CARE, IIT Delhi
CRL 707 Human and Machine Speech Communication
Major Exam 2006-07/I

- **Time allowed: Two hours.**
- **Maximum marks: 90; Q1: 20, Q2: 12, Q3: 20, Q4: 12; Q5: 10; Q6: 16.**
- **Give precise technical or mathematical explanations wherever needed.**

1. (a) Two pure tones at 350 Hz and 3500 Hz respectively are played at 50 dB SLP for several seconds, one after the other with a gap of few seconds in between. Explain clearly what will be the relative loudness of the two tones.
   (b) A pure tone at 3500 Hz and 50 dB SPL is played for duration of 10 ms, 100 ms, 1 s and 10 s respectively, one after the other with gaps of few seconds in between. Will the loudness be the same in each case or will it vary? Explain clearly.
   (c) Explain precisely two differences in the decoded speech signal when perceptual weighting filter is used in a CELP encoder.
   (d) A signal consists of a sum of two sinusoids with frequencies of 400 Hz and 2000 Hz. Assuming high SNR and a sampling frequency of 8 KHz, what should be the order of a linear predictor that will predict the signal accurately?

2. It is required to design a speech recognition system that is robust to channel errors and noise. To begin with, a reasonable design approach is to mimic the processing of the human ear, and attempt to attain its performance efficiency. This would require us to process the received electrical signal in a similar manner as the corresponding acoustic signal is processed by the ear, and to design features from the processed signal that are analogous to the internal representation of the signal in the ear. From your knowledge of the important auditory phenomena and the processing in the ear, give the mathematical steps of an algorithm to determine features from the received signal s[n] that can be used for the purpose. Explain the steps clearly.

3. Consider an isolated word speech recognizer with a vocabulary of two words. Let word 1 be modeled by a HMM λ_1, and word 2 be modeled by a HMM λ_2. The state transition values are given as follows: HMM λ_1: a11 = 0.5, a12 = 0.5, a21 = 0.4, a22 = 0.6; HMM λ_2: a11 = 0.4, a12 = 0.6, a21 = 0.2, a22 = 0.8. The observation (feature) sequence is a scalar variable x, modeled as a continuous density GMM. For λ_1, $p_1(x)$ = 0.5 N(0, 10) + 0.5 N(4, 4), and for λ_2, $p_2(x)$ = 0.7 N(4, 10) + 0.3 N(-4, 4), for both the states in each HMM. Here, N(m, s) denotes a normal pdf with mean m and variance s. It is assumed that the initial state Π is 1 for both the HMM models.
   (a) Given an observation sequence O = {o1, o2} = {2, 2} of a test word from the 2 word vocabulary, find the probability of the observation sequence given the two word models, i.e. $P(O/\lambda_1)$ and $P(O/\lambda_2)$, using the "forward" algorithm. Which word is more likely to have produced the given observation sequence?
   (b) For the same observation sequence as in part (a), use the Viterbi algorithm to find the single best state sequence for the HMM, λ_2. Show the steps clearly.

4. Show mathematically that the scalar gain factor of the fixed (or adaptive) codebook in CELP encoding is required to be computed only once for a subframe, for the best codebook index, and is not required to be computed for every index, as part of the best codebook index search.

5. Suppose that an adult male has generated a sustained vowel for duration of 1 second. Assume F0 = 125 Hz, F1 = 400 Hz, F2 = 1800 Hz, and F3 = 2600 Hz and that an 8 KHz sampled version of the signal is available. Consider a 25 ms Hamming windowed segment extracted from the middle of this signal. A complex cepstrum analysis is performed. Next two "frequency invariant linear filtering" or "liftering" operations are performed on the complex cepstrum c[n] according to: $y_i[n] = c[n]l_i[n]$, $i = 1,2$, where, $l_1[n] = 1$, $|n| < K$, and 0 otherwise, and $l_2[n] = 1$, $|n| \geq K$, and 0 otherwise.
   (a) Suggest a suitable value of K such that $y_1[n]$ retains the formant frequency information and $y_2[n]$ retains the fundamental frequency information of the speech signal. Explain your choice.
   (b) In two separate graphs, plot the magnitude spectrum of the two "liftered" signals $y_1[n]$ and $y_2[n]$. Explain the nature of the plots.

6. The accompanying figure gives the spectrogram of a sentence utterance.
   (a) Mark the phone boundaries accurately between 0 and 900 ms.
   (b) Identify the most likely phone class for each case from the following list: stop consonant, vowel, diphthong, unvoiced fricative, voiced fricative, whisper. Give reasons.
   (c) Give all articulatory information (such as position of tongue, vocal tract shape, excitation signal characteristics) for the utterances at (i) 150 ms, (ii) 250 ms, and (iii) 500 ms.