Indian Institute of Technology, Kharagpur
Centre for Educational Technology

End Semester Examination 2018

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

Time: 3:00 Hours PART-A:-10*2=20; PART-B:-5*16=80 **Full Marks =100**

Answer all the questions of PART-A and PART-B

PART-A

1. What is equal loudness curve or phone curve? Draw an equal loudness curve for ***15 dB***.

2. 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***?

3. What is the perceived pitch (in Mels) for the following tones:

   a) 100Hz b)1.50 kHz

4. To produce a voiced speech signal of bandwidth ***6 KHz*** how many section of lossless tubes are required? Where length of the tube is 17.5cm and c=35000cm/s.

5. Write the the place and manner of articulation of the following phonemes?
   (i) */p/*, (ii) /ɖ/, (iii) $/k^h/$ (iv) */l/*

6. ***2 kHz*** sinusoid signal is sampled at ***8 kHz*** determine the number of zero crossing in ***50 ms*** segment

7. Figure-1 in annexure-1 shows a spectrogram of a CVC segment (In annexure-1). Where C represents consonant and V represents the vowel. Segment the following region

   i) Consonant vowel transition, ii) vowel consonant transition and iii) steady-state vowel

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

   (a) 20dB level at 500Hz and 20 db at 200 Hz (b) 5dB level at 1 KHz and 5dB level at 8 KHz

9. Write three time domain methods for extraction of F_0?

10. ***5sec.*** speech segment is encoded using LPC coefficient and the LPC coefficient are extracted for each frame (frame length (L) = 5 pitch period) with a frame rate ***100 frame/s***. How many frame's LPC coefficient can be extract from the above speech signal? Where the F0 of the speech segment is 250 Hz and sampling frequency Fs=16 kHz

PART-B

1. 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 roots of the 6-order prediction error filter, A (z).

Table-1

Sl.	Root magnitude	Root angle
1	0.938	-10.36
2	0.9317	25.88
3	0.7837	35.13

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

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

2. (a) LPC coefficients {α_1, α_2, α_3} are extracted from a signal *x[n] = {1,-2,-1,2}* using 3rd order LPC analysis. If the values of the coefficients are $\alpha_1 = 0.52; \quad \alpha_2 = -0.25; \quad \alpha_3 = 0.36$ compute the model gain.

(b) 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) and draw the signal flow diagram of All-Pole Lattice Filter H(z) and error filter A(z)

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

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

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

Where e[m] represents the forward prediction error, b[m] represents the backward prediction error and k is the PARCOR coefficient

3. (a) Figure-2 in annexure-1 shows plots of 5 speech short-time log magnitude spectra. The set of 5 spectra include vowel and consonant regions by a male, a female and a child speaker.

(i) Which of the 6 spectra are most likely to have been uttered by a child? What leads you to this conclusion?

(ii) Which of the spectra correspond to unvoiced sounds?

(iii) Which of the voiced speech spectra most likely come from an adult male; which from an adult female?

(b) A signal is sampled at 16 KHz, 16 bit, encoded with minimum required LPC order. Each of the LPC coefficients is encoded with 2 byte, Gain in 2 byte. Voiced unvoiced F0 information is encoded using 1 byte. Calculate the compression ratio if frame rate is 100 frame /sec?

4. (a) Draw the block diagram of MFCC parameter extraction method. (b) Write two differences between a spoken language and written language and how does it affect automatic speech recognition? (c) Write the name of three Supra-segmental Speech parameters
5. (a) Draw a functional block diagram of concatinative speech synthesis system and describe the function of Grapheme to Phoneme conversion block (b) Write the phonetic transcription of the last word of your surname and list required di-phone for synthesized your name using di-phone concatinative synthesis method. (c) write two limitations of the Statistical Approach based automatic speech recognition (ASR)

Annexure-1

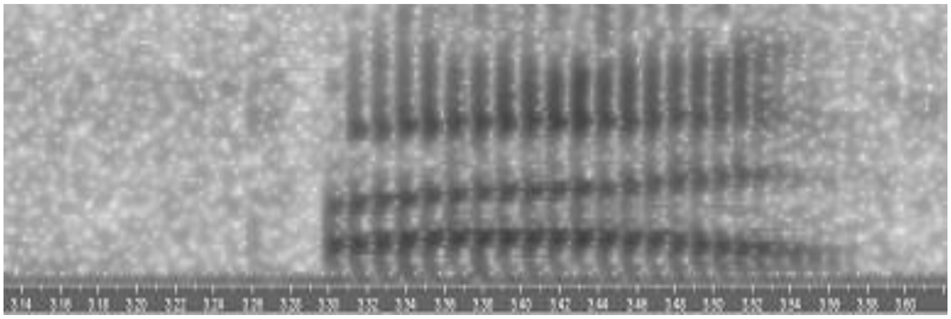
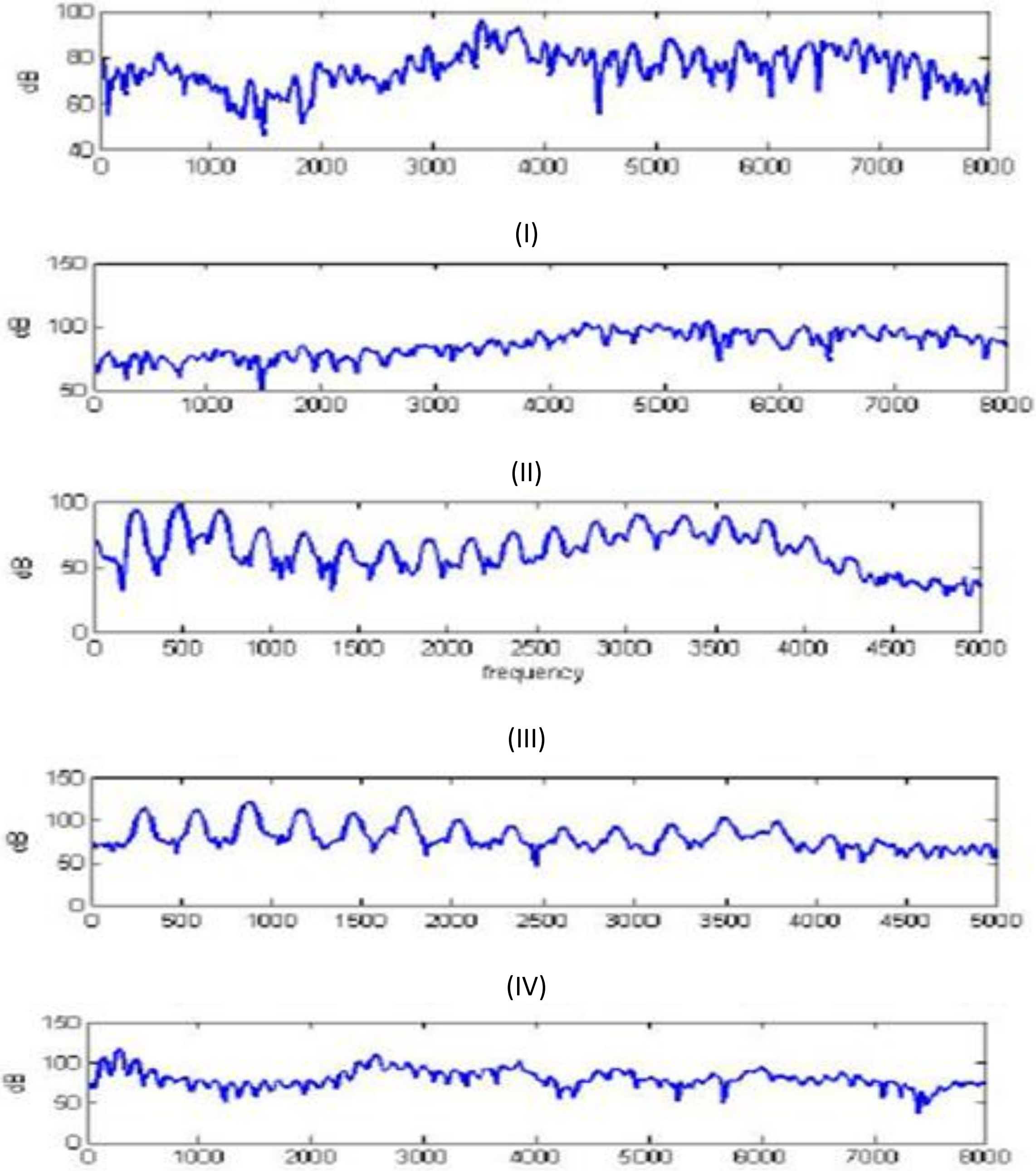


Figure-1



(I)

(II)

(III)

(IV)

(V) Figure-2