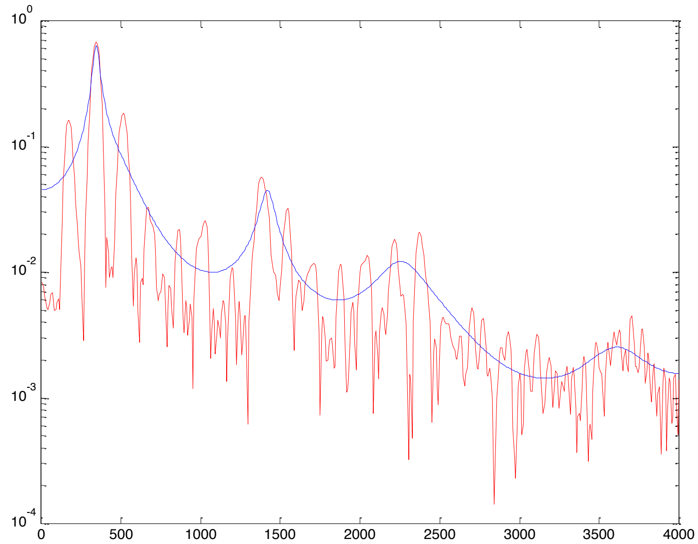
Q1 (a) Fig. Q1 shows the frequency spectrum of a frame of speech signals.



Frequency (Hz)

Fig. Q1

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

(3 marks)

(ii) Estimate the pitch period (if any) of the frame.

(3 marks)

(iii) Explain how the spectral envelope (the smooth curve in Fig. Q1) can be obtained from the linear prediction coefficients $\{a_k; k = 1, \ldots, P\}$ of this frame, where P is the prediction order.

(5 marks)

(b) The vocal tract of human can be considered as a digital filter with impulse responses $\{h(n); n = 0, 1, \ldots\}$. Denote $e(n)$ as the excitation (input) signal applied to this digital filter. Then, the filter's output can be written as

$$s(n) = e(n) * h(n),$$

where '$*$' is the convolution operator.

(i) Show that in the cepstral domain, the convolution operation becomes a summation, 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.

(5 marks)

(ii) Assume that $h(n)$ is unknown. Explain how you would obtain the spectral envelope of $s(n)$ based on $c_s(n)$.

(5 marks)

(iii) Draw on Fig. Q1 the spectral envelope that you may obtain in Q1(b)[ii]. You may detach Page 2 of this exam paper and attach it to your answer book.

(4 marks)

Q2 (a) In spectral subtraction for speech enhancement, the noisy speech signal $y(n)$, clean speech signal $x(n)$ and background noise $b(n)$ have the following relationship:

$$y(n) = x(n) + b(n).$$

(i) Show that the magnitude spectra of $y(n)$, $x(n)$, and $b(n)$ have the following relationship:

$$|Y(\omega)| = \sqrt{|X(\omega)|^2 + |B(\omega)|^2}.$$

State the assumption that you have made.

(5 marks)

(ii) Show that the clean spectrum can be estimated by

$$\hat{X}(\omega) = \begin{cases} \left[|Y(\omega)|^2 - |B(\omega)|^2\right]^{\frac{1}{2}} e^{j\varphi_y(\omega)} & \text{if } |Y(\omega)|^2 > |B(\omega)|^2 \\ 0 & \text{otherwise} \end{cases} \quad \text{(Eq. Q2–1)}$$

where $\varphi_y(\omega)$ is the phase spectrum of $y(n)$.

(5 marks)

(iii) What is the potential problem in using Eq. Q2–1 for speech enhancement when the signal-to-noise ratio of the noisy speech is less than 0dB? Briefly explain your answer and suggest a modification to Eq. Q2–1 so that the problem becomes less severe.

(10 marks)

(b) Fig. Q2 shows a code-excited linear prediction (CELP) decoder.

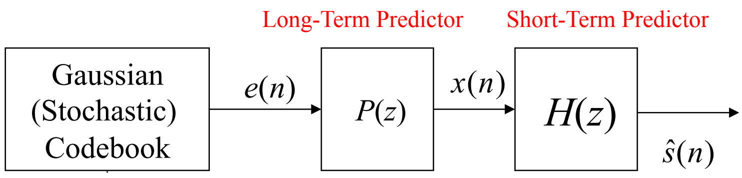


Fig. Q2

(i) What is the purpose of the long-term predictor $P(z)$?

(2 marks)

(ii) Discuss the quality of the decoded speech $\hat{s}(n)$ when $H(z) = 1$.

(3 marks)

Q3 (a) Fig. Q3 shows four hidden markov models (HMMs) of a speech recognizer. Each HMM represents one of the 46 phones (including silence) in English. Each circle in an HMM represents a state.

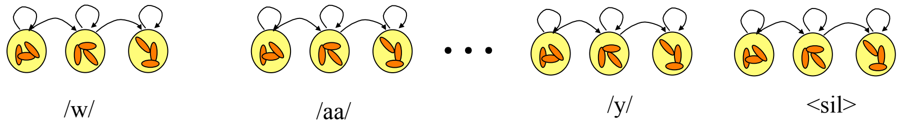


Fig. Q3

(i) As shown in Fig. Q3, each phone-based HMM has three states. Briefly explain what these states represent.

(4 marks)

(ii) What do the arrows in Fig. Q3 represent? Why are they important for acoustic modeling in speech recognition?

(5 marks)

(iii) In conventional GMM-HMM speech recognizers, each state is a Gaussian mixture model (GMM). Describe how these GMMs can be trained from a speech corpus. You may assume that the utterances in the corpus have been phonetically transcribed, i.e., the positions of phonetic events in each utterance are known.

(6 marks)

(b) Assume that the acoustic vectors of an utterance are given by $\mathcal{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$, where T is the number of frames in the utterance. Denote the parameters of a Gaussian mixture model (GMM) as $\mathbf{\Lambda} = \{\pi_j, \boldsymbol{\mu}_j, \mathbf{\Sigma}_j\}_{j=1}^{M}$, where π_j, $\boldsymbol{\mu}_j$, and $\mathbf{\Sigma}_j$ are the mixture coefficient, mean vector, and covariance matrix of the j-th Gaussian, respectively. To use maximum-likelihood linear regression (MLLR) for adapting the mean vectors of the GMM, we may estimate the transformation parameters $(\hat{\mathbf{A}}, \hat{\mathbf{b}})$ as follows:

$$\begin{aligned}(\hat{\mathbf{A}}, \hat{\mathbf{b}}) &= \arg\max_{\mathbf{A},\mathbf{b}} \sum_{t=1}^{T} \log p(\mathbf{o}_t | \mathbf{A}, \mathbf{b}, \mathbf{\Lambda}) \\ &= \arg\max_{\mathbf{A},\mathbf{b}} \left\{ \sum_{t=1}^{T} \log \sum_{j=1}^{M} \pi_j \mathcal{N}(\mathbf{o}_t | \mathbf{A}\boldsymbol{\mu}_j + \mathbf{b}, \mathbf{\Sigma}_j) \right\},\end{aligned}$$

where $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma})$ represents a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$.

(i) If the dimension of $\mathbf{o}_t$ is 39, what is the dimension of $\hat{\mathbf{A}}$?

(2 marks)

(ii) Express the adapted mean vector $\hat{\boldsymbol{\mu}}_j$ in terms of $\hat{\mathbf{A}}$, $\hat{\mathbf{b}}$, and $\boldsymbol{\mu}_j$.

(4 marks)

(iii) If we only want to adapt the mean vectors that are very close to the acoustic vectors in $\mathcal{O}$, should we use MLLR for the adaptation? Briefly explain your answer.

(4 marks)

Q4 (a) In i-vector based speaker verification, the GMM-supervector $\vec{\boldsymbol{\mu}}$ representing an utterance is assumed to follow a factor analysis model:

$$\vec{\boldsymbol{\mu}} = \vec{\boldsymbol{\mu}}^{(b)} + \mathbf{T}\mathbf{w} \qquad \text{(Eq. Q4–1)}$$

where $\vec{\boldsymbol{\mu}}^{(b)}$ is a GMM-supervector corresponding to a universal background model (UBM), $\mathbf{T}$ is a low-rank total variability matrix, and $\mathbf{w}$ is a low-dimensional latent factor.

(i) Discuss the purpose of matrix $\mathbf{T}$. Why is it important?

(5 marks)

(ii) To extract the i-vector from an utterance, we need to align the acoustic vectors of the utterance against the UBM and then compute the posterior mean of $\mathbf{w}$. Assuming that the acoustic vectors are of dimension 60 and that the UBM comprises 1024 Gaussians, what are the dimensions of $\boldsymbol{\mu}$ and $\mathbf{T}$ when the dimension of $\mathbf{w}$ is 500?

(4 marks)

(iii) Given the acoustic vectors $\mathcal{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$ of an utterance, its i-vector $\mathbf{x}$ is the posterior mean of $\mathbf{w}$ in Eq. Q4–1, i.e., $\mathbf{x} = \mathbb{E}\{\mathbf{w}|\mathcal{O}\}$. State one advantage of using i-vectors for speaker recognition. What is the advantage of using i-vectors rather than the GMM-supervectors ($\vec{\boldsymbol{\mu}}$ in Eq. Q4–1) for speaker recognition?

(6 marks)

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

$$\mathbf{x} = \mathbf{m} + \mathbf{V}\mathbf{z} + \boldsymbol{\epsilon} \qquad \text{(Eq. Q4–2)}$$

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 with covariance matrix $\boldsymbol{\Sigma}$.

(i) Explain why it is necessary to model i-vectors by Eq. Q4–2.

(4 marks)

(ii) Denote $\mathbf{x}_s$ and $\mathbf{x}_t$ as the i-vectors of target-speaker s and test speaker t, respectively. Assume that the target-speaker and the test speaker are the same person. Show that the joint likelihood of $\mathbf{x}_s$ and $\mathbf{x}_t$ is given by

$$p(\mathbf{x}_s, \mathbf{x}_t|\text{Same speaker}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}_s \\ \mathbf{x}_t\end{bmatrix} \middle| \begin{bmatrix}\mathbf{m} \\ \mathbf{m}\end{bmatrix}, \begin{bmatrix}\mathbf{V}\mathbf{V}^\mathsf{T} + \boldsymbol{\Sigma} & \mathbf{V}\mathbf{V}^\mathsf{T} \\ \mathbf{V}\mathbf{V}^\mathsf{T} & \mathbf{V}\mathbf{V}^\mathsf{T} + \boldsymbol{\Sigma}\end{bmatrix}\right)$$

where $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ denotes a Gaussian density function with mean vector $\boldsymbol{\mu}_x$ and covariance matrix $\boldsymbol{\Sigma}_x$. *Hint*: The convolution of Gaussians is also a Gaussian, i.e.,

$$\begin{aligned}\int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} &= \int \mathcal{N}(\mathbf{x}|\mathbf{m} + \mathbf{V}\mathbf{z}, \boldsymbol{\Sigma})\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})d\mathbf{z} \\ &= \mathcal{N}(\mathbf{x}|\mathbf{m}, \mathbf{V}\mathbf{V}^\mathsf{T} + \boldsymbol{\Sigma}).\end{aligned}$$

(6 marks)

– END –