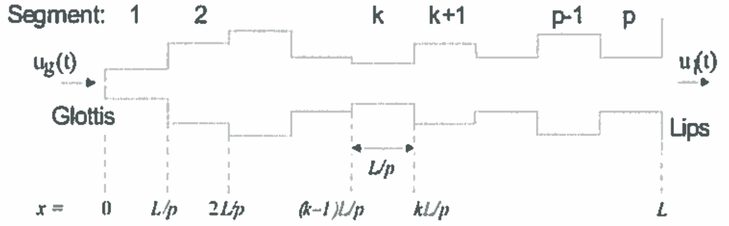
Solutions

SPEECH PROCESSING

1. Consider a p^{th} order lossless tube model of the vocal tract in the human speech production system.

a) Draw a fully labelled sketch of this model and briefly explain its key characteristics. [4]

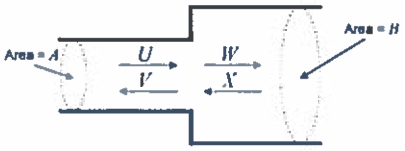
Solution:



The lossless tube model represents the vocal tract using a concatenation of lossless tube sections each of constant cross-sectional area. The length of each section corresponds to the distance travelled by sound in one half sampling period. The number of sections corresponds to the order of the model. Within each section, we consider a forward wave representing the flow of air from left to right as well as a reverse wave representing the flow of air from right to left. The flows are modelled in terms of the volume velocity. At each section junction, reflections occur according to the nature of the change of cross-sectional area.

b) Consider the junction between two sections of the lossless tube model for which the cross-sectional area either side of the junction is different. Derive expressions relating the forward and reverse acoustic waves in the two tube sections. Write your expressions also in matrix form. [5]

Solution: The interface between two sections of the lossless tube model can be illustrated as shown in the figure.



Let the forward waves be U and W and let the reverse waves be V and X. Then from flow continuity we have

$$(U-V)=(W-X)$$

and from pressure continuity we have

$$\frac{\rho c}{A}(U+V)=\frac{\rho c}{B}(W+X).$$

Combining these in matrix form leads to

$$\begin{pmatrix} 1 & -1 \\ B & B \end{pmatrix}\begin{pmatrix} U \\ V \end{pmatrix}=\begin{pmatrix} 1 & -1 \\ A & A \end{pmatrix}\begin{pmatrix} W \\ X \end{pmatrix}$$

c) State the definition of the reflection coefficients in terms of the cross-sectional area of the tubes in this model. State an appropriate value for the reflection coefficient at the lips. [3]

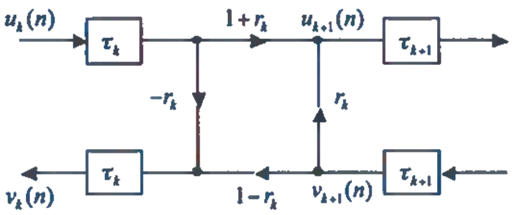
Solution: Denote the cross-sectional areas either side of the junction as A and B. The reflection coefficients are defined as

$$r=\frac{B-A}{B+A}$$

Reflection coefficient at the lips: r_p is from the last tube section to the air of effectively infinite cross-sectional area, meaning that the reflection coefficient approaches $r_p=+1$.

d) Sketch a signal flow graph for a complete lossless tube model employing 2 tube sections. The signal flow graph should contain delay elements, multipliers and addition nodes. [4]

Solution: The signal flow graph for a single interface is shown in the figure. For a model comprising two sections, two of these elements will be connected in cascade.



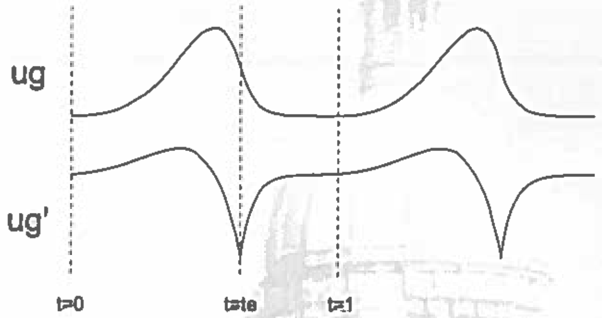
e) Now consider the glottal volume velocity as a function of time t, denoted $u_g(t)$. Draw an illustrative sketch of $u_g(t)$ over a representative time duration for a vowel.

The definition of the LF Model includes

$$u_g'(t) = \begin{cases} e^{at}\sin(bt) & 0 \le t < t_e \\ c + de^{-ft} & t_e \le t < 1. \end{cases}$$

Add to your above sketch of $u_g(t)$ a time-aligned sketch of the corresponding function $u_g'(t)$ and label $t = 0,\ t_e,\ 1$. [4]

Solution: The sketch must show at least 2 cycles and indicate that $u_g(0) = u_g(1) = 0$ and $u_g(t),\ u_g'(t)$ continuous at $t = t_e$.



2. In a speech recognition task, a particular hidden Markov model (HMM) with S states includes the transition probability from state i to state j which are denoted a_{ij}. Consider features computed from T frames representative of the speech, $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_T$ which are compared with the HMM. The output probability density is denoted by $d_i(\mathbf{x}_t)$ for frame t in state i. It can be assumed that frame 1 is in state 1.

a) 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, given that frame t of the speech signal is in state s.

[3]

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

b) Let $B(t,s)$ be defined as the highest probability density that the model generates frames $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_t$. 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]

Solution: The best path for $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_t$ with frame t in state s must have frame $t-1$ in one of the earlier states, state i. The partial path $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_{t-1}$ must also be optimum, so that
$B(t,s) = B(t-1,i) \times a_{is} \times d_s(\mathbf{x}_t)$
Now $B(t,s)$ represents the probability density of the best path so that
$B(t,s) = \max_{i=1,2,\ldots S} B(t-1,i) \times a_{is} \times d_s(\mathbf{x}_t)$.
Finally, given the assumption that frame 1 is in state 1, $B(1,s) = d_1(\mathbf{x}_1)$ if $s = 1$ and 0 otherwise.

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]

Solution: The maximization process in the above expression requires a very large amount 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, prune all those values less than $k \times \max_s B(t,s)$. 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) 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 2.1 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 $k = 0.15$. 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]

Solution:

$B(1,1) = 0.5$

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

$B(2,2) = 0.5 \times 0.8 \times 0.7 = \mathbf{0.24}$

$B(3,2) = 0.28 \times 0.3 \times 0.3 = \mathbf{0.0216}$

$B(3,3) = 0.28 \times 0.7 \times 0.1 = 0.0168$

$B(4,2) = 0.0252 \times 0.3 \times 0.1 = 0.000648$

$B(4,3) = \max(0.0216 \times 0.6,\ 0.0168 \times 0.5) \times 0.6 = \mathbf{0.009072}$

$B(4,4) = 0.0168 \times 0.5 \times 0.1 = 0.00084$

$bw = 0.15 \times 0.009072 = 0.0013608$ hence prune B(4,4) and B(4.2)

$B(5,3) = 0.009072.5 \times 0.6 = \mathbf{0.002722}$

$B(6,4) = 0.002722 \times 0.5 \times 0.4 = \mathbf{0.00054432}$

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

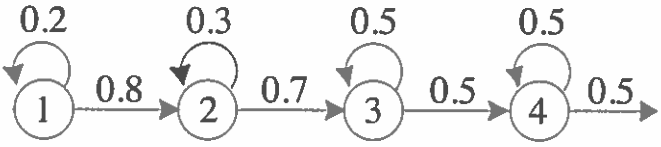


Figure 2.1

	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	0.6	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

3. a) Consider a linear predictor with order p and prediction coefficients a_k. Consider this linear predictor being applied to a segment containing N samples of a speech signal $s(n)$, indexed beginning with $n = 0$. Let the prediction error be denoted E.

i) Show that

$$E = \sum_{n=p}^{N-1} \left(s(n) - \sum_{k=1}^{p} a_k s(n-k) \right)^2 .$$

[3]

Solution: E is the sum squared error between the true and predicted speech over the range of predicted samples. The formula to compute the prediction must be given.

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

ii) In the above expression for E, explain the reason why the lower limit of the first summation begins at $n = p$. [2]

Solution: The predictor requires p previous samples and therefore the number of predicted samples is $N - p$ beginning with sample p.

iii) Show that the prediction coefficients that minimize E satisfy the equation

$$\mathbf{Ra} = \mathbf{b}$$

and, for this solution, give expressions for the elements of matrix $\mathbf{R}$ and vectors $\mathbf{a}$ and $\mathbf{b}$. [4]

Solution:

$$\frac{\partial E}{\partial a_k} = \sum_{n\in\{F\}} \frac{\partial(e^2(n))}{\partial a_k} = \sum_{n\in\{F\}} 2e(n)\frac{\partial e(n)}{\partial a_k} = -\sum_{n\in\{F\}} 2e(n)s(n-k)$$

The solution must satisfy

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

leading to

$$\sum_{k=1}^{p} a_k \sum_{n\in\{F\}} s(n-k)s(n-j) = \sum_{n\in\{F\}} s(n)s(n-j)$$

so that

$$\sum_{k=1}^{p} r_{jk}a_k = r_{j0}$$

with

$$r_{jk} = \sum_{n\in\{F\}} s(n-j)s(n-k).$$

The matix **R** has element j,k given by r_{jk}. Vector **b** has elements r_{j0} and vector **a** has elements a_k.

b) i) State the main application of Line Spectral Frequencies (LSFs) in the context of speech processing and explain the advantages and disadvantages of LSFs in that application. [2]

Solution:
LSFs are used for speech coding applications principally. LSFs are expensive to compute but very good for quantization in transmission systems.

ii) Consider a particular segment of speech for which the vocal tract transfer function is $V(z)$. Describe how LSFs would be obtained from $V(z)$. Include the mathematical details in your description. [4]

Solution:
LSFs 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 $f_1,\ f_2,\ \ldots$.

iii) Consider the vocal tract transfer function given by

$$V(z) = \frac{1}{1 - 0.95z^{-1} + 0.45z^{-2}}.$$

Find the LSFs corresponding to $V(z)$ and draw a representative sketch of the LSFs. [5]

Solution:
For this $A(z)$ we have $a_1 = 0.95$ and $a_2 = -0.45$ so

$$P(z) = 1 - 0.5z^{-1} - 0.5z^{-2} + z^{-3}$$

and

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

with roots given by

$$P(z) = (z+1)(z^2 - 1.5z + 1) = 0$$

giving $z = -1$ and $z = 0.75 \pm j0.6614$

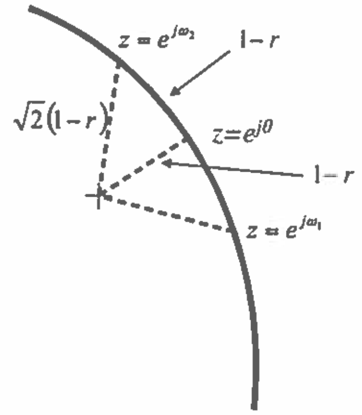
$$Q(z) = (z-1)(z^2 - 0.4z + 1) = 0$$

giving $z = +1$ and $z = 0.2 \pm j0.9798$.

The plot of the LSFs should show the roots of $P(z)$ and $Q(z)$, that they are interleaved and on the unit circle, having normalized frequencies of:
Roots of P: 0.1150, 0.5
Roots of Q: 0, 0.2180.

4. a) i) Consider a single complex pole forming part of an all-pole system function $H(z)$. This single complex pole has radius $r < 1$. Show that the 3 dB bandwidth b of the corresponding resonant peak in the magnitude frequency response can be approximated by $b = 2(1-r)$. Include an illustrative sketch and state the units of b. [3]

Solution:
The approximation comes from the geometric relationships for the 3 dB bandwidth arising from the right angle triangle shown.



ii) Consider a linear time-invariant system with system function

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{a_0 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + a_4 z^{-4}}.$$

Show how the process known as *bandwidth expansion* can be applied to $H(z)$. Describe the main application of such bandwidth expansion in the context of speech processing. [3]

Solution:
From the original $H(z)$, we can form a new filter by multiplying coefficients a_i and b_i by k^i for some $k < 1$ to give

$$G(z) = H(z/k) = \frac{b_0 + b_1 k z^{-1} + b_2 k^2 z^{-2}}{a_0 + a_1 k z^{-1} + a_2 k^2 z^{-2} + a_3 k^3 z^{-3} + a_4 k^4 z^{-4}}$$

such that if $H(z)$ has a pole/zero at z_0 then $G(z)$ will have one at kz_0 so that all poles and zeros will be moved towards the origin by a factor k.

iii) With reference to part (a)(i), apply bandwidth expansion using the approximate expression for b as given in part a) to write a new expression for the bandwidth of a resonant peak after bandwidth expansion. Clearly define all symbols in the new expression. [3]

Solution:
Given a pole in $H(z)$ with bandwidth $b = 2(1-r)$, the corresponding pole after bandwidth expansion has bandwidth

$$2(1-kr) = b + 2r(1-k).$$

b) Consider a speech recognition system using mel-frequency cepstrum coefficients, $\mathbf{c}_t$ as features, computed in appropriately short time-frames t, with

$$\mathbf{c}_t = [c_{t,0},\ c_{t,1},\ \ldots,\ c_{t,P}].$$

i) Show mathematically how the mel-frequency cepstrum coefficients are calculated. [4]

Solution:
At each time-frame t

$$MF_{t,r} = \frac{1}{A_r} \sum_{k=L_r}^{U_r} |V_r((k)X_t(k)|^2 \quad r = 1,2,\ldots,R$$

Where $X_t(k)$ is the DFT of the speech signal in time-frame t and frequency index k, R is the number of filters in the Mel filterbank and $V_r(k)$ is the weighting factor for filter r which spans DFT bins L_r to U_r and A_r is a normalization term

$$A_r = \sum_{k=L_r}^{U_r} |V_r(k)|^2.$$

Then

$$c_{t,p} = \frac{1}{R} \sum_{r=1}^{R} \log(MF_{t,r}) \cos\left(\frac{2\pi p}{R}\left(r + \frac{1}{2}\right)\right).$$

ii) It is decided also to compute the first-order time derivatives of the mel-frequency cepstrum coefficients. The simple difference approximation of the first-order time derivative can be written

$$\Delta c_{t,p} = c_{t,p} - c_{t-1,p}$$

for $p = 0,\ 1,\ \ldots,\ P$. However, this was found to be too inaccurate.

Propose and derive a formula for an alternative improved scheme to find $\Delta c_{t,p}$ [3]

Solution:
A first-order linear fitting operation can be used. Considering T points before and following t to give a total of $2T + 1$ points we the derivative as

$$\Delta c_{t,p} = \frac{\sum_{t=-T}^{T} t c_{t,p}}{\sum_{t=-T}^{T} t^2}$$

for $p = 0,\ 1,\ \ldots,\ P$.
Typical value: $T = 5$.

iii) State the 2 important advantages of using time-derivates of the features in a speech recognition system. [4]

Solution:

- It is advantageous to represent some aspects of the dynamic nature of speech signals.
- The differencing operation removes any effects of simple linear filtering such as the channel response.