INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

End-Autumn Semester Examination 2023-24

Advanced Technology Development centre

Subject: INTRODUCTION TO DIGITAL SPEECH PROCESSING **Code: ET60007**

Time: 3:00 Hours Full Marks [10x2+5x16] = 100

Answer all the questions

PART-A

1. Uniform Filter Banks analysis is used to extract the parameters of a speech segment, if the bandwidth of each filter is **100 Hz** and speech signal is recorded with sampling frequency **12 KHz** determine the required number filter to cover the entire spectrum of the speech segment
2. Write the manner of articulation of the phonemes /ʃ/, /g^h/.
3. Number of zero crossing is extracted from ***25ms*** speech segment of a fricative sound and ***25ms*** speech segment of a voiced sound which one has higher number of zero crossing and why?
4. Figure-1 represent the LPC Spectrum of a speech segment determine the order of the required LPC analysis. If 2 poles are used for radiation and 2 poles are used glottal pulse modeling

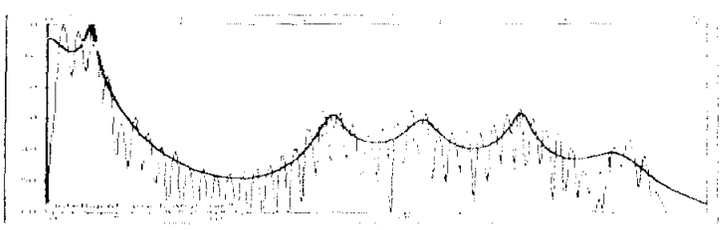

Figure-1

5. A signal is sampled at ***16 KHz, 16 bits***, and encoded with ***16th order LPC***. Each of the LPC coefficients is encoded with ***2 bytes***, Gain in ***2 bytes***. Voiced unvoiced and F_0 information is encoded using ***1 byte***. Calculate the compression ratio if frame rate is ***100 frame /sec***.
6. Write the name of two-time domain methods for F_0 extraction?
7. An audio signal is recorded using the ***F_S= 16 kHz***, encoded with16 bit and recoded in MONO format. To store 80 ms signal in PCM WAV format how much memory is required? If the above signal fundamental frequency is ***250Hz***. How many samples will be in one pitch period?
8. Figure-2 represent plot of the Normalized cross correlation Coefficients of speech segment. If the ***3rd pick*** occurred at ***k= 180*** sample determine the F_0 of the speech segment. Where sampling frequency ***Fs=8KHz***

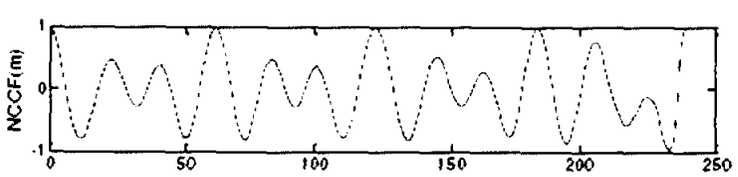

Figure-2

9. An audio signal is recorded using ***Fs=8000***. The fundamental frequency is extracted using window size of ***20 ms.*** If the fundamental frequency is ***125 Hz*** how many periods will be there within the window.
10. Write **two parameters** name which controls the speech prosody?

PART-B

1. (a) The following signal x[n] is sampled with ***10 kHz*** sampling frequency. How many time the signal will be cross the zero line for ***80ms*** segment.

$$x[n] = 10\sin 0.2\pi n + 2sign0.5\pi n$$

(b) 4 PARCOR coefficients {k_1, k_2, k_3, k_4} are extracted from a speech signal frame using autocorrelation method. Find the energy of the liner prediction residual that would obtain by inverse filtering the speech signal frame. The inverse filter is designed using the above PARCOR coefficient. Where speech signal frame has energy E^0=3000 and k_1=0.5, k_2=0.3, k_3= - 0.2, k_4=0.18

(c) If the same speech signal segment is generated using lossless tube modelling and above ***4 PARCOR*** coefficients are used to estimate the vocal tract cross-section area, calculate the value of the cross-section areas of the connected tubes. [Where initial tube cross-section area = ***0.45cm^2***]

2. A causal LTI system has system function is given in equation-1. Equation 2 represents the expression of prediction error filter. Lattice Formulations of Linear Prediction as given in equation 3(a) and 3(b). Where e[m] represents the forward prediction error, b[m] represents the backward prediction error and k_i is the PARCOR coefficient.

(a) If the signal ***s[n] = {1, -2, 3, -4}*** applied in the design error filter ***A(z)*** (as in question no. 2) calculate the value of the forward prediction error at the output of the second lattice.

(b) If the error signal at the output of the second lattice applied in the input of 1/A(z) determine the output signal and draw the signal flow diagram of 1/A(z).

$$H(z) = \frac{A}{1-\sum_{k=1}^{p}\alpha_k z^{-k}} \quad (1) \qquad A(z) = 1-\sum_{k=1}^{p}\alpha_k z^{-k} \quad (2)$$

$$e^i[m] = e^{i-1}[m] - k_i b^{i-1}[m-1] \quad (3a)$$

$$b^i[m] = b^{i-1}[m-1] - k_i e^{i-1}[m] \quad 3(b)$$

$$k_i^{\text{PARCOR}} = \frac{\sum_{m=0}^{L-1+i} e^{i-1}[m]b^{i-1}[m-1]}{\left(\sum_{m=0}^{L-1+i}[e^{i-1}[m]]^2 \sum_{m=0}^{L-1+i}[b^{i-1}[m-1]]^2\right)^{1/2}}$$

3. Linear prediction analysis is used to obtain a ***6-order*** all-pole model for a voiced speech segment which was sampled at ***F_S = 10000 Hz***. The system function of the model is given in equation-1:

$$H(z) = \frac{G}{A(z)} = \frac{G}{1-\sum_{k=1}^{11}\alpha_k z^{-k}} = \frac{G}{\prod_{i=1}^{11}(1-z_i z^{-1})} \qquad 1$$

Table 1 shows 3 pole of the 6-order prediction error filter, A (z).

Table-1

Sl.	Root magnitude	Root angle [in degree]
1	0.91	12
2	0.93	-25
3	0.78	35

(a) Determine where the other three pole of *H(z)*. Plot the pole location in z plane

(b) Determine all the formant frequency and formant bandwidth of the voiced speech segment

(c) Autocorrelation method based ***20th*** order LPC analysis was performed for a voiced speech signal with the frame rate of ***100 frame/sec***. If the length of the window used for this analysis is ***20 ms*** determine length of the error signal [where sampling frequency of the speech signal is ***16 kHz***]

4. (a) Draw the functional block diagram of text to speech conversion system and explain the function of the grapheme to phoneme conversion block.

(b) How does spoken language affect automatic speech recognition? Give an example of two homophones in English.

(c) Determine the number of vowel to consonant transition present in the phonetic representation of the first word of your name

(d) The following equation represents the Fujisaki model for speech prosody which part of the equation related to the accentuation. As per the Fujisaki modelling accentuation is produce due to what kind of movement of vocal card.

$$F_0(t) = \ln(F_b) + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}\{G_a(t - T_{1j}) - G_a(t - T_{2j})\}$$

5. (a) Draw the Cepstral Transform Coefficients (CC) feature extraction block diagram

(b) Write name two speech features which are formant like features but not exactly formant features.

(c) Write two Limitations of the Statistical Approach based automatic speech recognition (ASR) and definition of prosodic word

(d) Figure 3 represents the cepstral plot of a speech segment. The most visually prominent feature in this cepstrum is the peak near quefrency ***7 ms***. Determine the fundamental frequency (F_0) of the speech segment where the sampling frequency (Fs) is ***8 kHz***.

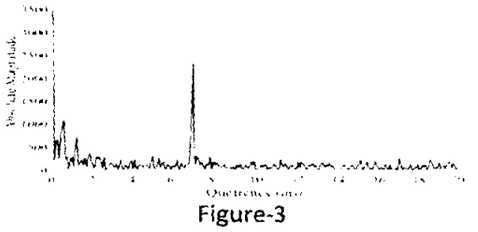

Figure-3