Indian Institute of Technology, Kharagpur
Centre for Educational Technology

End Semester Examination 2019

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. Numbers of zero crossing are extracted from ***10ms*** speech segment of two speech signal (segment-A and segment-B). Segment-A has more numbers of zero crossing then segment-B. Which of the segment is most likely to be voiced speech segment and why?

2. The frequency response of a uniform tube is as given in the following equation (1). The length of the tube ***l=17.5*** cm and speed of sound ***c=350m/s***. Determine the first two formant frequency and formant bandwidth.

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

3. ***2 kHz*** sinusoid signal is sampled at ***10 kHz*** determine the number of zero crossing in ***60 ms*** segment

4. Figure-1 represent the LPC Spectrum of a speech segment determine the order of the LPC analysis. If 2 poles are used for radiation and 2 poles are used glottal pulse modeling

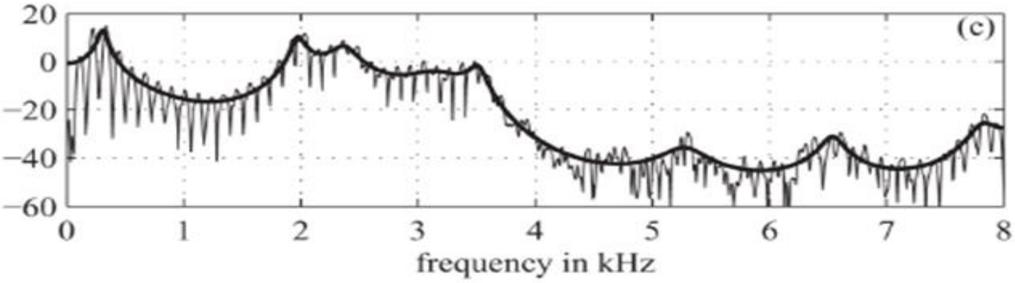


Figure-1

5. Short-Time Fourier Transform Magnitude *|S(nL,ω)|* is compute for a speech signal segment with time decimation rate ***L=128*** sample. If the signal is recover with modify decimation rate ***M=32*** sample. Determine the speed-up of factor.

6. STFT analysis of a speech segment is required for noise reduction. If the STFT analysis is done based on the rectangular Window of length ***20 ms*** determine the maximum possible temporal decimation factor so that signal is completely invertible. Where sampling frequency ***F_s=16 kHz***

7. Write the manner of articulation of the phonemes **/p/, /t^h/, /s/, /n/.**

8. Non-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 ***16 KHz*** determine the required number filter to cover the entire spectrum of the speech segment.

9. Write two advantages of having two Ears for sound perception

10. Write the name of the three Supra-segmental speech parameters that control the speech prosody

PART-B

1. A speech signal frame has energy $E_n^0 = 5000$ using the autocorrelation method the frame is analyzed and 3 PARCOR coefficients $k_1 = 0.6; \quad k_2 = -0.4; \quad k_3 = 0.2$ are extracted. **[7+4+5]**

(a) If the same speech signal segment is generated using lossless tube modelling and above 3 PARCOR coefficients are used to estimate the vocal tract cross-section area, calculate the value of thecross-section areas of the connected tubes. [Where initial tube cross-section area = ***0.65cm²***]

(b) Figure-2 represent plot of the Normalized cross correlation Coefficients of speech segment. If the ***L= 180*** sample determine the F$_0$ of the speech segment. Where sampling frequency ***Fs=16*** KHz

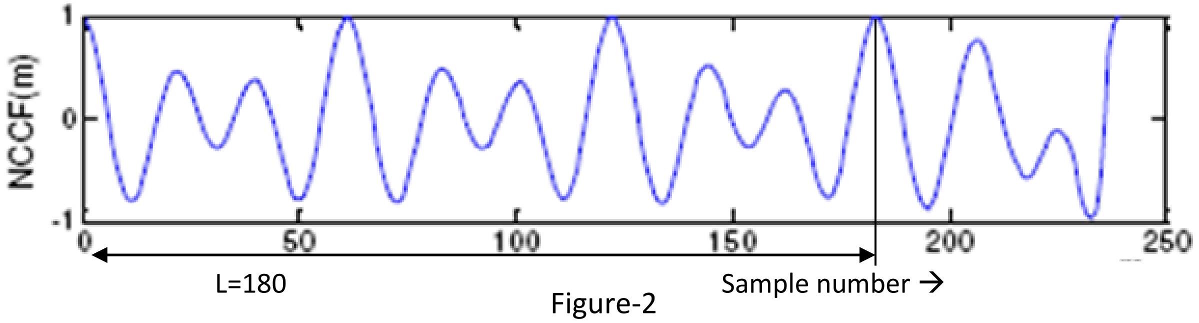


Figure-2

(c) Length of a vocal tract is ***17.5 cm*** and the speed of sound ***c=350 m/s***. Determine the number of tube sections required to produce a voice of ***6 kHz*** bandwidth.

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) **[6+10]**
Where *e[m]* represents the forward prediction error, *b[m]* represents the backward prediction error and *k$_i$* is the PARCOR coefficient

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

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

(a) Draw the signal flow diagram of the Filter ***H (z).***

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

Where

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

3. (a) Draw the MFCC feature extraction block diagram. Write one reason regarding the use of delta and double delta MFCC features for speech signal classification. [6]

(b) MFCC features are extracted from a speech signal if the speech signal is sampled at ***16 kHz*** and initial filter bandwidth is ***100Hz*** what will be the bandwidth of ***14th*** filter. Where [6]

$$Pitch(mels) = 3322\log_{10}(1 + \frac{f}{1000})$$

(c) Write name two speech features which are formant like features but not exactly formant features. [4]

4. (a) Draw the functional block diagram of text to speech conversion system and explain the function of the grapheme to phoneme conversion block. [6]

(b) How does spoken language affect automatic speech recognition? Give an example of two homophones in English. [4]

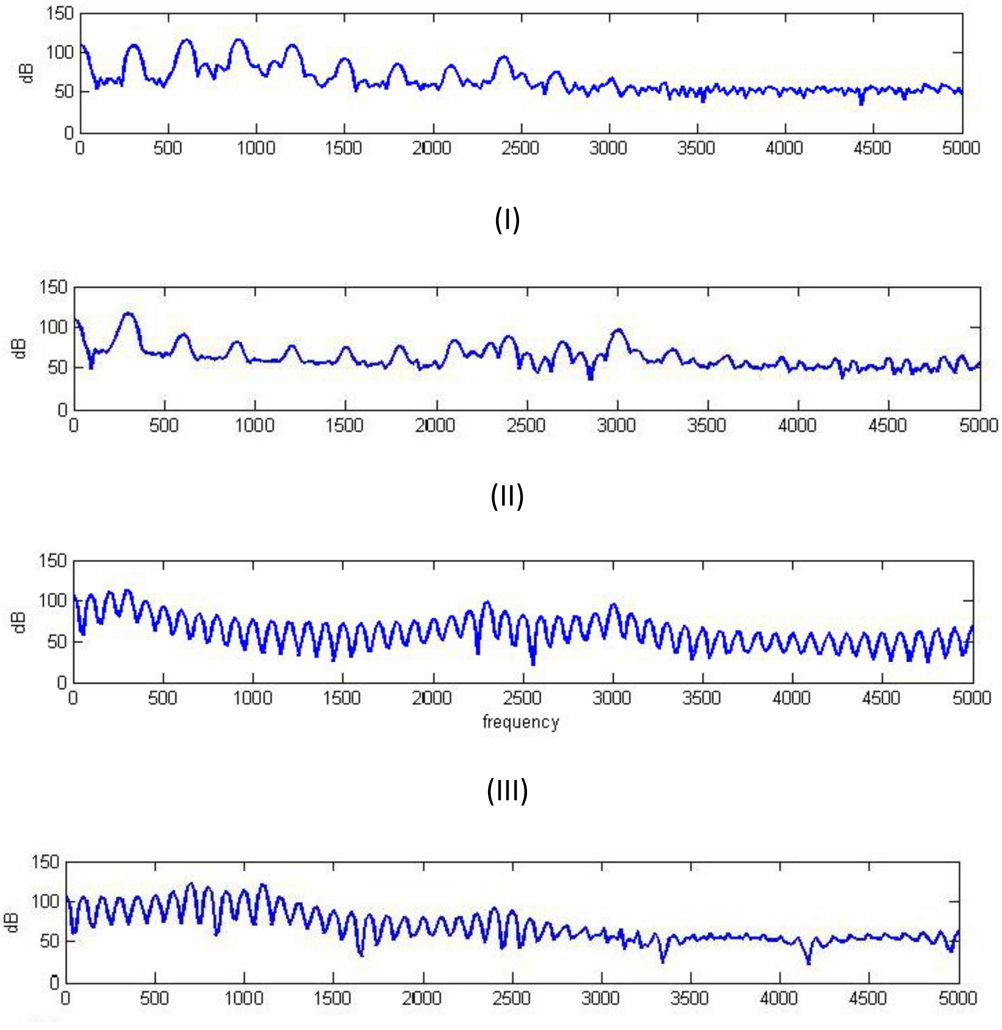
(c) Determine the number of consonant to vowel transition present in the phonetic representation of the first word of your name [6]

5. (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 [4]

(c) Figure 3 shows plots of 4 vowels segment short-time log magnitude spectra as obtained using a Henning window of an appropriate length. Determine the pitch frequency of each of the vowel segment. Which of the vowel segment most likely come from an adult male and why? [8]



(I)

(II)

(III)

(IV)

Figure-3