DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2017 / 18)

COURSE: EIE558 YEAR: MSc SUBJECT: Speech Processing and Recognition

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER
K: Knowledge A: Application E: Extrapolation	M.W. Mak		

Q1 (a) (i) The bottom panel is the prediction error because the relative magnitude at the positive or negative peaks in the bottom panel is much larger than those in the upper panel. (4 marks, A)

(ii) The speech frame is voiced because of the strong periodicity in the prediction error.

(3 marks, K)

(iii) Pitch period is about 8ms.

(3 marks, K)

(iv) The prediction error is given by

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{P} a_k s(n-k).$$

Taking z-transform on both sides, we have

$$E(z) = S(z) - \sum_{k=1}^{P} a_k S(z) z^{-k} \quad \Rightarrow \quad \frac{S(z)}{E(z)} = H(z) \quad \Rightarrow \quad s(n) = e(n) * h(n),$$

which means that using $e(n)$ as the input to the LP filter

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

we obtain $s(n)$.

(6 marks, AE)

(b) (i) Sampling frequency is 8kHz.

(2 marks, K)

(ii) The upper panel is the narrow-band spectrogram because narrow-band spectrograms use long windows, which can show the harmonic structure (the horizontal stripes) of the speech signal.

(4 marks, K)

(iii) The interval between the horizontal stripes is about 200Hz. So, the pitch period is about $1/200 = 5$ms.

(3 marks, E)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2017 / 18)

COURSE: EIE558 YEAR: MSc SUBJECT: Speech Processing and Recognition

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER
K: Knowledge A: Application E: Extrapolation	M.W. Mak		

Q2 (a) (i) Computing the magnitude spectrum of both sides of $S(\omega) = E(\omega)H(\omega)$ and then taking logarithm, we obtain

$$\begin{aligned} C_s(\omega) = \log|S(\omega)| &= \log|E(\omega)H(\omega)| \\ &= \log|E(\omega)| + \log|H(\omega)| \\ &= C_e(\omega) + C_h(\omega), \end{aligned}$$

where $C_e(\omega)$ and $C_h(\omega)$ are the log-spectrum of $s(n)$ and $h(n)$, respectively. Taking inverse Fourier transform on both sides of this equation, we have

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

(4 marks, K)

(ii) Apply a low-time lifter to $c_s(n)$ so that $c_s(n) = 0\ \forall n > L$, where L is the cut-off quefrency such that L is smaller than the pitch period. Then, apply Fourier transform on the truncated cepstrum $\tilde{c}_s(n)$ to obtain the envelope $C_h(\omega) = \log|H(\omega)|$.

(5 marks, A)

(iii) We apply a high-time lifter to $c_s(n)$ by setting $c_s(n) = 0\ \forall n < L$, where L is smaller than the possible pitch period. Then, the value of n at which the first peak in the liftered cepstrum $\tilde{c}_s(n)$ occurs is the pitch period.

(5 marks, E)

(b) (i) When the input has small amplitude ($|s| < 0.5$ in the figure), the tangents to the curve have slope larger than 1.0. Therefore, small signal amplitude will be expanded. On the other hand, when the input signal has large amplitude ($|s| > 0.5$ in the figure), the tangents to the curve have slope less than 1.0. As a result, large signal will be compressed.

(5 marks, KA)

(ii) Because for human speech, most of the $s(n)$ has small amplitude, expanding them is equivalent to having a finer quantization in the linear quantizer. This ensures that most of the speech samples are quantized with low quantization error, which improve speech quality. On the other hand, compressing the large-amplitude $s(n)$ is equivalent to having a coarser quantizer. However, because only a small fraction of $s(n)$ have large amplitude, it will not cause significant loss in speech quality even if these samples have higher quantization errors. (6 marks, AE)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2017 / 18)

COURSE: EIE558 YEAR: MSc SUBJECT: Speech Processing and Recognition

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER
K: Knowledge A: Application E: Extrapolation	M.W. Mak		

Q3 (a) $a_{12} = 1 - a_{11} = 1 - 0.4 = 0.6$.
If a_{11} is close to zero, the duration of the first part of the phoneme is very short.
(4 marks, KA)

(b) The states can be divided into initial, middle, and final. They represent the spectral distributions of the beginning, middle and ending parts of the corresponding phoneme.
(3 marks, KA)

(c)

$$\begin{aligned} p(\mathcal{O}, \mathbf{q}) &= P(\mathbf{q}) p(\mathcal{O}|\mathbf{q}) \\ &= a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T} \prod_{t=1}^{T} b_{q_t}(\mathbf{o}_t) \\ &= b_{q_1}(\mathbf{o}_1) a_{q_1 q_2} b_{q_2}(\mathbf{o}_2) a_{q_2 q_3} b_{q_3}(\mathbf{o}_3) \cdots a_{q_{T-1} q_T} b_{q_T}(\mathbf{o}_T) \end{aligned}$$

(5 marks, AE)

(d)

$$\begin{aligned} p(\mathcal{O}) &= \sum_{\text{all } \mathbf{q}} p(\mathcal{O}, \mathbf{q}) \\ &= \sum_{q_1} \sum_{q_2} \cdots \sum_{q_T} b_{q_1}(\mathbf{o}_1) a_{q_1 q_2} b_{q_2}(\mathbf{o}_2) a_{q_2 q_3} b_{q_3}(\mathbf{o}_3) \cdots a_{q_{T-1} q_T} b_{q_T}(\mathbf{o}_T). \end{aligned}$$

(5 marks, E)

(e) $p(\mathcal{O}) > p(\mathcal{O}')$ because $\mathcal{O}$ matches the model better than $\mathcal{O}'$.
(3 marks, K)

(f) Use a GMM–HMM to align the training frames of this phone to the 3 states of the HMM. Then, use the alignment results as target labels to train a DNN to output the posteriors of the HMM states, i.e., $P(\text{state} = j|\mathbf{o})$, where $j = 1, 2, 3$. Then, given a test frame $\mathbf{o}$, $b_j(\mathbf{o})$ can be obtained by using Bayes rule as follows:

$$\begin{aligned} b_j(\mathbf{o}_t) &= \frac{P(\text{state} = j|\mathbf{o}) p(\mathbf{o})}{P(\text{state} = j)} \\ &\propto \frac{P(\text{state} = j|\mathbf{o})}{P(\text{state} = j)} \\ &= \frac{\text{DNN}_j(\mathbf{o})}{P(\text{state} = j)}, \quad j = 1, 2, 3, \end{aligned}$$

where $\text{DNN}_j(\mathbf{o})$ is the value of the j-th output node of the DNN given $\mathbf{o}$ as its input. The prior probability $P(\text{state} = j)$ can be obtained by dividing the

number of training frames aligned to state j by the total number of frames.
(5 marks, KA)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2017 / 18)

COURSE: EIE558 YEAR: MSc SUBJECT: Speech Processing and Recognition

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER
K: Knowledge A: Application E: Extrapolation	M.W. Mak		

Q4 (a) (i) The no. of enrollment utterances is at least 1, because there is only one non-zero Lagrange multiplier from the speaker class.

(3 marks, K)

(ii) This is because in practical GMM–SVM systems, the dimensionality of the GMM-supervectors is much larger than the number of training utterances. In such situation, linear SVMs can perform the classification. Non-linear SVMs can easily lead to overfitting.

(5 marks, A)

(iii)
- In GMM–UBM, the client-speaker's GMM and the UBM are trained separately, i.e., once the UBM has been trained, it will be fixed for all client speakers. As a result, no attempt is made to maximize the discrimination between the client-speaker's GMM and the UBM. On the other hand, in GMM–SVM, the SVM of each client speaker is trained to discriminate his/her GMM-supervector(s) and the background speakers' GMM-supervectors. As a result, the discriminative information between speakers can be leveraged.
- In the GMM–UBM score function, the scores from the client-speaker model and the UBM are equally weighted. On the other hand, the scores from the client-speaker and the background speakers are optimally weighted by the Lagrange multipliers, which are trained to maximally separate the client speaker and the background speakers in the supervector space.

(8 marks, AE)

(b) (i) Because the i-vectors contain all sort of variability, we need to remove the non-speaker variabilities and focus on the speaker variability via the speaker subspace matrix $\mathbf{V}$.

(4 marks, KA)

(ii) The mean vector of $\mathbf{x}$'s is

$$\begin{aligned}\mathbb{E}\{\mathbf{x}\} &= \mathbb{E}\{\mathbf{m} + \mathbf{V}\mathbf{z} + \boldsymbol{\epsilon}\} \\ &= \mathbb{E}\{\mathbf{m}\} + \mathbf{V}\mathbb{E}\{\mathbf{z}\} + \mathbb{E}\{\boldsymbol{\epsilon}\} = \mathbf{m}.\end{aligned}$$

The covariance matrix of $\mathbf{x}$'s is

$$\begin{aligned}\mathbb{E}\{(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\mathsf{T}}\} &= \mathbb{E}\{(\mathbf{V}\mathbf{z}+\boldsymbol{\epsilon})(\mathbf{V}\mathbf{z}+\boldsymbol{\epsilon})^{\mathsf{T}}\} \\ &= \mathbf{V}\mathbf{V}^{\mathsf{T}}\mathbb{E}\{\mathbf{z}\mathbf{z}^{\mathsf{T}}\} + \mathbb{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}\} + 2\mathbb{E}\{\mathbf{z}\boldsymbol{\epsilon}^{\mathsf{T}}\} \\ &= \mathbf{V}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma}.\end{aligned}$$

(5 marks, AE)