Q1 (a) Fig. Q1(a) shows a windowed speech frame $s(n)$ and its prediction error $e(n)$. The prediction error is obtained by applying P-th order prediction analysis on the speech frame.

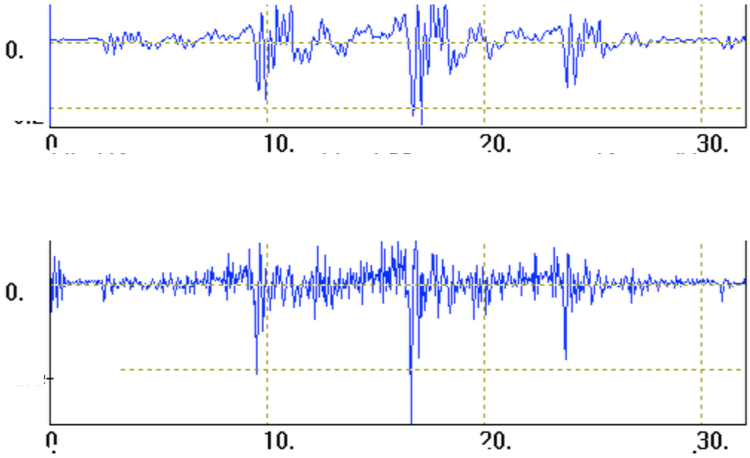

Fig. Q1(a)

(i) Which of the panels (upper or lower) in Fig. Q1(a) shows the prediction error? Briefly explain your answer.

(4 marks)

(ii) State whether the speech frame is voiced or unvoiced. Give reasons to support your answer.

(3 marks)

(iii) Given the unit on the horizontal axis in Fig. Q1(a) is millisecond (ms), estimate the pitch period (if any) of the speech frame.

(3 marks)

(iv) Express the prediction error $e(n)$ in terms of $s(n)$ and the linear prediction coefficients $\{a_1, \ldots, a_P\}$ of this frame. Hence, explain how you would obtain the speech signal $s(n)$ from $e(n)$ and $\{a_1, \ldots, a_P\}$.

(6 marks)

(b) Fig. Q1(b) shows a wide-band spectrogram and a narrow-band spectrogram of an utterance.

(i) Determine the sampling frequency of the speech signal.

(2 marks)

(ii) Which of the panels (upper or lower) corresponds to the narrow-band spectrogram? Briefly justify your answer.

(4 marks)

(iii) Estimate the pitch period at the instance indicated by the vertical dashed line in Fig Q1(b). Briefly explain your answer.

(3 marks)

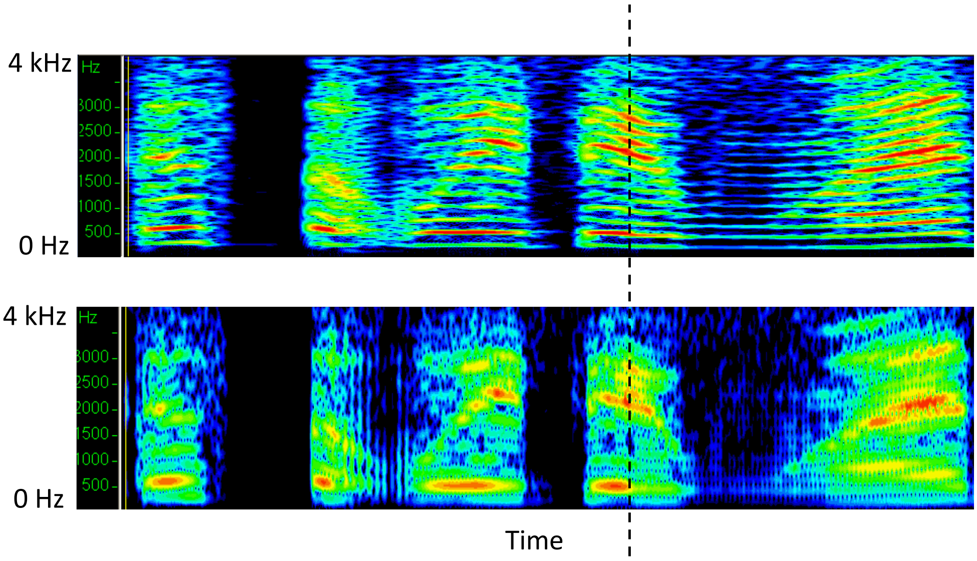

Fig. Q1(b)

Q2 (a) In the frequency domain, a speech spectrum $S(\omega)$ can be obtained by multiplying an excitation spectrum $E(\omega)$ and the frequency response of a vocal tract filter whose transfer function is $H(z)$, i.e.,

$$S(\omega) = E(\omega)H(\omega).$$

(i) Show that in the cepstral domain, the multiplication becomes an addition, i.e.,

$$c_s(n) = c_e(n) + c_h(n),$$

where $c_s(n)$, $c_e(n)$, and $c_h(n)$ are the cepstra of $s(n)$, $e(n)$, and $h(n)$, respectively.

(4 marks)

(ii) Assuming that $H(z)$ is unknown, explain how you would obtain the spectral envelope of $S(\omega)$ from $c_s(n)$.

(5 marks)

(iii) Suggest a method to estimate the pitch period from $c_s(n)$ if $S(\omega)$ is a voiced spectrum.

(5 marks)

(b) Fig. Q2 shows the A-law and μ-law nonlinear functions used in the ITU G.711 PCM coder. To perform encoding, speech signal $s(n)$ is passed through either of these nonlinear functions, followed by linear quantization.

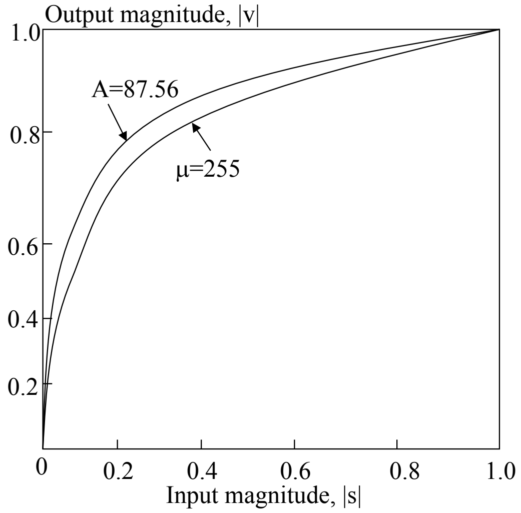

Fig. Q2

(i) Explain why these nonlinear functions can compress the amplitude of high-energy speech signals and expand the amplitude of low-energy speech signals.

(5 marks)

(ii) Explain why it is beneficial to quantize the signals after the nonlinear function instead of quantizing the signals before the nonlinear function.

(6 marks)

Q3 Fig. Q3 shows the structure of a 3-state hidden Markov model (HMM) that models the spectro-temporal characteristics of a phoneme. Each state comprises a Gaussian mixture model with M mixture components. Denote $q_t \in \{1, 2, 3\}$ as the state at frame t. The likelihood of an acoustic vector $\mathbf{o}_t$ condition on state q_t is

$$p(\mathbf{o}_t|\text{state} = q_t) \equiv b_{q_t}(\mathbf{o}_t) = \sum_{k=1}^{M} \omega_k \mathcal{N}(\mathbf{o}_t|\boldsymbol{\mu}_{q_t,k}, \boldsymbol{\Sigma}_{q_t,k}),$$

where $\{\omega_{q_t,k}, \boldsymbol{\mu}_{q_t,k}, \boldsymbol{\Sigma}_{q_t,k}\}_{k=1}^{M}$ are the GMM parameters of state q_t. Denote $\mathcal{O} = (\mathbf{o}_1, \ldots, o_T)$ and $\mathbf{q} = (q_1, \ldots, q_T)$ as the acoustic-vector sequence and the HMM-state sequence corresponding to a phone, where T is the number of frames in the phone. Then, the likelihood of $\mathcal{O}$ given the state sequence $\mathbf{q}$ is

$$p(\mathcal{O}|\mathbf{q}) = \prod_{t=1}^{T} p(\mathbf{o}_t|\text{state} = q_t) = \prod_{t=1}^{T} b_{q_t}(\mathbf{o}_t).$$

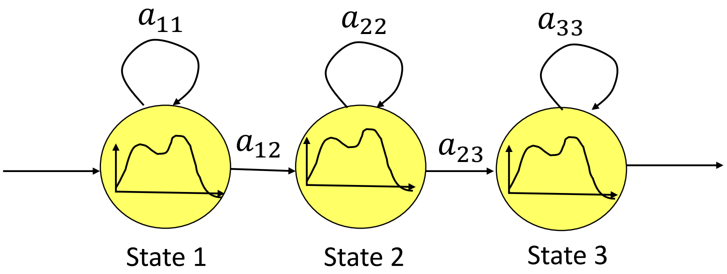

Fig. Q3

(a) In Fig. Q3, if $a_{11} = 0.4$, what is the value of a_{12}? What is the implication if a_{11} is close to zero?

(4 marks)

(b) Briefly explain the purpose of the three states in the HMM.

(3 marks)

(c) Given that the probability of the state sequence $\mathbf{q}$ is

$$P(\mathbf{q}) = a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T},$$

where a_{ij} is the probability of transiting from state i to state j. Express the joint likelihood of $\mathcal{O}$ and $\mathbf{q}$ (i.e., $p(\mathcal{O}, \mathbf{q})$), in terms of a_{ij} and $b_j(\mathbf{o}_t)$ for $i, j \in \{1, 2, 3\}$ and $t = 1, \ldots, T$. *Hint*: Use the product rule $P(A, B) = P(A|B)P(B)$.

(5 marks)

(d) Express the likelihood of $\mathcal{O}$ (i.e., $p(\mathcal{O})$) in terms of a_{ij} and $b_j(\mathbf{o}_t)$ for $i, j \in \{1, 2, 3\}$ and $t = 1, \ldots, T$. *Hint*: Use the sum rule $P(A) = \sum_B P(A, B)$.

(5 marks)

(e) If $\mathcal{O}$ is extracted from the phone for which this HMM is trained to model and $\mathcal{O}'$ is extracted from another phone, compare the values of $p(\mathcal{O})$ and $p(\mathcal{O}')$.

(3 marks)

(f) Fig Q3 is a classical GMM–HMM in which each state is represented by a GMM. Discuss how you would change it to a DNN–HMM in which the likelihood $b_j(\mathbf{o})$, $j = 1, 2, 3$, can be obtained from a DNN.

(5 marks)

Q4 (a) The scoring functions of GMM–UBM and GMM–SVM speaker verification systems are closely related. Denote the acoustic vectors extracted from a test utterance as $\mathcal{O}^{(t)} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$, where T is the number of frames in the utterance. Also denote $\Lambda^{(s)}$ and Λ^{ubm} as the GMM of client-speaker s and the UBM, respectively. The GMM–UBM score and the GMM–SVM score are respectively given by

$$S_{\text{GMM–UBM}}(\mathcal{O}^{(t)}|\Lambda^{(s)}, \Lambda^{\text{ubm}}) = \log p(\mathcal{O}^{(t)}|\Lambda^{(s)}) - \log p(\mathcal{O}^{(t)}|\Lambda^{\text{ubm}})$$

and

$$S_{\text{GMM–SVM}}(\mathcal{O}^{(t)}|\text{SVM}_s) = \alpha_0^{(s)} K\left(\vec{\boldsymbol{\mu}}^{(s)}, \vec{\boldsymbol{\mu}}^{(t)}\right) - \sum_{i \in \mathcal{S}_{\text{bkg}}} \alpha_i^{(s)} K\left(\vec{\boldsymbol{\mu}}^{(i)}, \vec{\boldsymbol{\mu}}^{(t)}\right) + b^{(s)},$$

where SVM_s is the SVM of speaker s, $\mathcal{S}_{\text{bkg}}$ comprises the support vector indexes of the background speakers, $\vec{\boldsymbol{\mu}}^{(s)}$ and $\vec{\boldsymbol{\mu}}^{(t)}$ are the GMM-supervector of speaker s and the test utterance, respectively, $\alpha_j^{(s)}$'s are the Lagrange multipliers, and $b^{(s)}$ is the bias term of the SVM.

(i) Based on the GMM–SVM scoring function, determine the lower-bound on the number of enrollment utterances from speaker s. Briefly justify your answer. You may assume that each enrollment utterance gives one GMM-supervector.

(3 marks)

(ii) In practical GMM–SVM systems, the kernel function $K(\cdot, \cdot)$ is always linear. Explain why it is the case. Also explain why it is inappropriate to use non-linear kernels such as the RBF kernel.

(5 marks)

(iii) Discuss the advantages of GMM–SVM systems over the GMM–UBM systems.

(8 marks)

(b) In i-vector/PLDA speaker verification, the i-vectors are modeled by a factor analysis model:

$$\mathbf{x} = \mathbf{m} + \mathbf{V}\mathbf{z} + \boldsymbol{\epsilon} \qquad \text{(Eq. Q4–1)}$$

where $\mathbf{x}$ is an i-vector, $\mathbf{m}$ is the global mean of all i-vectors, $\mathbf{V}$ represents the speaker subspace, $\mathbf{z}$ is a latent factor, and $\boldsymbol{\epsilon}$ is a residual term that follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \boldsymbol{\Sigma})$.

(i) Explain why it is necessary to model i-vectors by Eq. Q4–1.

(4 marks)

(ii) Assume that the prior of $\mathbf{z}$ follows a standard Gaussian, i.e., $\mathbf{z} \sim \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is an identity matrix. Use Eq. Q4–1 to explain why i-vectors follow a Gaussian distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{V}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma}$, i.e.,

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma})$$

Hints: Take the expectation of $\mathbf{x}$ and $(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}$ in Eq. Q4–1, i.e., $\mathbb{E}\{\mathbf{x}\}$ and $\mathbb{E}\{(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\}$.

(5 marks)

– END –