CARE, IIT Delhi
CRL707 Human and Machine Speech Communication (2019-20/I)
Assignment 5

- **Due date for computer assignment: 25th October 2019.**

1. **Pre-emphasis**. A pre-emphasis filter is used to reduce the spectral dynamic range (difference in log magnitude between maximum and minimum value of the spectral magnitude) of a speech signal. A common form for the transfer function of this filter, operating on the input speech is $H(z) = (1 - az^{-1})/G$, where G is a constant.
   a. If we decide the gain of the filter at the highest frequency present in the discrete-time input signal should be 9 times as great as the gain at DC, i.e. $\left|H(e^{j\pi})\right| = 9\left|H(e^{j0})\right|$, what should the value of *a* be? Explain.
   b. If it is desired that the total energy of the filter impulse response, $h(n)$, be unity, i.e., $\sum_{-\infty}^{\infty}\left|h(n)\right|^2 = 1$, what should the value of G be? Explain.
   c. Rather than use a fixed constant for *a*, it is often desirable to make *a* adaptive to the local character of the speech signal $x(n)$. One way is to treat $H(z)$ as a simple first-order LP inverse filter of the signal (thus making an attempt to whiten or flatten the spectrum to the limited extent possible with a first order FIR filter), in this way reducing the spectral dynamic range. Assume that $x(n)$ is already windowed so that it is non-zero over the range $0 \le n \le N-1$. Derive an expression for the filter coefficient *a* in terms of the values of *x(n)* and/or its autocorrelation coefficients using the total mean-squared error criterion by minimizing:

$$E = \sum_{-\infty}^{\infty} e^2(n)$$

   d. Following the spirit of part (c), the gain of the filter can also be determined adaptively. Specifically, use the criterion that the total energy in the output signal is normalized to unity, i.e.,

$$\sum_{-\infty}^{\infty} y^2(n) = 1$$

   Determine the gain in terms of $x(n)$ and/or its autocorrelation coefficients.

2. **Right filter, wrong input**. Let *x(n)* and *y(n)* be two different windowed speech segments and let the corresponding m[th] order optimal LP coefficients (from autocorrelation analysis) be denoted by a_i and b_i respectively. Let $\alpha_x = (-1, a_1, a_2, \ldots, a_m)^T$ and $\alpha_y = (-1, b_1, b_2, \ldots, b_m)^T$. Let R_x and R_y denote the autocorrelation matrices of order (m+1) for *x(n)* and *y(n)* respectively.
   a. Show that the sum squared prediction error for optimal m[th] order LP of x(n) is givenby $D = a_x^T R_x a_x$.
   b. If *y(n)* is applied as input to the LP inverse filter of m[th] order that is optimal for *x(n)*, show the output energy (sum squared values of output) is given by $F = a_x^T R_y a_x$.

c. What range of values can the ratio F / D taken on? In what way would this ratio indicate the degree of similarity between the spectral envelopes of the two speech segments?

d. Show that $F = r_y(0)r_a(0) + 2\sum_{i=1}^{m} r_y(i)r_a(i)$, where $r_y(j)$ are the autocorrelation values of $y(n)$ and $r_a(j) = \sum_{0}^{m-j} a_i a_{i+j}$ where a_o = -1.

3. **Algebraic codebook in G.729**. The fixed codebook in the 8 Kbps G.729 speech coding algorithm is based on an algebraic codebook structure in which each codebook vector contains 4 non-zero pulses. Each pulse can have either the amplitudes +1 or -1 and can assume the positions given in the accompanying table.

Structure of fixed codebook C

Pulse	*Sign*	*Positions*
i0	s0	0, 5, 10, 15, 20, 25, 30, 35
i1	s1	1, 6, 11, 16, 21, 26, 31, 36
i2	s2	2, 7, 12, 17, 22, 27, 32, 37
i3	s3	3, 8, 13, 18, 23, 28, 33, 38 4, 9, 14, 19, 24, 29, 34, 39

The codebook vector $c(n)$ is constructed by taking a zero vector, and putting the 4 unit pulses at the found locations, multiplied with their corresponding sign

$$c(n) = \sum_{j=0}^{j=3} s_i \delta(n - i_j) \quad n = 0,\ldots,39 \qquad \text{(A)}$$

where $\delta(n)$ is a unit pulse (and the subframe length is 40 samples). The fixed codebook is searched by minimizing the mean squared error between the weighted input speech and the weighted reconstructed speech. The target signal $x(n)$ used in the closed-loop pitch search is updated by subtracting the adaptive codebook contribution. That is,

$$x_2(n) = x(n) - g_p y(n), \quad n = 0,...,39$$

where $y(n)$ is the filtered adaptive codebook vector, and g_p is the corresponding gain. The matrix $\mathbf{H}$ is defined as the lower triangular Toeplitz convolution matrix with diagonal $h(0)$ and lower diagonals $h(1),..., h(39)$. If c_k is the algebraic codevector at index k, then the codebook is searched by maximizing the term

$$\frac{C_k^2}{E_k} = \frac{\left(\sum_{n=0}^{39} d(n)c_k(n)\right)^2}{\mathbf{c_k^t \Phi c_k}} \qquad \text{(B)}$$

where $d(n)$ is the correlation between the target signal $x_2(n)$ and the impulse response $h(n)$, and $\mathbf{\Phi = H^t H}$ is the matrix of correlations of $h(n)$. The signal $d(n)$ and the matrix $\mathbf{\Phi}$ are computed before the codebook search. The elements of d(n) are computed from

$$d(n) = \sum_{i=n}^{39} x_2(i)h(i - n), \; n = 0,..,39 \qquad \text{(C)}$$

and the elements of the symmetric matrix $\mathbf{\Phi}$ are computed by

$$\Phi(i, j) = \sum_{n=j}^{39} h(n-i)h(n-j), \; (j \geq i) \qquad \text{(D)}$$

a. Show that the numerator term on right hand side of equation (B) is equivalent to $(x_2^t \cdot w_j)^2$ where w_j is the filtered codevector c_k and x_2 is the target vector with entries $x_2(n)$. Only the elements actually needed are computed and an efficient

storage procedure speeds up the search procedure as detailed below. The algebraic structure of the code book *C* allows for a fast search procedure since the code book vector c_k contains only four nonzero pulses.

b. Show that the correlation in the numerator of equation (B) for a given vector c_k is given by

$$C = \sum_{i=0}^{3} a_i d(m_i) \quad \text{(E)}$$

where m_i is the position of the i^{th} pulse of c_k and a_i is the amplitude of *d(n)* at the corresponding location.

c. Show that the energy in the denominator of equation (B) is given by

$$E = \sum_{i=0}^{3} \phi(m_i, m_i) + 2\sum_{i=0}^{2} \sum_{j=i+1}^{3} a_i a_j \phi(m_i, m_j) \quad \text{(F)}$$

To simplify the search procedure, the pulse amplitudes are predetermined by quantizing the signal *d(n)*. This is done by setting the amplitude of a pulse at a certain position equal to the sign of *d(n)* at that position. Before the codebook search, the following steps are done. First, the signal d(n) is decomposed into two signals: the absolute signal $d'(n) = |d(n)|$ and the sign signal *sign*[d(*n*)]. Second, the matrix Φ is modified by including the sign information; that is

$$\Phi(i, j) = sign[d(i)]sign[d(j)]\Phi(i, j), \quad i = 0,..., 39, \; j = i,...,39 \quad \text{(G)}$$

To remove the factor 2 in equation (F),

$$\Phi'(i,i) = 0.5\Phi(i,i), \quad i = 0,...,39 \quad \text{(H)}$$

d. Show that the correlation in equation (E) is now given by

$$C = d'(m_o) + d'(m_l) + d'(m_2) + d'(m_3)$$

and,

$$E = \Phi'[(m_0, m_0)] + [\Phi'(m_1, m_1) + \Phi'[m_0, m_1)]$$
$$+ [\Phi'(m_2, m_2) + \Phi'(m_0, m_2) + \Phi'(m_1 + m_2)]$$
$$+ [\Phi'(m_3, m_3) + \Phi'(m_0, m_3) + \Phi'(m_1, m_3) + \Phi'(m_2, m_3)]$$

e. Get an estimate of the total number of multiply-adds required to get the best fixed code book index using this method. Compare this estimate with the multiply-adds that would be required if a brute-force method is used to compute the ratio in equation (B).

4. **Nothing in - something out**. A CELP encoder has a stochastic excitation codebook containing 128 vectors of dimension 4 (the subframe size is 4 samples). There is no adaptive codebook in this coder and no perceptual weighting. The synthesis filter for the current frame (consisting of subframes 26, 27, 28 and 29) is given by $A(z) = 1/(1- (1/4)z^{-2})$. The encoder has been operating for a while and the last completed search has led to the choice of excitation code vector c = (2, -1,0,1) and gain scaling factor g = 2 for subframe 27. The ZIR vector for subframe 27 was found to be (1,0,-1,0). The speech vector for subframe 28 is s = (1,2, -2, -1). Before beginning a new search to approximate this speech vector, the ZIR (zero input response) for subframe 28 is needed.
   a. Find this ZIR vector.
   b. Find the target vector to be used for this search.
   c. Find the matrix H that maps a gain-scaled excitation vector into a zero-state response (ZSR) vector for each subframe in the current frame.

5. **CELP coding.** An analysis-by-synthesis speech coding system uses dimension $k = 2$ excitation vectors selected from the codebook shown below of size $N = 4$. For the current frame of speech (consisting of two subframes or four samples) the synthesis filter (including gain) is given by

$$H(z) = \frac{2}{1 - \frac{1}{2} z^{-1}}$$

   a. Find the explicit input-output relation in the time domain between the output of the synthesis filter *y(n)* and the input *x(n).*
   b. Find the index of the optimal excitation vectors for the first subframe to synthesize the best approximation to the first two samples of the original speech *s(n)* whose first four samples *s(0)* to *s(3)* are respectively (1,0,-1/2,1/4). Assume that the initial state of the filter is zero at time $n = 0$.
   c. Find the zero input response (ZIR) for the second subframe.
   d. Find the index for the best excitation vector for the second subframe.
   e. Specify the complete synthesized speech samples for the entire frame and find the total mean-squared error between the original and synthesized speech for this frame.

Excitation Codebook		
Index	Binary Word	Codevector
1	00	1,0
2	01	0, 1
3	10	-1, 0
4	11	0,-1

6. **Computer Assignment:** Extract 90 ms or four frames of speech data at 22.5 ms per frame from voiced segments of recorded speech in your voice, at 8 KHz sampling rate. For each frame of speech,
   a. **Mimicking LPC-10**: Obtain the parameter values that are computed (and transmitted after quantization) by the LPC-l0 vocoder. Note that there is no need to quantize the parameters.
   b. **Line Spectral Frequencies (LSFs):** For each frame, form the *P(z)* and *Q(z)* polynomials from the inverse filter *A(z),* for each segment. On the same graph (one for each segment), plot the roots of polynomials *P(z)* and *Q(z)* in the z-plane. What are the LSFs in Hz? Plot the spectral envelope 1/*A(z)* for each frame and superimpose the LSFs on the plots? How are the two related?

   Label all axes properly. Explain in brief the calculations for parts (a) and (c).