EE4-14 Speech Processing Solutions Summer '20 v1

Speech Processing (Imperial College London)

Speech Processing

1. Let $s(n)$ be a speech signal with discrete time index n and with sample values distributed according to a continuous probability density function $p(s)$. Now consider a process $q(s)$ that quantizes $s(n)$ to give a quantized version, $s_q(n)$, using uniform quantization with N quantization levels s_i for $i = 1, 2, ..., N$.

   a) i) Explain the difference between uniform and non-uniform quantization of signals.

   ii) Explain whether uniform or non-uniform quantization is more advantageous for speech signals, and give reasons. [2]

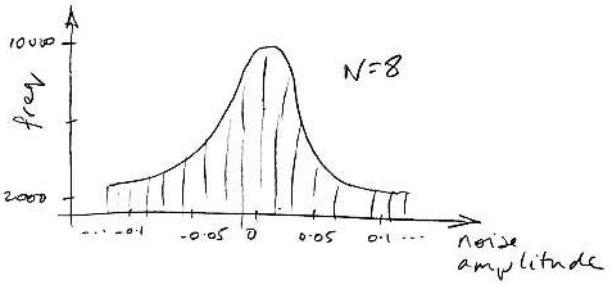
Solution: The difference concerns the distribution of quantization levels. It is most advantageous in terms of minimizing quantization noise when the quantization levels are distributed according to the pdf of the signal sample values. Hence, for typical speech, non-uniform quantization will give lower quantization error, particularly when N is relatively small.

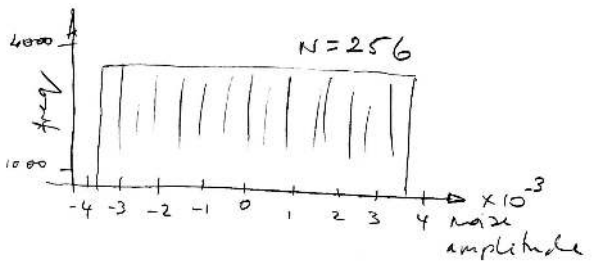
   b) Sketch labelled illustrative histograms of the samples of quantization noise obtained when uniform quantization is applied to a typical speech signal using

   i) $N = 8$,

   ii) $N = 256$. [4]

Solution:





   c) Derive an expression for the mean square quantization noise in $s_q(n)$. [4]

Solution: The range of signal from $\frac{1}{2}(s_{i-1}+s_i)$ to $\frac{1}{2}(s_i+s_{i+1})$ will be quantized to the value s_i. These limits are a_i and b_i respectively. MSE is then

$$E = \int_{-\infty}^{\infty} (s-q(s))^2 p(s)\, ds\,.$$

Splitting this integral into a sum of integrals over ranges for which $q(s)$ is constant gives

$$E = \sum_{i=1}^{N} \int_{a_i}^{b_i} (s-s_i)^2 p(s)\, ds\,.$$

d) Show that the mean square quantization noise is minimized when

$$\int_{(s_{i-1}+s_i)/2}^{(s_i+s_{i+1})/2} (s-s_i) p(s)\, ds = 0.$$

[4]

Solution: Considering the partial derivatives

$$\frac{\partial E}{\partial s_i} = \frac{1}{2}\left((b_{i-1}-s_{i-1})^2 p(b_{i-1}) - (a_i-s_i)^2 p(a_i) + (b_i-s_i)^2 p(b_i) - (a_{i+1}-s_{i+1})^2 p(a_{i+1})\right) - 2\int_{a_i}^{b_i} (s-s_i) p(s)\, ds$$

and

$$\frac{\partial b_{i-1}}{\partial s_i} = \frac{\partial a_i}{\partial s_i} = \frac{\partial b_i}{\partial s_i} = \frac{\partial a_{i+1}}{\partial s_i} = \frac{1}{2}$$

and

$$a_i = b_{i-1} = \frac{s_i + s_{i-1}}{2},$$

it can be seen that the terms in the expression for $\frac{\partial E}{\partial s_i}$ sum to zero except for the integral term.

e) Consider the case of $N = 3$ with

$$p(s) = \begin{cases} 1-|s| & \text{for } |s| \le 1, \\ 0 & \text{for } |s| > 1. \end{cases}$$

Find the quantization levels s_i giving minimum mean square quantization noise. [6]

Solution: Start by noting that $p(s)$ is symmetric around the origin and therefore the choice of quantization levels should also be symmetric, written as $\{-a, 0, a\}$ with a to be determined.
This problem can then be formulated, taking into account $p(s)$ and the expression for the quantization levels from part d), leads to finding the value of a that satisfies

$$\int_{a/2}^{1} (s-a)(1-s)\,ds = 0.$$

This leads to

$$2a^3 - 9a^2 + 12a - 4 = 2(a-0.5)(a-2)(a-2) = 0$$

with the only possible solution $a = 0.5$.

2. a) Let r_k for $k = 1$ to N denote the reflection coefficients in the lossless tube model of the speech production system, and r_G denote the reflection coefficient at the glottis.

i) Show that the transfer function of this model is given by

$$V(z) = \frac{0.5(1+r_G)\prod_{k=1}^{N}(1+r_k)z^{-N/2}}{D(z)}.$$

Give full details of your derivation. [2]

ii) Also derive an expression for the denominator $D(z)$ in terms of the reflection coefficients. [1]

Solution:
The length of each segment is the distance travelled by propagation of sound in 0.5 samples.
Segment delays are modelled by

$$\begin{bmatrix} U \\ V \end{bmatrix} = z^{+1/2} \begin{bmatrix} 1 & 0 \\ 0 & z^{-1} \end{bmatrix} \begin{bmatrix} W \\ X \end{bmatrix}.$$

Segment junctions are modelled by

$$\begin{bmatrix} U \\ V \end{bmatrix} = \frac{1}{1+r} \begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix} \begin{bmatrix} W \\ X \end{bmatrix}.$$

Combining these in cascade gives rise to a product and the given expression follows.
The expression for $D(z)$ follows from this as

$$D(z) = \begin{bmatrix} 1 & -r_G \end{bmatrix} \begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & -r_N \\ -r_N z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

b) Now consider the case when $r_G = 1$. In this case, $D(z)$ can be written $D(z) = \mathbf{P}_N \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, where bold typeface indicates vectors and matrices.

i) What does $r_G = 1$ imply for the physical system represented by the lossless tube model? Include an explanation in terms of characteristic acoustic impedance. [3]

ii) Find a recursive expression for $\mathbf{P}_k$ in terms of $\mathbf{P}_{k-1}$ for $k = 1, \ldots, N$. [2]

iii) Hence deduce a recursive expression for $D_k(z)$ in terms of $D_{k-1}(z)$. [2]

iv) Exploit the recursive expression to write out $D_1(z)$, $D_2(z)$ and $D_3(z)$. [2]

Solution: The case of $r_G = 1$ implies that the glottis is closed. The characteristic acoustic impedance goes to infinity as the cross-sectional area goes to zero.
Because we are seeking a recursive solution, and using $r_G = 1$, let us start with

$$\mathbf{P}_1 = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -r_1 \\ -r_1 z^{-1} & z^{-1} \end{bmatrix} = \begin{bmatrix} (1+r_1 z^{-1}), & -(r_1 + z^{-1}) \end{bmatrix}.$$

It follows that

$$\mathbf{P}_2 = \mathbf{P}_1 \begin{bmatrix} 1 & -r_2 \\ -r_2 z^{-1} & z^{-1} \end{bmatrix}.$$

By induction

$$\mathbf{P}_k = \mathbf{P}_{k-1} \begin{bmatrix} 1 & -r_k \\ -r_k z^{-1} & z^{-1} \end{bmatrix} \quad k = 1, \ldots, N.$$

Now concerning D, starting with $D_1(z) = 1 + r_1 z^{-1}$ it can be seen that

$$\mathbf{P}_1 = \begin{bmatrix} D_1(z), & -z^{-1} D_1(z^{-1}) \end{bmatrix}$$

and by induction similarly

$$\mathbf{P}_k = \begin{bmatrix} D_k(z), & -z^{-k} D_k(z^{-1}) \end{bmatrix}.$$

Substituting for $\mathbf{P}_2$, we find

$$\mathbf{P}_2 = \begin{bmatrix} D_1(z) + r_2 z^{-2} D_1(z^{-1}), & -r_2 D_1(z) - z^{-2} D_1(z^{-1}) \end{bmatrix}$$

which can be written as

$$\mathbf{P}_2 = \begin{bmatrix} D_2(z) & -z^{-2} D_2(z^{-1}) \end{bmatrix}$$

with

$$D_2(z) = D_1(z) + r_2 z^{-2} D_1(z^{-1}).$$

So we finally obtain

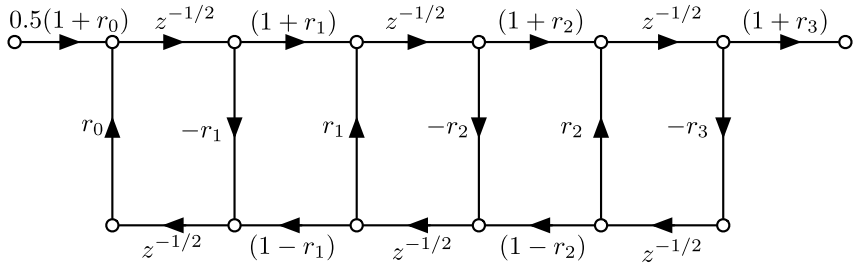
$$D_k(z) = D_{k-1}(z) + r_k z^{-k} D_{k-1}(z^{-1}).$$

This leads to

$$D_1(z) = 1 + r_1 z^{-1}$$
$$D_2(z) = 1 + (r_1 + r_1 r_2) z^{-1} + r_2 z^{-2}$$
$$D_3(z) = 1 + (r_1 + r_1 r_2 + r_2 r_3) z^{-1} + (r_2 + r_1 r_3 + r_1 r_2 r_3) z^{-2} + r_3 z^{-3}.$$

c) Ignoring latency, draw a labelled signal flow graph corresponding to the lossless tube model of order 3 using delays corresponding only to half a sampling period. [3]

Solution:



d) Using the recursion of part (b) in reverse order, find an expression for $D_{k-1}(z)$ in terms of $D_k(z)$, clearly indicating the range of k. [5]

Solution:

$$D_{k-1}(z) = \frac{D_k(z) - r_k D_k(z^{-1}) z^{-k}}{1 - r_k^2} \qquad k = N, N-1, \ldots, 2.$$

3. a) A speech signal s is processed to give

$$S(k,m) = \sum_{i=0}^{N-1} w(i)s(m-i)\exp\left(-j\frac{2\pi}{N}k(m-i)\right).$$

i) Explain this processing in overview.

ii) Define m, k and w.

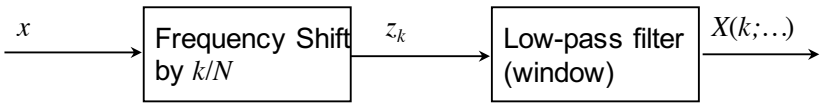
iii) Explain the purpose of the term $(m-i)$ in the exponent.

iv) State the frequency resolution.

v) For any particular m, how many terms of $S(k,m)$ are independent?

vi) Consider a particular value of k. Explain the effect of w. Support you explanation with any appropriate formulae and/or diagrams. [7]

Solution: $S(k,m)$ is a time-frequency analysis centred at sample $m-(N-1)/2$. The $(m-i)$ term in the exponent means that the phase origin remains consistent by cancelling out the linear phase shift introduced by a delay of m samples. m is the last sample in the analysis window, k is the frequency index, w is the window function. The frequency resolution is $1/N$ Hz (normalized units). There are $N/2+1$ independent frequency values.
At a particular k, the k^{th} frequency bin is a filtered version of $z_k(r) = s(r)\exp(-j\frac{2\pi}{N}kr)$ in which the filter has an impulse response of $w(i)$. The term $z_k(r)$ is a frequency shifted version of $s(r)$.



b) In a real-world application of speech processing, a signal model is used comprising $s(n)$ representing a speech signal and $\nu(n)$ representing additive noise. The signal model is written in the time domain as $y(n) = s(n)+\nu(n)$ and in the frequency domain as $Y(l,k) = S(l,k)+V(l,k)$ for time-frame index l and frequency index k.

i) Explain what is meant by spectral variance of a signal in this context. [2]

ii) Write down the mathematical definition for the spectral variance of $y(n)$, denoted $\phi_y(l,k)$. [2]

Solution: For any signal $x(n)$ with time-frequency domain representation $X(l,k)$, the spectral variance is a measure of the signal power at time-frame l and frequency k.

$$\phi_y(l,k) = E\left[|Y(l,k)|^2\right]$$

c) Consider a device which performs noise reduction on a noisy speech signal. The device uses the method of amplitude spectral subtraction.

i) Show that the output signal of the device at time-frame l and frequency index k can be written

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

and write down an expression for $H(l,k)$ in terms of $\hat{\phi}_v(l,k)$.
[4]

ii) Describe and explain the minimum statistics method for determining the estimate $\hat{\phi}_v(l,k)$. State the assumptions that are required for this estimation method to be accurate. State any specific advantages of the minimum statistics method. [5]

Solution:
For amplitude spectral subtraction, we have

$$\begin{aligned} Z(l,k) &= Y(l,k) - \sqrt{\hat{\phi}_v(l,k)} \\ &= Y(l,k)\left(1 - \frac{\sqrt{\hat{\phi}_v(l,k)}}{Y(l,k)}\right) \\ &= Y(l,k)H(l,k) \end{aligned}$$

Minimum statistics approach: This technique is based on the assumption that during a speech pause, or within brief periods between words and even syllables, the speech energy is close to zero. As a result, a short-term power spectrum estimate of the noisy signal, even during speech activity, decays frequently due to the noise power. Thus, by tracking the temporal spectral minimum without distinguishing between speech presence and speech absence, the noise power in a specific frequency band can be estimated.

$$\begin{aligned} \hat{\phi}_y(l,k) &= \alpha\hat{\phi}_y(l-1,k) + (1-\alpha)|Y(l,k)|^2 \\ \hat{\phi}_v(l,k) &= \min\left\{\hat{\phi}_y(l,k),\ \hat{\phi}_y(l-1,k),\ \ldots,\ \hat{\phi}_y(l-D+1,k)\right\} \end{aligned}$$

Advantage: can normally track time-varying noise well, and better than approaches based on VAD.

4. a) i) Using a few sentences and illustrative examples, explain the difference between a phone and a phoneme. [2]

ii) Explain how an automatic speech recognizer could be designed to be robust to coarticulation. [2]

Solution:
A phoneme can be thought of a the set phones for which the meaning of a word using that phoneme doesn't change for any phone in the set.

The answer should point to the following factors to aid robustness: use of triphone models instead of uniphone models to capture explicitly the coarticulation; adequate representation in the training data of all relevant forms of coarticulation.

b) An example of an isolated word speech recognition system uses S states for each word. A hidden Markov model (HMM) is applied for which the transition probability from any state i to next state j is denoted a_{ij}. The speech signal to be recognized is segmented into frames with each frame represented by a feature vector $\mathbf{x}_i$ for $i = 1, 2, \ldots, T$.

i) Describe an appropriate training procedure for this HMM.

Include in your answer a definition of what data is required by the training procedure, any relevant mathematical analysis and supporting diagrams. [5]

ii) For a particular state q in the HMM, let the probability of a transition from state q to the next state be denoted p. Derive an expression for the average time duration (measured in frames) spent in state q. Show your working. [4]

iii) A simplified speech recognition system is designed using an HMM with 3 states. The state diagram of the HMM is shown in Figure 4.1 in which the labels on the arrows indicate the state transition probabilities.

In an experiment, this speech recognition system receives as input the feature vectors

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_5.$$

Table 1 shows the output probability densities of the feature vectors for each state.

Let

$$B(s,t)$$

be the maximum probability density that the model generates the sequence of feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t\}$ from any sequence of states for which frame 1 is in state 1 and frame t is in state s.

Construct a lattice diagram showing all the feasible paths from frame 1 being in state 1, denoted $(1,1)$, to frame 5 being in state 3, denoted $(3,5)$.

Determine the maximum probability alignment of the feature vectors to the states of the HMM and the value of $B(3,5)$ for this alignment. [7]

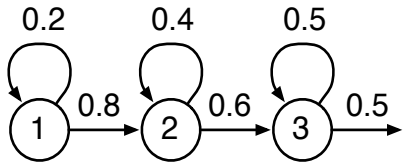


Figure 4.1: Hidden Markov model.

	frame $\mathbf{x}_1$	frame $\mathbf{x}_2$	frame $\mathbf{x}_3$	frame $\mathbf{x}_4$	frame $\mathbf{x}_5$
state 1	0.5	0.2	0.6	0.4	0.5
state 2	0.5	0.7	0.3	0.1	0.5
state 3	0.5	0.5	0.1	0.6	0.6

Table 1: Output probabilities.

Solution:
(i) The training procedure should determine the state transition probabilities and the probability distributions of the features in each state. The probability distributions are typically assumed to be Gaussian and are represented in such cases by the mean and variance for each feature.
The training procedure comprises a first step of initial alignment such as uniform partition of frames to states or a partitioning based on the peaks in the Euclidean distance between neighbouring frames in the feature space. This is followed by a second step of re-estimation that could be performed by Viterbi re-estimation and/or Baum-Welch re-estimation. For full marks, students are expected to give complete descriptions of (at least) one of these methods including supporting analysis (bookwork). Additional credit is given for consideration of triphone models in preference to phones.
The required data is the training data itself (audio) and importantly should be accompanied by the associated labels.

Solution:
(ii) Let the duration in frames be denoted D. Then

$$pr(D=n) = p(1-p)^{n-1} \Longrightarrow E(D) = \sum_{n=1}^{\infty} np(1-p)^{n-1}$$

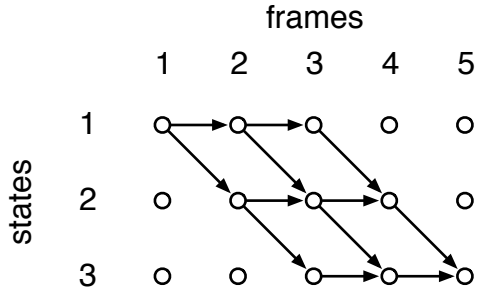
Using differentiation of both sides:

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \Longrightarrow \sum_{n=0}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2}$$

leads to the expression

$$pr(D=n) = \frac{1}{p}.$$

Solution:
(iii) The lattice is drawn as



The best probabilities for the lattice can be computed as

$B(1,1) = 0.5$
$B(1,2) = 0.5 \times 0.2 \times 0.2 = 0.02$
$B(2,2) = 0.5 \times 0.8 \times 0.7 = \mathbf{0.28}$
$B(1,3) = 0.02 \times 0.2 \times 0.6 = 0.0024$
$B(2,3) = \max(0.28 \times 0.4 \times 0.3, 0.02 \times 0.8 \times 0.3) = \mathbf{0.0336}$
$B(3,3) = 0.28 \times 0.6 \times 0.1 = 0.0168$
$B(2,4) = 0.0336 \times 0.4 \times 0.1 = 0.0013$
$B(3,4) = \max(0.0336 \times 0.6 \times 0.6, 0.0168 \times 0.5 \times 0.6) = \mathbf{0.0121}$
$B(3,5) = \max(0.0013 \times 0.6 \times 0.6, 0.0121 \times 0.5 \times 0.6) = 0.0036$

and the alignments path for the five frames is states $\{1,2,2,3,3\}$.