Indian Institute of Technology Kharagpur
School of Information Technology
IT60116: Advanced Topics in Speech Processing

Date: 1st Feb 2010 Quiz-1 Marks: 20 Time: 50 min

1. Why do you feel speech processing is difficult using machine? Mention some important issues to support your argument. (2 Marks)

2. Assume the speech production system is linear time invariant system. Let the excitation signal is e(n), impulse response of the vocal tract system is v(n) and the speech signal is s(n). Relate the above three signals in time domain and frequency domain. (2 Marks)

3. Mention different places of articulation, and give two sound units corresponds to each type of place of articulation. (2 Marks)

4. Specify the spectrogram details for the following sound units: (i) vowels, (ii) fricatives, (iii) nasals and (iv) stop consonants. (2 Marks)

5. Why speech signal to be processed at sub-segmental, segmental and supra-segmental levels? (3 Marks)

6. Draw the approximate time domain waveform and spectrogram for the speech utterance "mask". Indicate different regions in both time domain waveform and spectrogram. (4 Marks)

7. Derive the formula for determining the convolution using the fundamental properties of linear time invariant system, unit sample function and unit sample response of a system. (3 Marks).

8. Determine the convolution between two sequences x1(n) and x2(n). (2 Marks)
   Where x1(n) = [1,0,1,0], for n = 1 to 4, and 0 for all other values of n.
   x2(n) = [0,1,0,1], for n = 1 to 4, and 0 for all other values of n.

Indian Institute of Technology Kharagpur
Department of Computer Science and Engineering
IT 60116: Advanced Topics in Speech Processing

Date: 18-02-2019 Mid-sem examination Marks: 60 Time: 2hrs

Note:

1. Intermediate results and steps need to be presented.
2. All parts of the question (a,b,c,d) should be answered at a stretch.

---

1. (a) Highlight the salient points in speech production and perception mechanisms.
   (b) Draw an intuitive speech signal (in time domain) for the word "mask". Mark the boundaries of each sound unit present in the signal. Correlate the nature of each sound unit with the characteristics of the specific speech segment. (**4+6 = 10 marks**)

2. (a) What is spectrogram? How it is plotted?
   (b) Name two different types of spectrograms. Briefly discuss their salient features.
   (c) What are formants? Mention the discriminative characteristics of formants in view of discriminating the vowel categories. (**3+3+4 = 10 marks**)

3. (a) Name the various types of excitations present in human speech production.
   (b) Highlight the salient features of each type of excitation.
   (c) Name two sound unit examples to each type of excitation. (**1+6+3 = 10 marks**)

4. (a) What is a vowel triangle in view of speech analysis? What is the special about the corners of the vowel triangle? Specify the articulatory characteristics related to the corners of the vowel triangle.
   (b) Physiologically (production perspective) how nasal sounds (the last column of consonant varnamala) differ among themselves within a group? How they can be discriminated using signal processing techniques?
   (c) Name the different knowledge sources present in speech. (**5+3+2 = 10 marks**)

5. (a) Define an unit sample. Represent the digital signal using unit sample sequence.
   (b) From the basic properties of LTI system, derive the formula to compute the output of LTI system in terms of its input and unit sample response of a system.
   (c) Determine the convolution and correlation sequences using two discrete signals x(n) and y(n). Consider x(n) = {1, 2, 1} for n = {-1, 0, 1}, y(n) = {1, 0, 1} for n = {-1, 0, 1}. (**2+4+4 = 10 marks**)

6. Derive the DFT relations graphically from their respective continuous time Fourier transform relations. Plot the figures carefully, and discuss the salient points at each stage of transformation. Highlight the artifacts introduced in both time and frequency domains and suggest the appropriate solutions to minimize the affect of the artifact. (**10M**)

Indian Institute of Technology Kharagpur
Department of Computer Science and Engineering
IT 60116: Advanced Topics in Speech Processing

Date: 22-04-2019 End-sem examination Marks: 100 Time: 3hrs

Note: ANSWER ANY FIVE QUESTIONS. DO NOT ATTEMPT EXTRA QUESTIONS.

1. Intermediate results and steps need to be presented.
2. All parts of the question (a,b,c,d) should be answered at a stretch.

---

1. (a) Let x(n) and y(n) are the input and output of the system. Find the transfer function of the system using linear constant coefficient difference equations.

   (b) How poles and zeros are estimated from the transfer function?

   (c) What is meant by all-pole and all-zero systems?

   (d) What is a minimum phase system? Comment on poles and zeros of a minimum phase system.

   (e) Draw the frequency and impulse (time domain) responses of a system having a pair of complex conjugate poles.

   (f) With the help of appropriate formulae (relations), discuss how complex conjugate poles affect the formant frequencies and their bandwidths in the frequency response. Plot the frequency and impulse responses of the systems whose complex conjugate poles have the same angular frequency, but their radial distances (r_1, r_2, r_3, r_4) to the origin vary such that $r_1 < r_2 < r_3 < r_4; r_3 = 1, (r_1, r_2) < 1, r_4 > 1$.

   (g) Draw the frequency and impulse (time domain) responses of systems having real poles a_1, a_2, a_3, a_4 such that $a_1 < a_2 < a_3 < a_4, a_3 = 1, (a_1, a_2) < 1$, and $a_4 > 1$.

   [2+1+1+2+2+8+4 = 20 marks]

2. (a) Why short-term energy is preferred over overall energy in speech processing? How short-term energy feature is derived for a speech signal? What are the various speech tasks that can be performed using the above feature?

   (b) What information is conveyed through zero crossing rate (ZCR) of a signal? For a given speech signal, how short-term ZCR feature is estimated? What are the various speech tasks that can be performed using the above feature?

   (c) How short-term autocorrelation sequence is estimated? Mention various applications of short-term autocorrelation sequence in the context of speech. Explain, how autocorrelation feature is used in any one of the above mentioned speech applications.

   (d) What is AMDF? How it is computed and where it is used? Show the waveforms of voiced speech segment and its AMDF signal. In what aspects AMDF is preferred over short-term autocorrelation?

   [5+5+5+5 = 20 marks]

3. (a) With help of LP analysis model (with appropriate figure), explain how speech signal is parameterized? What will be the filter transfer function and its frequency response in case of LP analysis model?

   (b) Draw a block diagram representing the LP synthesis model. What will be the filter transfer function and its frequency response in case of LP synthesis model?

   (c) Derive the linear prediction (LP) filter coefficients using autocorrelation formulation by minimizing the mean square error.

   (d) Illustrate the LP analysis of a voiced and unvoiced speech segments by plotting the following figures: (a) speech signal, (b) error signal, (c) true spectrum (using DFT), (d) LP spectrum and (e) spectrum of the error signal.

   [3+3+6+8 = 20 marks]

4. (a) Discuss the significance of cepstrum analysis in speech processing.

   (b) List the steps for deriving the linear prediction cepstral coefficients (LPCCs) from speech signal.

   (c) Discuss, why the following post processing steps are incorporated on cepstral parameters: (i) Weighting and (ii) Temporal derivative. Also, explain how you incorporate these two post-processing operations?

   (d) Briefly, discuss about the need of perceptually-relevant distortion measure in the context of speech processing. Mention the speech features which are perceptually sensitive and insensitive. Suggest the perceptually relevant distance measure for speech processing applications.

   [2+6+6+6 = 20 marks]

5. (a) What is the difference between vector and scalar quantizers?

   (b) What is codebook in the context of vector quantization in the domain of speech signals.

   (c) Clearly discuss the steps involved k-means clustering method for deriving the code book in VQ.

   (d) Determine the data rates for the transmission of speech using the following methods:

      i. Sampling at 10 KHz and using a quantizer with 1024 levels

      ii. Sampling at 10 KHz, frame size of 20 ms, frame shift of 10 ms, 15 LPCCs per frame and each coefficient is represented by 10 bit accuracy.

      iii. Use the above spectral features and prepare the codebook of size 1024 using vector quantization.

   (e) Briefly explain how vector quantization (VQ) is used for speech compression?

   (f) Discuss how VQs are used for developing isolated word (limited vocabulary) speech recognition systems?

   [2+2+5+5+3+3 = 20 marks]

6. (a) In view of statistical pattern recognition, briefly explain the following:
   (i) Classification vs Regression
   (ii) Curse of dimensionality
   (iii) Bias and variance w.r.t polynomial curve fitting
   (iv) Bayee's classification rule

(b) Explain the principle of K-nearest-neighbours (KNN) based approach for density estimation. Explain with appropriate expressions, how KNN will be used for classification task?

(c) In view of Mixture models, give the expressions for the following: (i) $p(\mathbf{x}|j)$, (ii) Negative log-likelihood E, (iii) μ_j^{new}, (iv) $(\sigma_j^{new})^2$ and (v) $P(j)^{new}$. Briefly explain each expression in 1-2 lines.

[10+5+5 = 20 marks]

7. (a) Define the basic elements of HMM (N, M, A, B and Π).

(b) With the appropriate expressions, discuss the computation of $P(O/\lambda)$ using (i) direct method and (ii) forward variable $\alpha_t(i)$.

(c) With appropriate steps, explain the Viterbi algorithm to compute the optimal state sequence for the given observation sequence.

(d) How the parameters A, B and Π are estimated for deriving the optimal λ where $P(O/\lambda)$ is maximized.

[2+8+6+4 = 20 marks]

8. (a) Consider a scenario where there are 50 speakers who speak the phrase RESTRICTED AREA. The pronunciations of the words are as follows: RESTRICTED: /R/ /IY/ /S/ /T/ /R/ /IH/ /K/ /T/ /AH/ /D/; AREA: /AE/ /R/ /IY/ /AH/. Suppose you are required to build a p-state phone HMM with k-component GMM per state.

i. How many HMMs are required to build a phrase recognizer? Justify your answer.

ii. How many HMM states are involved in the process? Justify your answer.

iii. What is the dimension of the state transition matrix?

iv. What is the dimension of GMM prior probability for each state of HMM?

v. How many forward probabilities are computed per input speech frame? Justify your answer.

(b) Suppose there are three words W1, W2 and W3 belonging to the languages L1, L2 and L3 respectively. A five dimensional feature vector is extracted for each word. Draw the DNN architecture having 3 hidden-layers highlighting the number of nodes in each layer.

(c) Draw the LSTM unit highlighting different gated connections. Design a single LSTM network to classify the following cities to the respective states given below:

City	**State**
Kharagpur	West Bengal
Darjeeling	West Bengal
Bangalore	Karnataka
Mysore	Karnataka
Konark	Odissa
Puri	Odissa
Cuttack	Odissa

[10+5+5 = 20 marks]