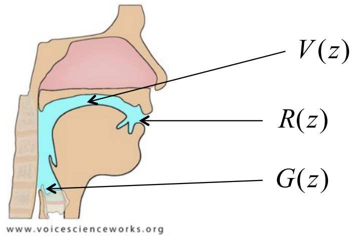
DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF ______EIE______

SOLUTION & MARKING SCHEME (2018 / 19)

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

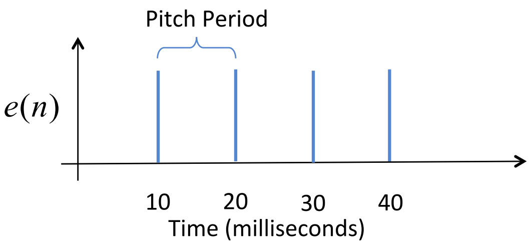
	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER
K: Knowledge A: Application E: Extrapolation	M.W. Mak		

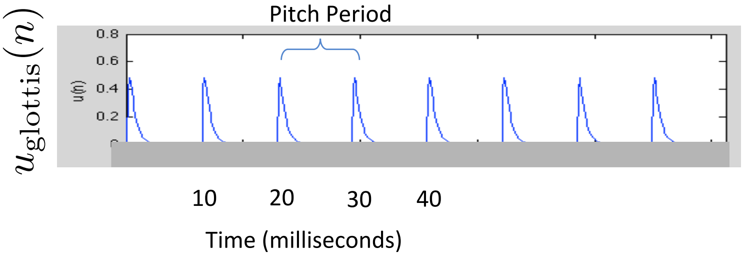
Q1 (a) 1 marks for each filter.



(3 marks, K)

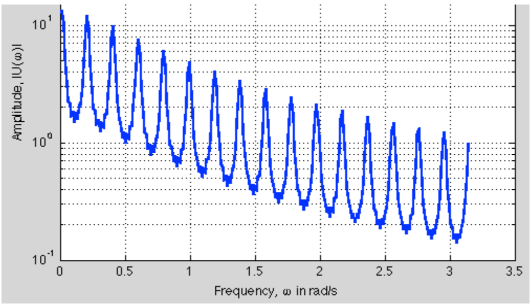
(b) 4 marks for each figure.





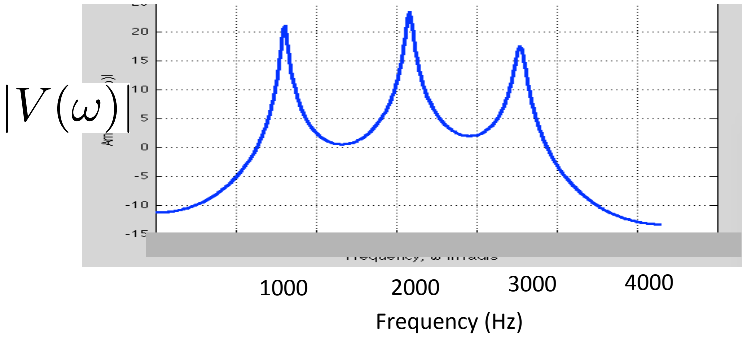
(8 marks, A)

(c) The frequency spectrum should be periodic and the spectral slope should be negative.



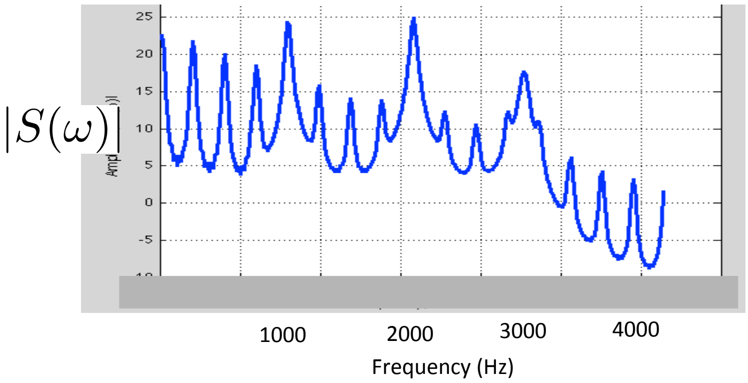
(5 marks, AE)

(d) $|V(\omega)|$ should peak at 1kHz, 2kHz, and 3kHz.



(4 marks, KA)

(e) $|S(\omega)|$ should peak at 1kHz, 2kHz, and 3kHz. It should also contains harmonics and magnitude decreases with frequency.



(5 marks, E)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2018 / 19)

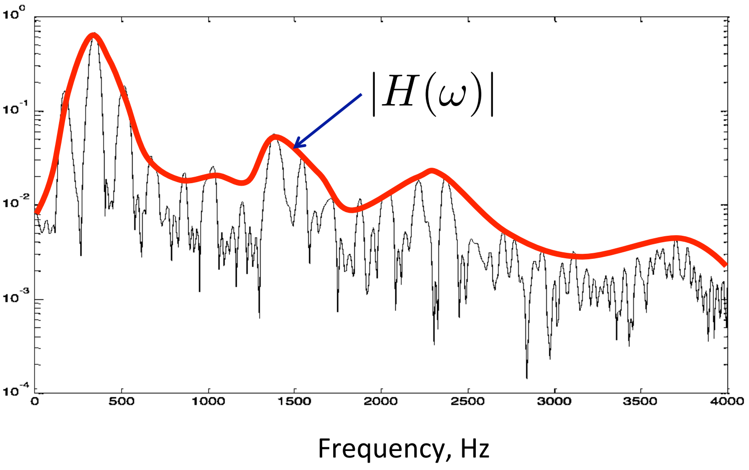
COURSE: EIE558 YEAR: MSc SUBJECT: Speech Processing and Recognition

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

Q2 (a) The spectrum is obtained from a voiced frame because the spectrum shows harmonics (periodicity) and the spectral slope is negative (magnitude decreases with frequency).

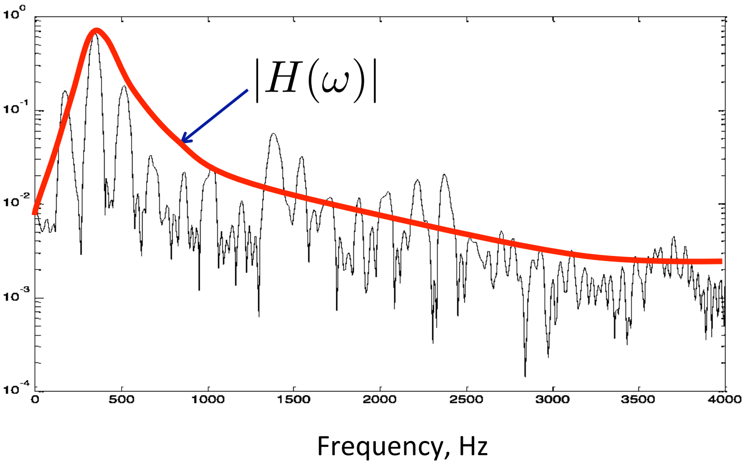
(3 marks, K)

(b) The envelope should generally track the peaks of $H(\omega)$. This is because the LP model can separate the speech spectrum into the source (harmonic information represented by the prediction error) and the vocal tract filter (represented by the LP coefficients). Using the LP coefficients to plot the frequency spectrum will lead to the spectral envelope.



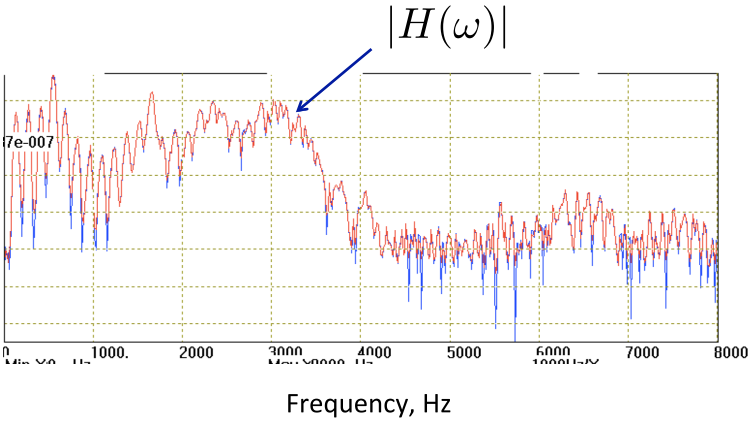
(4 marks, KA)

(c) When $P = 2$, there is only one complex pole pair. As a result, there is only one peak in the envelope and it should track the main peak in $|H(\omega)|$.



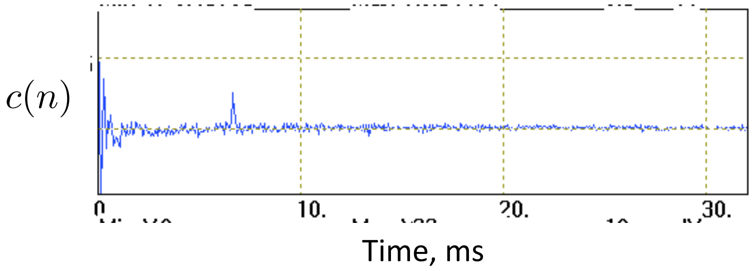
(4 marks, A)

When $P = 512$, there are more complex pole pairs than the number of peaks in the speech spectrum $|S(\omega)|$, As a result, the LP envelop can tract almost all of the peaks in the speech spectrum. *Note*: Students do not have to draw exactly like the one show below. But they need to demonstrate that the spectral envelope follows the speech spectrum closely.



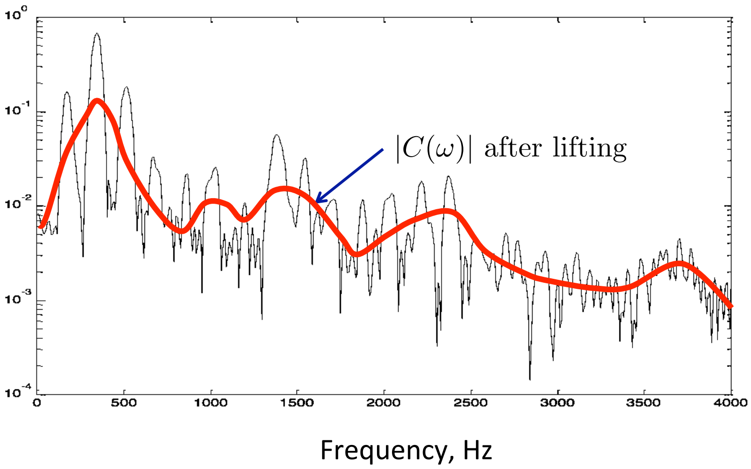
(4 marks, AE)

(d) There should be a small peak in $c(n)$ corresponding to the pitch period.



(5 marks, AE)

(e) The envelope should track the middle between the peaks and troughs of $|S(\omega)|$. Because $c(n)$ is not obtained from an all-pole model, it will not track the peaks of the speech spectrum. The envelop given by $|C(\omega)|$ is smooth because the low-time lifter remove the part of the cepstrum corresponding to the harmonic information of the speech signal.


10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
500
1000
1500
2000
2500
3000
3500
4000
$|C(\omega)|$ after lifting
Frequency, Hz

(5 marks, E)

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

Q3 (a) (i) D is the number of samples in a pitch period. When the pitch period is 8ms and sampling rate is 8kHz,

$$D = 0.008 \times 8000 = 64.$$

(2 marks, K)

(ii) b is the gain of the long-term predictor, which aims to model the periodicity of the synthetic speech. As unvoiced speech is weak in periodicity, the value of b must be much smaller than 0.9 (say $b = 0.1$) but it must be non-negative.

(4 marks, A)

(b) (i) $a_{11} + a_{12} = 1$

(1 mark, K)

(ii) a_{11} will be close to 0 because the chance of transiting to the next state should be very high if the initial part of the phoneme is very short.

(3 marks, KA)

(iii) This means that state skipping is permitted. In other words, frames may be generated by the GMM in State 1 and State 3 only.

(3 marks, A)

(iv) An acoustic vector comprises 12 MFCCs and log-energy plus their first and second derivative:

$$\mathbf{o}_t = [e, o_{t,1}, \ldots, o_{t,12}, \Delta e, o_{t,1}, \ldots, \Delta o_{t,12}, \Delta\Delta e, o_{t,1}, \ldots, \Delta\Delta o_{t,12}]^{\mathsf{T}}$$

(3 marks, K)

(v) They are different. $p(\mathcal{O}) > p(\mathcal{O}')$ because the vector sequence in $\mathcal{O}'$ does not match the 3 states and the transition probabilities of the HMM.

(4 marks, E)

(vi) We can replace the conditional likelihoods $p(\mathbf{o}_t|\text{State} = j)$ by the DNN's output. As the DNN can only output posterior probabilities of states, we may turn posterior into likelihood by using the Bayes rule:

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

where $\text{DNN}_j(\mathbf{o}_t)$ is the value of the j-th output node of the DNN given $\mathbf{o}_t$ as its input.

(5 marks, AE)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2018 / 19)

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) Dimensionality of $\mathbf{m}$ is $1024 \times 60 = 61440$.

(2 marks, K)

(ii) Dimensionality of $\mathbf{m}$ is the same as $\hat{\mathbf{m}}$, which is 61440. The reason is that $\mathbf{V}\mathbf{V}^{\mathsf{T}}$ is a symmetric square matrix of size 61440×61440.

(3 marks, A)

(iii) The columns of $\mathbf{V}$ represent the directions of maximum variability in the supervectors due to the presence of nuisance attributes, e.g., noise and channel effects.

(3 marks, E)

(iv) 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.

(4 marks, AE)

(v) In i-vector systems, variable-length utterances are represented by a low-dimensional vectors called i-vectors. Because of the low-dimensionality, many classical machine learning methods can be used to remove the channel effects and domain mismatch. On the other hand, the GMM–SVM systems use SVMs to classify GMM-supervectors. As the dimensionality of GMM-supervectors is extremely high, they are typically classified by SVMs. Also, in GMM–SVM sytems, each target speaker has his/her own SVM. As SVMs require a lot more memory than the i-vectors for scoring, GMM–SVM systems require a lot more memory to operate.

(5 marks, AE)

(b) (i) Dimension of $\mathbf{x}$ is 500.

(2 marks, K)

(ii) $\sum_{c=1}^{C} \gamma_c(\mathbf{o}_t) = 1$

(2 marks, K)

(iii) The i-vectors are compact because their dimension is typically very small (500) when compared with the GMM-supervectors (61440). i-vectors can represent utterances of any duration because given the acoustic vectors of an utterance, the zeroth- and first-order sufficient statistics (n_c and $\tilde{\mathbf{f}}_c$) provide a summary of all of the frames in the entire utterance.

(4 marks, E)