Indian Institute of Technology, Kharagpur
Advanced Technology Development Centre

Class Test-2 (2024-25)

Subject: INTRODUCTION TO DIGITAL SPEECH PROCESSING Code: ET60007

Time: 0:45 Hours **Full Marks: 10+10 =20**

---

1. (a) Using autocorrelation method, a voiced signal frame is analyzed and 3 PARCOR coefficients $\{k_1, k_2, k_3\}$ are calculated. Now, the same speech signal segment is generated using lossless tube modeling. If the cross-sectional area of the first tube section is $2\ cm^2$, the calculate the cross-sectional area of the other tubes. Given $k_1 = 0.62$; $k_2 = -0.15$; $k_3 = 0.46$

   (b) If the voiced signal is *x[n] = {1, 2, 1, -1,2}* order of the LPC analysis is *3* and LPC coefficients are as given in question 1(a) determine the model gain.

2. A causal LTI system has system function is given in equation-1. Equation 2 represents the expression of prediction error filter ***A(z)***. 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 ***kᵢ*** is the PARCOR coefficient

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

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

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

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

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

Indian Institute of Technology, Kharagpur
Advanced Technology Development Centre

Class Test-1(2024)

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

Time: 0:45 Hours **Full Marks: 10+5+5 =20**

1. Consider a two tube lossless vocal tract model. **Draw** the signal flow diagram Including Boundary condition at lips and glottis. The tube cross section area and length are **A_1=1 cm², l_1=9 cm and A_2 = 7 cm² and l_2= 8 cm**. If the glottis and lips are completely lossless determine the transfer function of the system in z domain. Where velocity of sound **c=340m/s** and the sampling frequency F_s=10kHz

$$V(z) = \frac{0.5(1+r_g)(1+r_1)(1+r_2)z^{-1}}{1+(r_1r_2+r_1r_G)z^{-1}+r_2r_gz^{-2}}$$

2. Let length of the vocal tract *l=17cm* and the velocity of sound *c=340m/s* find the number of section required to generate *5 kHz* bandwidth voiced signal

3. The frequency response of a loss less uniform tube is as given in the following equation. The length of the tube ***l=17.5 cm*** and speed of sound ***c=350m/s***. **Draw** the volume velocity vs. Frequency curve for first **4 roots**?

$$\frac{U(l,\Omega)}{U_g(\Omega)} = V_a(\Omega) = \frac{1}{\cos(\Omega l / c)}$$

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Mid-Autumn Semester Examination 2024-25

Advanced Technology Development centre

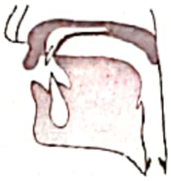
Subject: INTRODUCTION TO DIGITAL SPEECH PROCESSING **Code: ET60007**

Time: 2:00 Hours Full Marks [10x2+5x6] = 50

PART-A(Answer all the questions)

1. An audio signal is recorded using the ***F_S= 16 kHz***, encoded with16 bit MONO format. If the audio signal fundamental frequency is ***250Hz***, then how many samples will be in one pitch period?

2. Following figure represents the production of a consonant. Write the **place of articulation** of the consonant.



3. Time-varying glottal impedance is function of glottal opening AG(t). If the glottis is completely closed, then what will be the value of glottal impedance and volume velocity?

4. Write the ***state of the Glottis*** during the pronunciation of the following phoneme?

   /gʰ/, /o/,/l/,/p/

5. Which of the following pair of tones is perceived as louder tone and why?

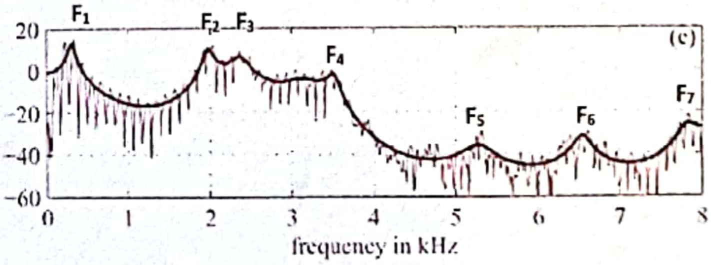
   (a) 20dB level at 600Hz and 20 db at 1.5kHz (b) 5dB level at 1.2 KHz and 5dB level at 8 KHz

6. Suppose an electric fan produces a noise intensity of ***60 dB***. How many times more intense is the sound of a conversation if it produces an intensity of ***80 dB***?

7. Why the women speech has high F_0 and formant frequencies compare to men speech?

8. A uniform tube is closed in both end find out the 2nd and 4th formant frequency if the tube is excited at one end, where the length of tube is ***17.5cm and*** speed of sound ***c=350m/s***.

9. The frequency response of a uniform tube is as given below figure. Determine the number of complex poles in the tube transfer function.



10. Draw an equal loudness curve for ***15 dB***.

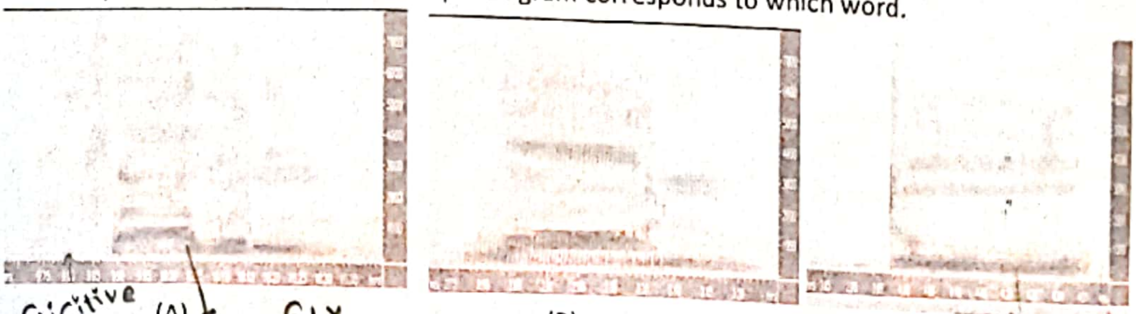
PART-B(Answer all the questions)

1. (a) Write your first name in ***IPA*** and syllabify it. Write the place and manner of articulation of each of the phonemes of your first name.

   (b) Let length of the vocal tract ***l=17cm*** and the velocity of sound ***c=340m/s*** find the number of section required to generate ***6 kHz*** bandwidth voiced signal. If the voiced signal is modeled with all-pole model how many complex conjugate poles will be there?

2. A voiced based telephone dialing system is designed using the following words
[DIAL; STOP; ONE; TWO; THREE; FOUR; FIVE; SIX; SEVEN; EIGHT; NINE; ZERO]

Following figure shows spectrograms of one version of any three words. Using your knowledge of acoustic phonetics, determine which spectrogram corresponds to which word.



fricitive (A) vowel SIX

(B) DIAL

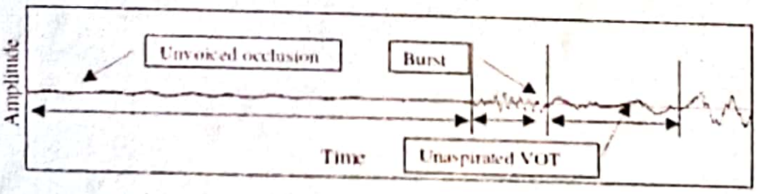
TWO (C)

3. Consider the two tube lossless vocal tract model is used to produce the vowel sounds. Assume that the termination at glottis and lips are completely lossless. The table-1 represents tube parameters for production of vowel ***/o/*** and ***/e/***. Determine the formant frequencies and formant bandwidth of both the vowels when sampling frequency ***F_s=8 kHz***. Where speed of sound ***c=350m/s***

Table-1

Vowel	Tube-1 parameters		Tube-2 parameters	
	Length l_1	Area A_1	Length l_1	Area A_1
/ae/	4 cm	3 cm^2	13 cm	7 cm^2
/a/	8.5 cm	4 cm^2	8.5 cm	6 cm^2

4. (a) The below figure is the time domain representation of a phoneme acoustic signal. Write the manner of articulation of the phoneme



(b) For the ***16-tube model***, determine the number of zeros and the number of poles present in the transfer function of the tube model

(c) Draw a schematic diagram of the upper palate and mark the place of articulation of the following phonemes

i. /p/, ii. /kʰ/, iii. /ɖ/, iv. /ʃ/

5. (a) A signal is sampled at ***20 KHz, 16 bits***, and encoded with ***24th 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 ***50 frame /sec***.

(b) Second formant frequency of a steady state vowel F_2 is ***1200Hz***. Consider that the vowel is produced using a single lossless acoustic tube. What will be the length of the vocal tract? The speed of sound ***c=350 m/s***.

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

End-Autumn Semester Examination 2024-25

Advanced Technology Development centre

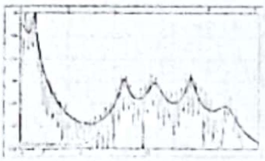
Subject: INTRODUCTION TO DIGITAL SPEECH PROCESSING **Code: ET60007**

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

Answer all the questions

PART-A

1. LPC spectrum of a speech segment is shown in figure 1. Calculate the order of the LPC analysis.



2. Consider a two-tube lossless vocal tract model. Determine reflection coefficient for the case in which tube cross section area and length are A_1= 1 cm^2, l_1=9 cm and A_2 = 7 cm^2 and l_2= 8 cm. Consider velocity of sound c = **340m/s.**
3. If the length of the vocal tract is **17.5 cm** and the velocity of sound is **350 m/sec**, then find the time delay of sound to reach from the glottis to the lips (vocal tract).
4. Acoustic intensity (I) of an audio system is **140dB**. Find out the Loudness (L) in sones where $L = 445\ I^{0.33}$
5. **500Hz** sinusoid signal is sampled at ***16 kHz***. Determine the number of zero crossing in ***50ms*** segment.
6. A ***3*** second audio signal is recorded and stored in a computer in uncompressed format. The size of the store signal is ***48.0 Kbyte***. Calculate the sampling frequency of the store audio signal if it encoded with ***16 bit***.
7. 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.
8. Write the name of the speech perceptual parameters.
9. Write the manner of articulation of the following phonemes.

   /bʰ/, /t/, /k/, /ɖ/
10. 50% overlap uniform Filter Banks analysis is used to extract the parameters of a speech segment, if the bandwidth of each filter is ***200Hz*** and speech signal is recorded with sampling frequency ***12 KHz*** determine the required number filter to cover the entire spectrum of the speech segment.

PART-B

Q1. For a given signal ***x[n] = {1, 2, 1, -1, 2} 3rd*** order LPC analysis is done based on the following set of LPC analysis equation.

(i) Calculate the value of the LPC coefficients {α_1, α_2, α_3}

(ii) Compute the value of model gain. **[10+6]**

$$k_i = \frac{R_n[i] - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R_n[i-j]}{E^{i-1}} \quad (1) \qquad E_n^{(i)} = E_n^{(i-1)}(1-k_i^2) \quad (2)$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \quad (3) \qquad \alpha_i^{(i)} = k_i \quad (4)$$

Q2. (a) Draw the MFCC feature extraction block diagram

(b) Complex cepstrum $\hat{x}(n)$of a digital signal *x[n]* is the inverse Fourier transform of the complex log spectrum

$$\hat{X}(e^{j\omega}) = \log\left|X(e^{j\omega})\right| + j\arg[X(e^{j\omega})] \qquad \hat{x}(n)$$

Show that cepstrum *c[n]* define as the inverse Fourier transform of the log magnitude is the even part of

$$c[n] = \frac{\hat{x}[n] + \hat{x}[-n]}{2}$$

(c) A signal $X_n[e^{jw}]$ is the STFT of a signal $x_n[n]$ using hamming window w[n]. Determine the minimum frequency sampling rate where length of the window ***N_w = 800*** sample, and ***F_s = 16 KHz***.

[3+8+5]

Q3. (a) Draw a functional block diagram of a text to speech conversion system and explain the function of text normalization and grapheme to phoneme conversion block.

(b) What are the different techniques for speech synthesis?

(c) Name the signal segment that required for synthesized your first name using part name based Concatenative synthesizer. **[8+3+5]**

Q4. 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 is the PARCOR coefficient

(a) Draw the signal flow diagram of All-Pole Lattice Filter H(z).

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

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

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

(b) If the error signal ***e[n] = {0.1, 0.3, -0.4, 0.2, 0.3, 0.2}*** applied in the design filter ***H(z)*** (as in question no. 1) determine the output signal of H(z). where ***k_1=0.3, k_2= - 0.5, k_3=0.2*** **[5+11]**

Q5. (a) 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. **[4]**

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

(b) Write two Limitations of the Statistical Approach based automatic speech recognition (ASR) and definition of prosodic word **[3+3]**

(c) What are the differences between a spoken language and written language and how does it affect automatic speech recognition? **[3+3]**