**Indian Institute of Technology Guwahati** [*Name:* ] 17

**Mid Semester Examination CS-566 Speech Processing** [*Roll No:* ]

***Full Marks:* 20** ***Time: 60 minutes*** ***19th September 2023*** ***14:00 to 15:00 hours*** **PART A**

**No Queries** **Objective questions** carry ***negative marking*** for incorrect/incomplete answers

**Calculators not permitted**

19/09

1. Linguistically distinct speech sounds in a language are called "Phonemes". (True / False)? 1

① TRUE ✓

2. Give an example of a semivowel. 1

② Examples of semi vowels are: /w/, /r/, /l/, /y/ ✓

3. The definition of auto correlation of lag k is given by... 1

③ Auto correlation of lag k is given by

$$R(k,x) = \frac{1}{N}\sum_{i=0}^{N-1} x_i\, x_{i+k}$$ where ...

4. What is the difference between vowels and diphthongs? Give examples and explain in 1 line. 1

④ Vowels are those distinct speech sounds where there is constant uninterrupted flow of air. eg: /a/, /e/, /i/, /o/, /u/ ✓

Diphthongs are those sounds which seem to start with one vowel and then glide to another vowel, eg /ay/, /uy/, /oy/. ✓ ~~1~~

5. Why does the spectrogram of nasal sounds have gaps?

⑤ The spectrogram of nasal sounds contain gaps because nasal sounds contain 2 components, from the vocal tract, as well as the nose, the component coming from nasal tract does not contribute to the spectrogram ✗ and hence we see a gap in the spectrogram.

6. What information from a speech signal do we get from ZCR (Zero Crossing Rate)? 1

⑥ The zcr of speech signal tells us how many times the signal has crossed the x-axis. ZCR is useful to identify hissing sounds using comparison. ✓

7. A speech sample is sampled at a rate of 40,000 samples per second. A 20-msec window is used for short term cepstral analysis, and the window is moved by 15 msec in consecutive analysis frame. 2 + 3

a) How many speech samples are used in each segment?

7a) No of speech samples in 1 segment = $40000 * 20 \times 10^{-3}$

$= 40 \times 20 = \underline{800 \text{ samples}}$. ✓

b) What is the frame rate of the short-time LP analysis?

7b) Frame rate = no of frames in 1 sec $= \left\lfloor \frac{1s - 20msec}{15msec} \right\rfloor + 1$

$= \left\lfloor \frac{1000 \times 10^{-3} - 20 \times 10^{-3}}{15 \times 10^{-3}} \right\rfloor + 1$

$= \left\lfloor \frac{980 \times 10^{-3}}{15 \times 10^{-3}} \right\rfloor + 1$ P.T.O

$= \left\lfloor \frac{980}{15} \right\rfloor + 1 \quad = 65 + 1 \quad = \underline{66 \text{ frames/sec}}$.

$$\begin{array}{r|l} 15 & 980 \;|\; 65 \\ & 90 \\ \hline & 80 \\ & -75 \\ \hline & 5 \end{array}$$

The frame rate is 66 frames/sec. ✓

8. Let the discrete time signal be x(n) and its Fourier Transform be given. In order to get back the original signal from its Fourier Transform representation, what will be the lower limit and upper limits of the integration? 2

(8) The lower limit and upper limit is $-\pi$ and $\pi$ respectively.

$$\tilde{x}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) \cdot e^{j\omega n} \cdot d\omega$$

9. How do you estimate word boundaries in a speech word utterance? → Rest of answer on shot 2

(9) Once we have the speech sample, we can calculate the energy of each frame given by $\frac{1}{N}\sum_{i=1}^{N} x_i^2$, N is frame size. We set certain threshold for this energy. If the energy for a frame is greater than this threshold, we say that the word utterance has started. Initially the energy levels are low ~~hence we~~ for silence, once the utterance starts, the energy

10. Discuss the covariance method to derive values of LP coefficients. Show time complexity of this method. ~~4+1~~

(10) Covariance method for LP coefficients

This is another method to compute the LP coefficients. and the complexity of this method is $O(p^3)$.

Let $\phi_n(i,k) = \sum_{m=0}^{N-1-0} S_n(m-i)\, S_n(m-k)$. ~~$S_n$ → signal~~

we can shift the window $-i$ or $-k$ to the left to obtain,

$$\phi_n(i,k) = \sum_{m=-i}^{N-1-i} S_n(m)\, S_n(m-k+i)$$

or

$$\phi_n(i,k) = \sum_{m=-k}^{N-1-k} S_n(m)\, S_n(m-i+k)$$

here, the ~~letters~~ limits of the summation are not same anymore, ~~as~~ we require samples from $-p$ to $N-1$ for the derivation. Therfore here, no tapering window is used.

hence, this can be written as.

$$\sum_{k=1}^{p} \phi_n(i,k)\, a_k = \phi_n(i,0) \qquad i = \dots$$

Submit and collect Part B

in matrix form, we get.

$$\begin{bmatrix} \phi(1,1) & \phi(1,2) & \phi(1,3) & \cdots & \phi(1,p-1) \\ \phi(2,1) & \phi(2,2) & \phi(2,3) & \cdots & \phi(2,p-1) \\ \phi(3,1) & \phi(3,2) & \phi(3,3) & \cdots & \phi(3,p-1) \\ \vdots & \vdots & \vdots & & \vdots \\ \phi(p-1,1) & \phi(p-1,2) & \phi(p-1,3) & \cdots & \phi(p-1,p-1) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} \phi(1,0) \\ \phi(2,0) \\ \phi(3,0) \\ \vdots \\ \phi(p-1,0) \end{bmatrix} \quad -(1)$$

This matrix is a symmetric matrix as $\phi(i,j) = \phi(j,i)$

The function $\phi_n(i,k)$ is also an even function.

The matrix equation obtained in (1) can be solved by Cholesky's Decomposition method

## Cholesky's Decomposition method

(3)

Given matrix equation

$$\phi_{pxp} \cdot \alpha_{px1} = \psi_{px1}, \quad \text{—— (2)}$$

we can write $\phi$ as $VDV^{-1}$

V → Lower triangular matrix

D → Digonal matrix.

eq(2) becomes.

$$V_{pxp} D V^{-1} \cdot \alpha = \psi_{px1}$$

Let $DV^{-1}\alpha = Y$ — Y will be px1 matrix.

$$\Rightarrow V_{pxp} \cdot Y_{px1} = \psi_{px1}$$

This matrix equation can be solved using LU Decomposition and the $\alpha$ can be produced.

The Storage complexities are

| | |
|---|---|
| Frame : | N+P |
| Matrix : | $\alpha p^2/2$ |
| Windowing : | 0 |

No of multiplications

| | |
|---|---|
| Windowing : | 0 |
| Matrix : | $\alpha p^2$ |
| Solution : | $\alpha p^3$ |

# Indian Institute of Technology Guwahati

(Supplementary Answer Sheet )

| Name of Student : | [redacted] | Roll N[redacted] |
|---|---|---|
| Course No. | CS-566 | Signature of the stud[redacted] |

Part A

9) → Increases sharply and crosses threshold, we can mark this point as the start marker. Similarly once the word utterance is completed, there is a sharp decrease in energy and it goes below threshold value. we can mark this as end marker for the word. Hence we can obtain the word boundaries for the given word utterance. ✓

**Indian Institute of Technology Guwahati** [*Name:* _____] 15

**Mid Semester Examination CS-566 Speech Processing** [*Roll No.* _____]

***Full Marks:* 20** ***Time: 60 minutes*** ***19th September 2023*** ***15:00 to 16:00 hours*** **PART B**

**No Queries** **Objective questions** carry ***negative marking*** for incorrect/incomplete answers

**Calculators not permitted**

1. Why is Hamming window required in LP speech analysis using Auto-correlation method? (1)

(1) ~~Hamm~~ Autocorrelation method assumes that the signal is identically 0 outside the interval under consideration. to ensure this, the signal is multiplied with hamming window which makes the signal 0 outside the interval [0 to N−1] (1)

2. Why is "liftering" done on cepstral coefficients?

(2) We know that lower order cepstral coefficients are sensitive to spectral slope and higher order cepstral coefficients are sensitive to noise. to minimise these sensitivities, LIFTERING is performed on the cepstral coefficients. (1)

3. State the properties of a Toeplitz matrix.

(3) Properties of Toeplitz matrix are :-

(1) It is a symmetric matrix

(2) All the elements across any diagonal are equal. (1)

4. Discuss characteristics of epsilon in LBG algorithm.

(4) $\epsilon$ is the splitting parameter used in LBG algorithm. (1)

its values ranges from 0.01 to 0.05 $[0.01 \le \epsilon \le 0.05]$

the new ~~centroids~~ $\tilde{y}_n$ after splitting are $\tilde{y}_n^+ = \tilde{y}_n(1+\epsilon)$

ideal value we consider is 0.03. $y_n^- = \tilde{y}_n(1-\epsilon)$

5. Let us digress to an image processing problem using VQ. Let the size of an image be 512 x 512. If you decide to use a codebook size of 8 for vector quantization, how many bits per pixel are needed to represent the quantized image? (2)

(5) ~~No of bits per pixel~~ Size of codebook = 8

$\Rightarrow$ no of bits required to represent codebook index = $\log_2 8 = 3$

Image size = 512 × 512

$\Rightarrow$ total pixels = $2^9 \cdot 2^9 = 2^{18}$ → rest in extra sheet

→ Each of these pixels will be mapped to 1 vector from codebook.

∴ No of bits to represent 1 pixel = Size of codebook index = 3 bits per pixel

6. How do k-means and LBG algorithms differ? Which of the two is more favored and ~~for what~~ reason? Illustrate with diagram(s). (2+1+1)

(6) K means Algorithm

In the K means algorithm, the initial codebook size is fixed. and ~~the first iteration~~ the initial codebook is generated by an adequate method.

On each iteration Each vector from universe is classified into one of the K clusters. and centroid is computed for each cluster upon completion of classification

Rest on extra sheet →

7. Suppose you have a dataset of 8 data points in 2D space:
Data Points: (3, 2), (9, 7), (2, 3), (10, 8), (3, 3), (5, 5), (6, 4), (7, 5).

You want to apply the k-means algorithm to cluster these data points into two clusters (K=2) using the Euclidean distance metric. Start with initial cluster centroids at (3, 3) and (7, 5). Determine the final cluster centroids and the data points assigned to each cluster.

~~10~~ 5

(7) iteration $m=0$.

$C_1 = \{(3,3)\}$ $C_2 = \{(7,5)\}$. , $Cl_1 = [(3,3)]$ , $Cl_2 = [(7,5)]$

iteration $m=1$:

$x = (3,2) \rightarrow$ $D(x,C_1) = \sqrt{(3-3)^2 + (3-2)^2}$

$= \sqrt{0+1} = 1$

$D(x,C_2) = \sqrt{(3-7)^2 + (3-5)^2} = \sqrt{4^2 + (2)^2} = \sqrt{18}$

$D(x,C_1) < D(x,C_2)$

$\therefore (3,2)$ is put to $Cl_1$.

$x: (9,7) \rightarrow$ $D(x,C_1) = \sqrt{(3-9)^2 + (3-7)^2} = \sqrt{6^2 + 4^2} = \sqrt{52}$

$D(x,C_2) = \sqrt{(7-9)^2 + (5-7)^2} = \sqrt{(-2)^2 + (-2)^2} = \sqrt{8}$

$D(x,C_2) < D(x,C_1)$

$\therefore (9,7)$ is put to $Cl_2$.

$x: (2,3) \rightarrow$ $D(x,C_1) = \sqrt{(2-3)^2 + (3-3)^2} = \sqrt{1+0} = \sqrt{1}$

$D(x,C_2) = \sqrt{(2-7)^2 + (3-5)^2} = \sqrt{(-5)^2 + (-2)^2} = \sqrt{25+4} = \sqrt{29}$

$D(x,C_1) < D(x,C_2)$

$\therefore (2,3)$ is put to $Cl_1$

$x: (10,8) \rightarrow$ $D(x,C_1) = \sqrt{(10-3)^2 + (8-3)^2} = \sqrt{7^2 + 5^2} = \sqrt{49+25} = \sqrt{74}$

$D(x,C_2) = \sqrt{(10-7)^2 + (8-5)^2} = \sqrt{3^2 + 3^2} = \sqrt{18}$

$D(x,C_2) < D(x,C_1)$

$\therefore (10,8)$ is put to $Cl_2$

→ Rest done on extra sheet

**Indian Institute of Technology Guwahati**

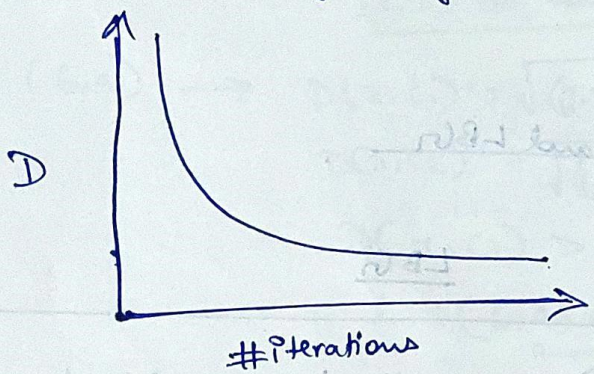
(Supplementary Answer Sheet)

| Name of Student : | | Roll N | |
|---|---|---|---|
| Course No. | CS-566 | Signature of the student : | |

Part B

6) the distortion D(n) is calculated and if D(m-1) − D(m) is less than threshold, the algorithm stops, otherwise the classification & codevector updation process is repeated.

Here is the graph of Distortion V/s #iterations.




The Distortion decreases exponentially until a local minimum is reached.

But Kmeans algorithm only allows us to reach a local minimum and not a global minimum.

LBG Algorithm.

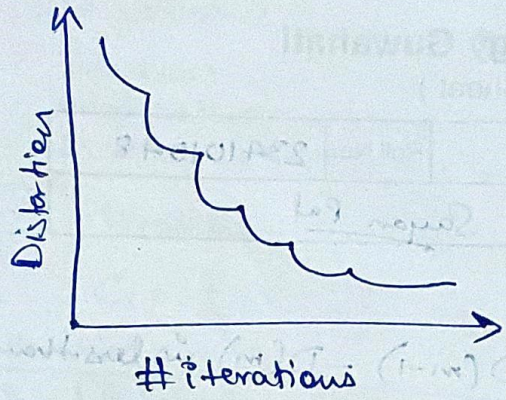
Here, we start with a 1vector codebook. [Centroid of the universe]

We split the codebook in each iteration with the rule

$$\tilde{y}_n^+ = \tilde{y}_n(1+\epsilon)$$
$$\tilde{y}_n^- = \tilde{y}_n(1-\epsilon) \qquad 0.01 \le \epsilon \le 0.05$$

Now Kmeans is applied to the split codebooks (Double the size) to obtain new centroids and distortion is calculated as well, If it is less than threshold, then we check if our codebook size desired is reached, and if we get the desired codebook, we stop the algorithm.

P.T.O




The Distortion v/s # of iterations graph is shown here, for LBG algorithm. The LBG algorithm helps us to cross the local minimum and get even lower values for difference in Distortion.

Although LBG algorithm is time consuming and intensive, it has proved to provide better result than regular K-means algorithm and hence it is preffered.

## Algorithmic steps for Kmeans and LBG

| Kmeans | LBG |
|---|---|
| 1) **Initialization** initialize random codebook of size K | 1) Start with 1 vector codebook |
| 2) **Classification** classify universe into K clusters | 2) Split codebook with $\tilde{y}_n$ into $\tilde{y}_n^+ = \tilde{y}_n(1+\epsilon)$, $\tilde{y}_n^- = \tilde{y}_n(1-\epsilon)$, $0.01 \le \epsilon \le 0.05$ |
| 3) **Code vector updation** update the code vectors for each cluster. | 3) Apply K means on the split codebook to get distortion |
| 4) **Termination.** Repeat steps 2 & 3 until the distortion is below the threshold. | 4) Continue steps 2 & 3 till desired codebook size is reached. |

LBG Algorithm gives us the flexibility of having dynamic sized codebook it helps in better classification and optimizes the centroids. Also since K means uses a random initial codebook, it induces randomness into the codebook but LBG uses the centroid of universe, ∴ it is more accurate.

7) $x : (3,3) \longrightarrow D(x,c_1) = \sqrt{(3-3)^2+(3,3)^2} = 0$

$D(x,c_2) = \sqrt{(3-7)^2+(3-5)^2} = \sqrt{4^2+(-2)^2} = \sqrt{16+4} = \sqrt{20}$

$(3,3)$ ~~this~~ is equal to centroid itself.

$\therefore$ it is part of $c_1$

$x : (5,5) \longrightarrow D(x,c_1) = \sqrt{(5-3)^2+(5-3)^2} = \sqrt{2^2+2^2} = \sqrt{8}$

$D(x,c_2) = \sqrt{(5-7)^2+(5-5)^2} = \sqrt{2^2+0} = \sqrt{4}$

$D(x,c_1) > D(x,c_2)$

$\therefore$ $(5,5)$ is added to $c_2$

$x : (6,4) \longrightarrow D(x,c_1) = \sqrt{(6-3)^2+(4-3)^2} = \sqrt{2^2+1^2} = \sqrt{5}$

$D(x,c_2) = \sqrt{(6-7)^2+(4-5)^2} = \sqrt{1^2+1^2} = \sqrt{2}$

$D(x,c_1) > D(x,c_2)$

$\therefore$ $(6,4)$ is added to $c_2$

$x : (7,5) \longrightarrow D(x,c_1) = \sqrt{(7-3)^2+(5-3)^2} = \sqrt{4^2+2^2} = \sqrt{20}$

$D(x,c_2) = \sqrt{(7-7)^2+(5-5)^2} = \sqrt{0} = 0$

$(7,5)$ is a centroid itself.

$\therefore$ it is part of $c_2$. ✓

$\therefore$ the final clusters are ?

$c_1 = [(3,2), (2,3), (3,3)]$

$c_2 = [(9,7), (10,8), (5,5), (7,5)]$

new centroid for $c_1 = \left(\frac{3+2+3}{3}, \frac{2+3+3}{3}\right) = (8/3, 8/3)$

new centroid for $c_2 = \left(\frac{9+10+5+7}{4}, \frac{7+8+5+5}{4}\right) = (31/4, 25/4)$

∴ the final clusters are:

$C1_1 = [(3,2), (2,3), (3,3)]$ with centroid $(8/3, 8/3)$

and

$C1_2 = [(5,5), (7,5), (9,7), (10,8)]$ with centroid $(31/4, 25/4)$

5) No of bits to represent image = $512 \times 512 \times 3$

$= 3 \times 2^{18}$

No of bits per pixel needed = 3 bits per pixel ✓