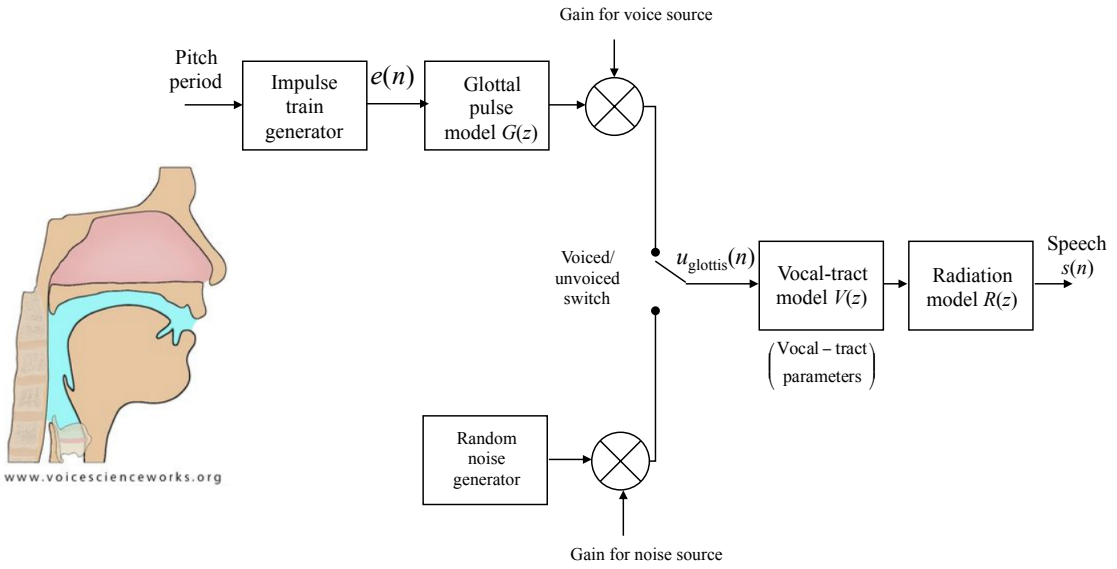
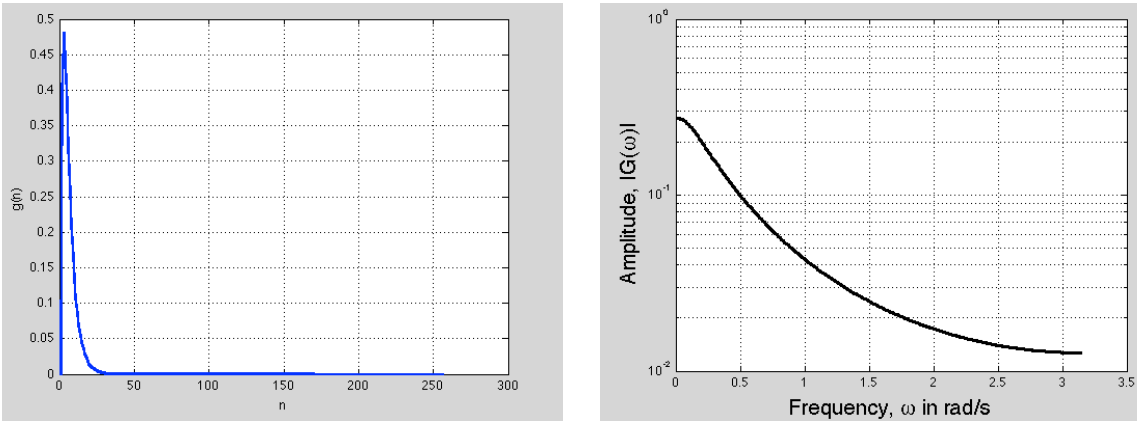
Q1 Fig. Q1(a) and Fig. Q1(b) show the human speech production system and its associated source-filter model, respectively. Fig. Q1(c) and Fig. Q1(d) show the impulse and frequency responses of the glottal shaping filter $G(z)$, respectively.



(a) (b)



(c) (d)

Fig. Q1: (a) Speech production system; (b) source-filter model; (c) impulse response of $G(z)$; (d) frequency response of $G(z)$.

(a) Indicate the parts in Fig. Q1(a) that correspond to the filters $V(z)$, $R(z)$, and $G(z)$ in Fig. Q1(b). You may draw on this page of the exam paper and attach it to your answer book.

(3 marks)

(b) Assume that the pitch frequency is 100Hz and that the impulse response of $G(z)$ is shown in Fig. Q1(c). Draw the signals $e(n)$ and $u_{\text{glottis}}(n)$ in Fig. Q1(b) for four pitch cycles. Indicate the pitch period (in milliseconds) in your answers.

(8 marks)

(c) Given the frequency response of $G(z)$ shown in Fig. Q1(d), draw the frequency spectrum of $u_{\text{glottis}}(n)$, i.e., $|U_{\text{glottis}}(\omega)|$. You may assume that $u_{\text{glottis}}(n)$ covers four pitch cycles.

(5 marks)

(d) Assume that the first three formant frequencies of the vocal tract are 1kHz, 2kHz, and 3kHz, respectively. Draw the frequency response of $V(z)$ when the sampling frequency is 8kHz.

(4 marks)

(e) Based on your answers in Q1(a)–(d), draw the frequency spectrum of the speech signal (i.e., $|S(\omega)|$) at the output of the source-filter model. You may assume that the sampling frequency is 8kHz.

(5 marks)

Q2 Fig. Q2 shows the frequency spectrum of a speech frame $s(n)$ with frame size $N = 512$ samples.

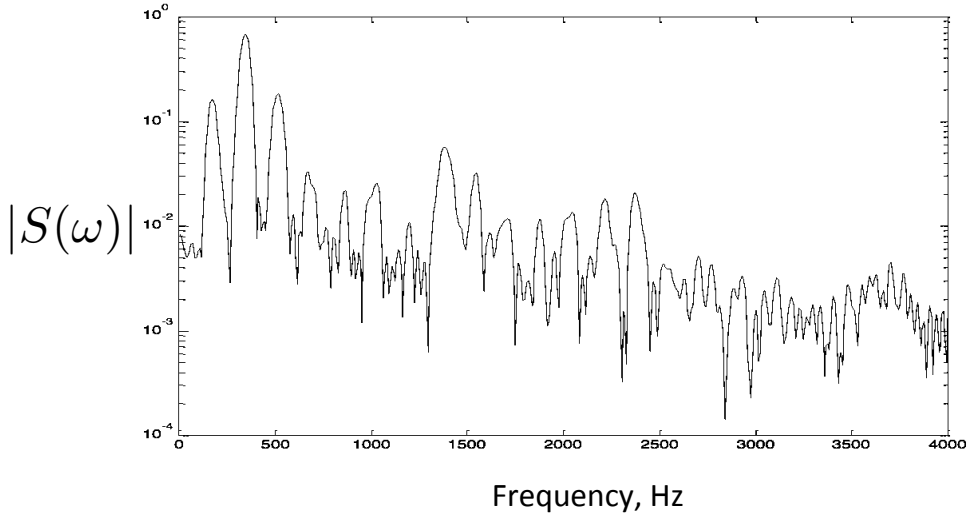


Fig. Q2

(a) State whether the frequency spectrum was obtained from a voiced frame or from an unvoiced frame. Briefly explain your answer.

(3 marks)

(b) Assume that you applied linear prediction (LP) analysis on this speech frame and obtained a set of linear prediction coefficients $\{a_k; k = 1, \ldots, P\}$, where P is the order of the prediction. If the LP filter has the form

$$H(z) = \frac{1}{1 + \sum_{k=1}^{P} a_k z^{-k}},$$

sketch the frequency spectrum

$$|H(\omega)| = \frac{1}{\left|1 + \sum_{k=1}^{P} a_k e^{-j\omega k}\right|}$$

when $P = 12$. Briefly explain your answer. You may draw $|H(\omega)|$ on top of the figure on next page and attach the page to your answer book.

(4 marks)

(c) Repeat Q2(b) for $P = 2$ and $P = 512$. Briefly explain your answer.

(8 marks)

(d) Assume that you applied cepstral analysis on $s(n)$ and obtained a set of cepstral coefficients $\{c(n); n = 0, \ldots, N-1\}$, i.e.,

$$c(n) = \text{real}\{\text{IFFT}(\log|\text{FFT}(s(n))|)\},$$

where FFT and IFFT stand for fast Fourier transform and inverse fast Fourier transform, respectively, and real$\{\cdot\}$ means extracting the real part of IFFT. Sketch the waveform of $c(n)$ for $n = 0, \ldots, N-1$.

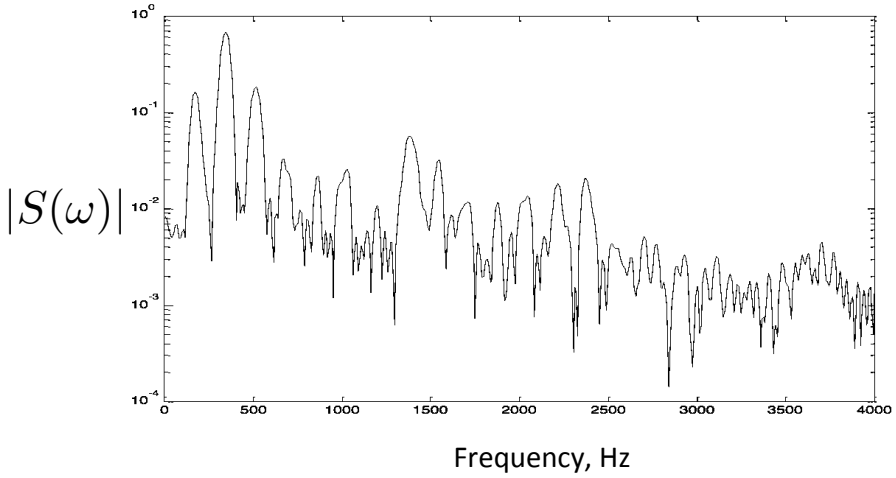
(5 marks)

(e) If a low-time lifter is applied to $c(n)$ such that $c(n) = 0$ for all $n > 20$, sketch the spectrum of $c(n)$ after this low-time liftering. Briefly explain your answer.

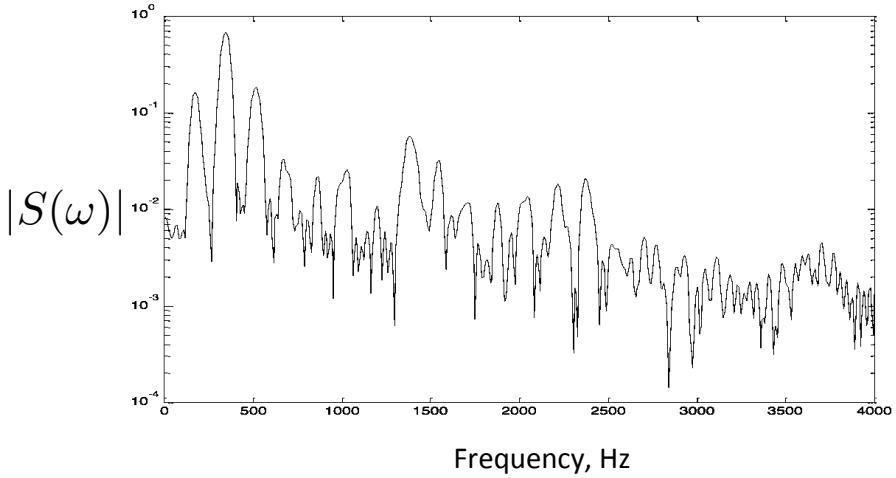
(5 marks)

Name: _____ Student No._____

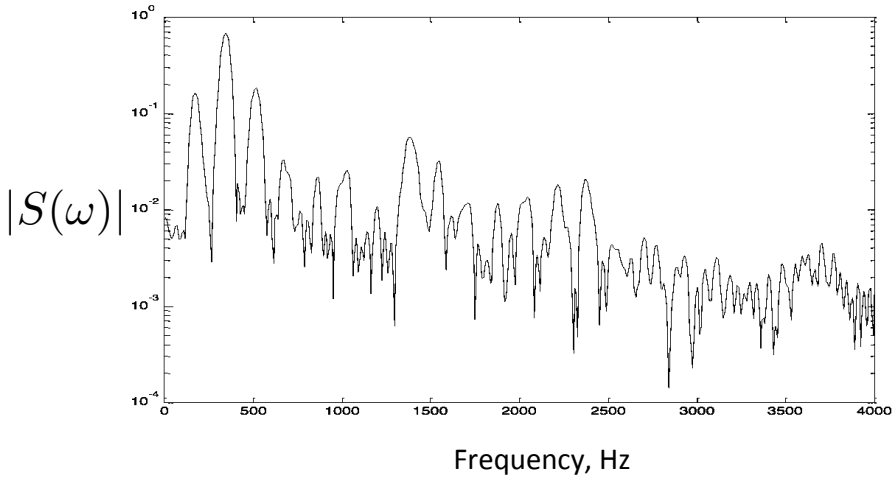
Put your answer of Q2(b) on top of the figure below



Put your answers of Q2(c) on top of the figure below.



Put your answer of Q2(e) on top of the figure below



Q3 (a) The long-term predictor of a code-excited linear prediction (CELP) coder has the form

$$P(z) = \frac{1}{1 - bz^{-D}}.$$

(i) Given that the pitch period of the current subframe is 8ms and that the sampling frequency of the coder is 8kHz, determine the value of D. Briefly explain your answer.

(2 marks)

(ii) The value of b depends on whether the subframe is voiced or unvoiced. If $b = 0.9$ for a voiced subframe, suggest a value of b for an unvoiced subframe. Briefly explain your answer.

(4 marks)

(b) Fig. Q3 shows a phone-based hidden Markov model (HMM). 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

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

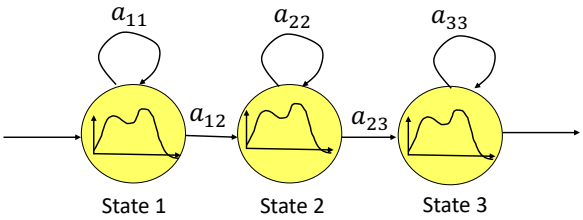


Fig. Q3

(i) If a_{ij} is the probability of transiting from state i to state j, what is the value of $a_{11} + a_{12}$?

(1 mark)

(ii) Assume that the waveform of a phoneme can be divided into three sections: initial, middle, and final. Assume also that these three sections can be modeled by State 1, State 2, and State 3 of the HMM in Fig. Q3, respectively. For a phoneme with very short duration at its initial section, will a_{11} be close to 1 or close to 0? Briefly explain your answer.

(3 marks)

(iii) The structure of the HMM in Fig. Q3 suggests that $a_{13} = 0$. What is the implication if a_{13} is non-zero?

(3 marks)

(iv) In most speech recognition systems, the dimension of $\mathbf{o}_t$ is 39. State the components of the acoustic vector $\mathbf{o}_t$.

(3 marks)

(v) Assume that we have two acoustic vector sequences: $\mathcal{O}$ and $\mathcal{O}'$. $\mathcal{O}$ is obtained from the realizations of the phoneme for which the HMM in Fig. Q3 is trained. $\mathcal{O}'$ contains the same set of acoustic vectors as $\mathcal{O}$ but its acoustic vectors are arranged in a reversed order, i.e.,

$$\mathcal{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$$
$$\mathcal{O}' = \{\mathbf{o}_T, \mathbf{o}_{T-1}, \ldots, \mathbf{o}_1\}.$$

Are the values of $p(\mathcal{O})$ and $p(\mathcal{O}')$ the same or different? If they are different, which one is larger?

(4 marks)

(vi) Explain how the GMMs in the HMM states can be replaced by a DNN.

(5 marks)

Q4 (a) To reduce the session variability in GMM-supervectors, nuisance attribute projection (NAP) performs the following linear transformation:

$$\hat{\mathbf{m}} = (\mathbf{I} - \mathbf{V}\mathbf{V}^{\mathsf{T}})\mathbf{m},$$

where $\mathbf{m}$ and $\hat{\mathbf{m}}$ are the original and transformed GMM-supervectors, respectively.

(i) If the universal background model (UBM) has 1024 mixture components (Gaussians) and the dimensionality of the acoustic vectors is 60, determine the dimensionality of $\mathbf{m}$.

(2 marks)

(ii) Based on the result in Q4(a)(i), determine the dimensionality of $\hat{\mathbf{m}}$.

(3 marks)

(iii) What do the columns of $\mathbf{V}$ represent?

(3 marks)

(iv) In GMM–SVM systems, the NAP-projected GMM-supervectors $\hat{\mathbf{m}}$'s from a target-speaker and background speakers are used to train a speaker-specific support vector machine (SVM). Explain why a linear kernel is often used in the speaker-specific SVMs.

(4 marks)

(v) Discuss the advantages of i-vector systems over GMM–SVM systems.

(5 marks)

(b) Fig. Q4 shows the process of extracting GMM i-vectors from a sequence of acoustic vectors $\mathcal{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$.

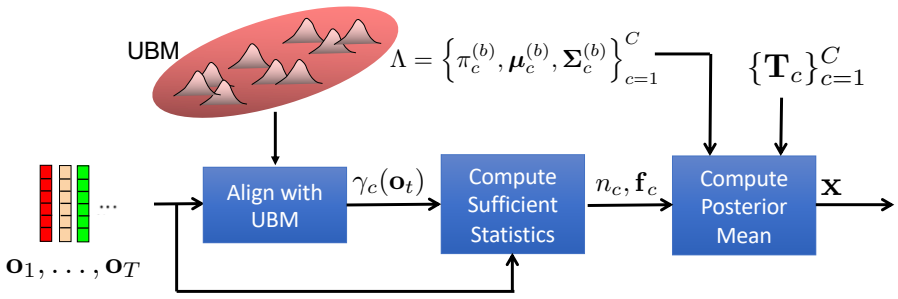


Fig. Q4

In the figure, $\mathbf{T}_c$ is the c-th sub-matrix of the total variability matrix $\mathbf{T}$, i.e., $\mathbf{T} = [\mathbf{T}_1^{\mathsf{T}}\ \mathbf{T}_2^{\mathsf{T}}\ \cdots\ \mathbf{T}_C^{\mathsf{T}}]^{\mathsf{T}}$, where the superscript $^{\mathsf{T}}$ stands for matrix transpose. Also, the mixture coefficients, mean vectors and covariance matrices of the UBM are denoted by $\pi_c^{(b)}$, $\boldsymbol{\mu}_c^{(b)}$,and $\boldsymbol{\Sigma}_c^{(b)}$ for $c = 1, \ldots, C$, respectively. The formula for extracting GMM i-vectors is given by

$$\mathbf{x} = \left(\mathbf{I} + \sum_{c=1}^{C} n_c \mathbf{T}_c^{\mathsf{T}} (\boldsymbol{\Sigma}_c^{(b)})^{-1} \mathbf{T}_c\right)^{-1} \sum_{c=1}^{C} \mathbf{T}_c^{\mathsf{T}} \left(\boldsymbol{\Sigma}_c^{(b)}\right)^{-1} \tilde{\mathbf{f}}_c,$$

where

$$n_c = \sum_{t=1}^{T} \gamma_c(\mathbf{o}_t) \quad \text{and} \quad \tilde{\mathbf{f}}_c = \sum_{t=1}^{T} \gamma_c(\mathbf{o}_t)(\mathbf{o}_t - \boldsymbol{\mu}_c^{(b)}),$$

where

$$\gamma_c(\mathbf{o}_t) = \frac{\pi_c^{(b)} \mathcal{N}(\mathbf{o}_t | \boldsymbol{\mu}_c^{(b)}, \boldsymbol{\Sigma}_c^{(b)})}{\sum_{j=1}^{C} \pi_j^{(b)} \mathcal{N}(\mathbf{o}_t | \boldsymbol{\mu}_j^{(b)}, \boldsymbol{\Sigma}_j^{(b)})}, \quad c = 1, \ldots, C,$$

where $\mathcal{N}(\mathbf{o}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for a Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

(i) If $\mathbf{T}_c$ has 500 columns, determine the dimensionality of the i-vector $\mathbf{x}$.

(2 marks)

(ii) Determine the value of $\sum_{c=1}^{C} \gamma_c(\mathbf{o}_t)$.

(2 marks)

(iii) Explain why i-vectors are compact representation of utterances and they can represent utterances of any duration.

(4 marks)

– END –