigma_3^2 &= E\left[e_3^2(n)\right] \\ &= E\left[(ae_1(n) + (1-a)e_2(n))^2\right] \\ &= a^2 E\left[e_1^2(n)\right] + (1-a)^2 E\left[e_2^2(n)\right] \\ &= a^2\sigma_1^2 + (1-a)^2\sigma_2^2.\end{aligned}$$

The min is obtained for optimal $a = a_o$ found by differentiation w.r.t. a and setting to zero, giving

$$\begin{aligned}\frac{\partial\sigma_3^2}{\partial a} = 2a\sigma_1^2 - 2(1-a)\sigma_2^2 &= 0 \\ 2a_o\sigma_1^2 + 2a_o\sigma_2^2 &= 2\sigma_2^2 \\ a_o &= \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.\end{aligned}$$

The value of the resulting min variance of $e_3(n)$ can then be written

$$\begin{aligned}E\left[e_3^2(n)\right] &= \left(\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^2\sigma_1^2 + \left(\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2\sigma_2^2 \\ &= \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}.\end{aligned}$$

4. Consider a speech signal that has been segmented into time frames. A feature vector $\mathbf{x}_t$ is computed from each frame $t = 1,2,\ldots,T$ of the speech signal, where T is the number of frames. The set of feature vectors representing the speech signal is then denoted $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_T\}$. Next consider a speech recognition system based on a hidden Markov model containing S states $\{s_1,s_2,\ldots,s_S\}$.

a) Explain with the use of any appropriate diagrams how the hidden Markov model can be used to recognize speech. Include a clear list and explanation of the parameters involved. Also include an explanation of an *alignment* in this context. Further include the definitions of the output probability density and transition probability in the hidden Markov model. [6]

Solution:
The explanation is bookwork. Credit will be given for demonstration of understanding as well as presentation of factual information and diagrams.

A hidden Markov model for a word must specify the following parameters for state s: The mean and variance for each of the F elements of the parameter vector: μ_s and σ_s^2. These allow us to calculate $d_s(\mathbf{x})$: the output probability density of input frame $\mathbf{x}$ in state s. The transition probabilities a_{sj} to every possible successor state is often zero for all j except $j = s$ and $j = s+1$. It is then called a left-to-right, no skips model. For a hidden Markov model with S states we therefore have around $(2F+1)S$ parameters. A typical word might have $S = 15$ and $F = 39$ giving 1200 parameters in all. The alignment refers to estimation of the (hidden) alignment of frames of data into the various states of the HMM. In speech recognition this is equivalent to segmentation of the audio into the units of the model, such as uni-, bi-, or triphone models, for example.

b) Let $P(t,s)$ denote the total probability density of all the possible alignments of $\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_t\}$ given that $\mathbf{x}_1$ is aligned to state s_1 and $\mathbf{x}_t$ is aligned to state s. Then let $Q(t,s)$ denote the total probability density of all the possible alignments of $\{\mathbf{x}_{t+1},\mathbf{x}_{t+2},\ldots,\mathbf{x}_T\}$ given that $\mathbf{x}_t$ is aligned to state s and $\mathbf{x}_T$ is aligned to state S.

Now show how $P(t,s)$ and $Q(t,s)$ could be computed recursively and explain your reasoning. Clarify your approach by expressing $P(t,s)$ in terms of $P(t-1,k)$ for $k = 1,2,\ldots,S$, and also by expressing $Q(t,s)$ in terms of $Q(t+1,k)$ for $k = 1,2,\ldots,S$. Also state the initial conditions for P and Q for recursive computation.

[6]

Solution

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 $d_s(\mathbf{x}_t)$ to include the output probability at time t. There is a similar argument for Q but in reverse time.

The recursive forms are given by

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

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

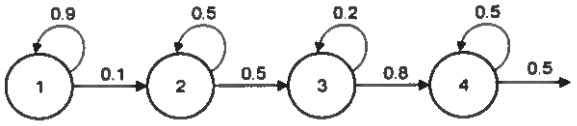
The required initializations are

$$P(1,1) = d_1(\mathbf{x}_1)$$
$$Q(T,S) = a_S.$$

c) A hidden Markov model with 4 states is going to be used for a simplified speech recognition task.

i) Sketch the state diagram of a hidden Markov model with 4 states. Label the diagram and the state transitions such that the model is constrained to be a *'left-to-right, no skips'* model. [2]

Solution



ii) The state transition probabilities are

$$a_{12} = 0.1,\ a_{23} = 0.5,\ a_{34} = 0.8$$

and the exit probability is $a_{4=} = 0.5$.

Label these transition probabilities on the state diagram.

The output probability densities for each of 6 observed feature vectors are shown in Table 1. Determine the total probability of all alignments of the observation with the model for which frame 3 is in state 2 given that frame $\mathbf{x}_1$ is in state s_1 and frame $\mathbf{x}_6$ is in state s_4. Show your answer to 6 decimal places.

[6]

	x_1	x_2	x_3	x_4	x_5	x_6
s_1	0.5	0.3	0.5	0.2	0.1	0.1
s_2	0.4	0.5	0.8	0.6	0.3	0.8
s_3	0.2	0.8	0.2	0.8	0.2	0.2
s_4	0.5	0.4	0.5	0.2	0.5	0.8

Table 1 Output probability densities.

Solution
We need to calculate P(3,2) ×Q(3,2)
P(1,1) = 0.5
P(2,1) = 0.3×(0.9×0.5) = 0.135
P(2,2) = 0.5×(0.1×0.5) = 0.025
P(3,2) = 0.8×(0.1×0.135 + 0.5×0.025) = 0.0208
Q(6,4)=0.5
Q(5,4)=0.5×0.8×0.5=0.2
Q(5,3)=0.8×0.8×0.5=0.32
Q(4,4)=0.5×0.5×0.2 = 0.05
Q(4,3)=0.2×0.2×0.32+0.8×0.5×0.2 = 0.0928
Q(4,2)=0.5×0.2×0.32 = 0.032
Q(3,2)=0.5×0.6×0.032+0.5×0.8×0.0928 = 0.0467
Hence P(3,2) ×Q(3,2) = 0.000971