DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2015 / 16)

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			

1. (a) (i) The signal is voiced because (1) there is a spectral tilt (high-frequency components have lower energy than the low-frequency component) and (2) the frequency spectrum shows periodicity.

(3 marks, K)

(ii) Pitch period $\approx 1/160 = 6.25$ms

(3 marks, K)

(iii) Given the LP coefficients $\{a_k; k = 1, \ldots, P\}$, we obtain the inverse filter

$$A(z) = 1 - \sum_{k=1}^{P} a_k z^{-k}.$$

Then, apply FFT to the sequence $\{1, -a_1, -a_2, \ldots, -a_P, 0, \ldots, 0\}$, we obtain the $|A(e^{j\omega})|$. Taking the inverse, we obtain the spectral envelope $|H(e^{j\omega})| = 1/|A(e^{j\omega})|$.

(5 marks, E)

(b) (i) Computing the magnitude spectrum of both side of $s(n) = e(n) * h(n)$ and then take 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 side of the equation above, we have

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

(5 marks, A)

(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 $c_h(n)$ to obtain the envelope $C_h(\omega) = \log|H(\omega)|$.

(5 marks, A)

(iii) The spectral envelope from cepstral analysis is given in the blue line shown in the diagram below.

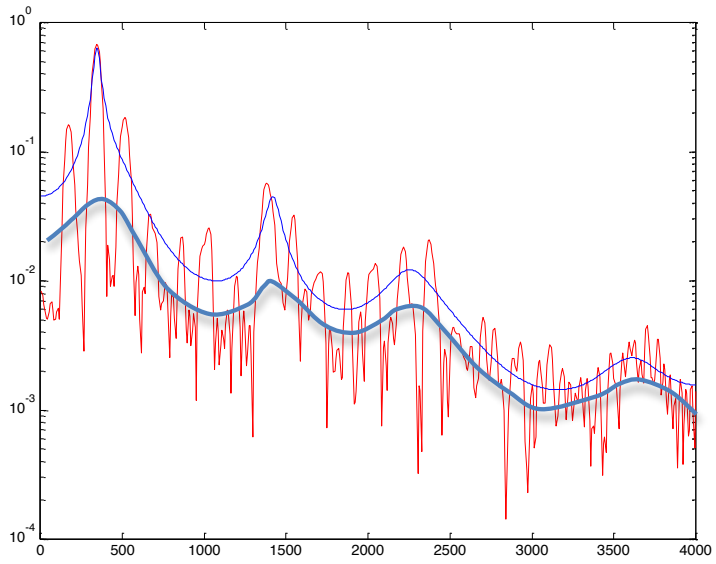
(4 marks, E)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2015 / 16)

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			



Frequency (Hz)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2015 / 16)

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			

2. (a) (i) Taking Fourier transform of $y(n) = x(n) + b(n)$ then computing the squared magnitude, we obtain

$$\begin{aligned} |Y(\omega)|^2 &= |X(\omega) + B(\omega)|^2 \\ &= |X(\omega)|^2 + |B(\omega)|^2 + 2|X(\omega)B(\omega)| \\ &\approx |X(\omega)|^2 + |B(\omega)|^2 \end{aligned}$$

if $x(n)$ and $b(n)$ are uncorrelated, i.e., $R_{xb}(\tau) = 0$.

(5 marks, AE)

(ii) Using the equation in Q2(a)[i], the clean power spectrum can be estimated from $|Y(\omega)|^2 - |B(\omega)|^2$ when $|Y(\omega)|^2 > |B(\omega)|^2$. To obtain the denoised spectrum, we use the phrase spectrum of noisy speech as the phrase of the denoise spectrum, which gives $\hat{X}(\omega) = [|Y(\omega)|^2 - |B(\omega)|^2]^{\frac{1}{2}} e^{j\varphi_y(\omega)}$. When $|Y(\omega)|^2 \leq |B(\omega)|^2$, we set the denoised spectrum to 0 because spectrum cannot be negative.

(5 marks, K)

(iii) When SNR is very low, $|Y(\omega)|^2$ will be less than $|B(\omega)|^2$ many frequencies and many frames, which will cause musical noise in the enhanced speech $\hat{x}(n)$.

(3 marks, A)

The reason is that when $|Y(\omega)|^2 \leq |B(\omega)|^2$, the lower branch of Eq. Q2-1 will be used, causing many sudden changes in frequency across time and frequency axes of the denoised spectrogram. These sudden changes will be perceived as musical noise by listeners. (3 marks, E)

A possible solution is to modify Eq. Q2-1 to

$$\hat{X}(\omega) = \begin{cases} [|Y(\omega)|^2 - \alpha|B(\omega)|^2]^{\frac{1}{2}} e^{j\varphi_y(\omega)} & \text{if } |Y(\omega)|^2 - \alpha|B(\omega)|^2 > \beta|B(\omega)| \\ \beta|B(\omega)|e^{j\varphi_y(\omega)} & \text{otherwise} \end{cases}$$

where $\alpha \in [1, 3]$ is an over-subtraction factor and $\beta > 0$ is a spectral floor factor. (4 marks, K)

(b) (i) $P(z)$ is to generate periodicity in the decoded speech.

(2 marks, K)

(ii) The decoded signal (not sound like speech signal at all) will be tonal and noisy.

(3 marks, E)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2015 / 16)

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			

3. (a) (i) The states can be divided into "initial", "middle", and "final". They represent the spectral characteristics in the beginning, middle and ending parts of the corresponding phoneme.

(4 marks,A)

(ii) The arrows represent state transitions, meaning that at each time step t, they allow the HMM to transit to the next state or remain in the same state with different probabilities. They are important because they enable the GMMs to model the temporal structure of the acoustic vectors.

(5 marks, AE)

(iii) The spectral features (MFCCs) are firstly extracted from a large speech corpus. Then, for each phoneme, say /aa/, we collect the MFCCs corresponding to /aa/ based on the phonetic transcriptions. Then, we apply the EM algorithm to estimate the parameters of the GMMs and transition probabilities corresponding to /aa/ using the MFCCs of /aa/ as the input sequence. We repeat this procedure for all phoneme, including silence.

(6 marks,E)

(b) (i) Dimension $= 39 \times 39$.

(2 marks,K)

(ii) $\hat{\boldsymbol{\mu}}_j = \hat{\mathbf{A}}\boldsymbol{\mu}_j + \hat{\mathbf{b}}$.

(4 marks,A)

(iii) We should not use MLLR, because all Gaussian means will be adapted by $\hat{\boldsymbol{\mu}}_j = \hat{\mathbf{A}}\boldsymbol{\mu}_j + \hat{\mathbf{b}}$, $\forall j = 1, \ldots, M$, regardless of whether they are close to the acoustic vectors.

(4 marks, E)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2015 / 16)

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			

4. (a) (i) The total variability matrix $\mathbf{T}$ defines the subspace in which the GMM-supervectors can vary. It defines the speaker and channel subspaces. It is important because the dimension of GMM supervectors is huge and among these dimensions (axes), only a few hundred are important for describing the speaker and channel variability.

(5 marks, K)

(ii) Dimension of $\boldsymbol{\mu}_i$ is $60(1024) = 61440$
Dimension of $\mathbf{T}$ is 61440×500.

(4 marks, A)

(iii) I-vectors are compact representation of the MFCC vectors. I-vector extraction can be considered as a feature extraction process in which a variable-length acoustic sequence is converted to a fix-size vector.

The dimension of i-vectors is much lower than that of the GMM-supervectors. The low-dimensionality facilities the use of machine learning techniques, such as WCCN, LDA and PLDA, to suppress the channel variability.

(6 marks, AE)

(b) (i) Because the total variability matrix $\mathbf{T}$ represents the combined subspace describing both speaker variability and channel variability, we need to remove the channel variability and focus on the speaker variability via the speaker subspace matrix $\mathbf{V}$ in Eq. Q4-2.

(4 marks, A)

(ii) The joint-likelihood of $\mathbf{x}_s$ and $\mathbf{x}_t$ is

$$\begin{aligned} p(\mathbf{x}_s, \mathbf{x}_t|\text{Same speaker}) &= \int p(\mathbf{x}_s, \mathbf{x}_t, \mathbf{z}) d\mathbf{z} \\ &= \int p(\mathbf{x}_s, \mathbf{x}_t|\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \\ &= \int \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix} + \begin{bmatrix} \mathbf{V} \\ \mathbf{V} \end{bmatrix} \mathbf{z}, \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \end{bmatrix} \right) \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) d\mathbf{z} \\ &= \mathcal{N}\left(\begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_t \end{bmatrix} \middle| \begin{bmatrix} \mathbf{m} \\ \mathbf{m} \end{bmatrix}, \begin{bmatrix} \mathbf{V}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma} & \mathbf{V}\mathbf{V}^{\mathsf{T}} \\ \mathbf{V}\mathbf{V}^{\mathsf{T}} & \mathbf{V}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma} \end{bmatrix} \right) \end{aligned}$$

(6 marks, E)