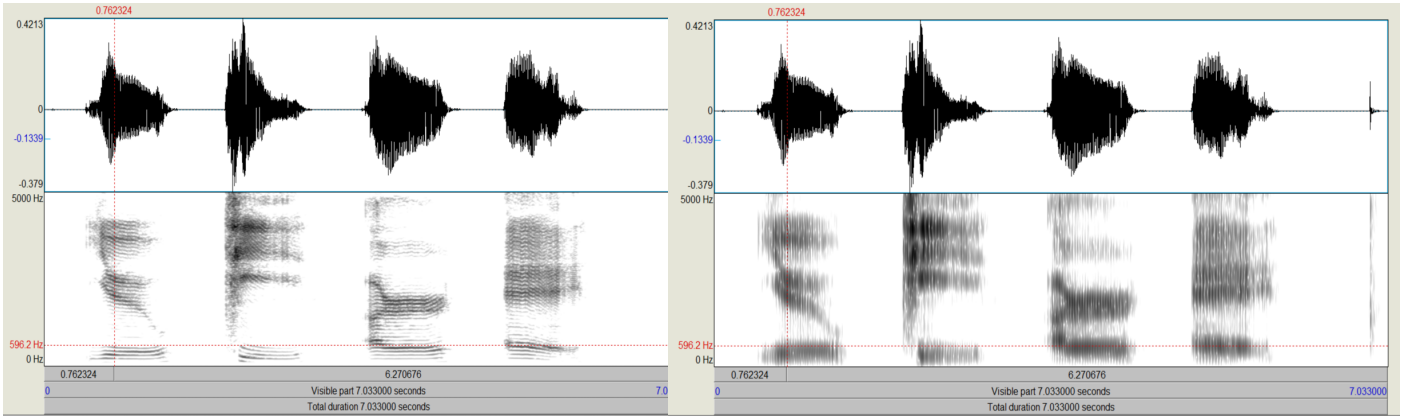
UCLA — Electrical and Computer Engineering Dept.
ECE114: Speech and Image Processing — Take-home Speech Exam
Due November 6, 2020 at 4pm (pacific standard time)

This exam has 4 questions, for a total of 100 points.

Open book. Calculators allowed, but you must show all work. When asked to explain your reasoning, please give a TYPED explanation. Full credit will not be given without proper justification where asked.
Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page.

Question	Points	Score
1	30	
2	24	
3	30	
4	16	
Total:	100	

Please go to next page...



(a) One spectrogram. **(b)** Another spectrogram.

Figure 1: Two spectrograms

1. Fig. 1 shows two spectrograms taken using different window lengths of a person saying "UCLA." However, the person misspoke, and two of the phonemes are incorrect.
   (a) (4 points) Transcribe the phrase "UCLA" in the Arpabet.
   (b) (6 points) Mark the time regions that correspond to each phoneme in the spectrogram. You can do this by drawing vertical lines on the spectrogram that segment it into regions containing one phoneme each.
   (c) (6 points) Identify each region you drew as voiced or unvoiced.
   (d) (4 points) Which phonemes were said incorrectly? How do you know?
   (e) (4 points) Which spectrogram is a wideband and which is a narrowband spectrogram? How do you know?
   (f) (6 points) Identify the pitch of the speaker? Which spectrogram did you use, and how did you identify the pitch?

2. Construct a filter to approximate the vocal tract transfer function of an /i/ sound.

   (a) (2 points) List the first three formants of an /i/ sound.

   (b) (6 points) Assuming that we sample at 44 100 Hz, write a stable transfer function that has poles at the locations of these formants. You may select any reasonable magnitude for the poles.

   (c) (6 points) Draw a zero pole diagram for the system.

   (d) (10 points) What is the smallest filter order possible such that if this vocal tract impulse response, $v(n)$, were convolved with a glottal pulse signal, $x(n)$, from someone with a pitch of 150 Hz, the pitch harmonics would be resolved. Assume that the filter order is the size of the DFT of $x(n)$.

3. Consider the three-tube model of Fig. 2 (the asterisk indicates the position of the vocal cords). Assume that the length of the vocal tract is $l = l_1 + l_2 + l_3 = 17.5\,\text{cm}$ and the speed of sound is $c = 360\,\text{m/s}$. The cross-sectional areas are assumed to be $A_1 = A_3 = \pi d_1^2/4$ and $A_2 = \pi d_2^2/4$.

Figure 2: Three-tube model for vowels.

(a) (10 points) Compute the nomogram for $0 < l_1 < 12.5\,\text{cm}$, $l_2 = 5\,\text{cm}$, with the choice of $d_1 = 3.4\,\text{cm}$ and $d_2 = 1\,\text{cm}$. For this problem, compute the resonance frequencies F_1, F_2, and F_3 of the entire system for each value of l_1 from 0 to 12.5cm in increments of 0.5 cm. Make a table of the resonances from each decoupled tube. Clearly label the columns and rows of the table and submit it. Then plot the values in the table corresponding to each of the first three formants to form a nomogram as shown in class. You may find the min() function helpful for this. Label which curve corresponds to which tube and to which formant frequency in the plot. You may label it how you wish as long as it is clear. You may do this problem by hand or in a graphing program like MATLAB or Excel. You need not include any code, but you must explain your calculations and write out all equations used. We assume no acoustic coupling in this part of the problem.

(b) (5 points) On your plot, revise the nomogram taking into account the phenomenon of acoustic coupling. Draw (by hand is OK) the lines corresponding to where you expect the three formants F_1, F_2, and F_3 to be for all values of l_1. You do not need to be precise.

(c) (10 points) As you did in part a, plot another nomogram for $0 < l_1 < 12.5\,\text{cm}$, $l_2 = 5\,\text{cm}$, with the choice of $d_1 = 3.4\,\text{cm}$ and $d_2 = 2.4\,\text{cm}$. How is this nomogram different from what you drew before?

(d) (5 points) The formants of the vowels /a/, /o/, /u/, and /i/ can be estimated using this three-tube model with $d_2 = 1\,\text{cm}$ for values of $l_1 = 0.5\,\text{cm}, 2.5\,\text{cm}, 7.5\,\text{cm}, 9.1\,\text{cm}$, respectively. The vowel /e/ is obtained with the same parameters as the vowel /i/, but with a wider constriction, i.e., $d_2 = 2.4\,\text{cm}$. Based on these considerations, list the formants F_1 and F_2 for the vowels /a/, /o/, /u/, /i/, and /e/.

4. Consider the speech segment
$s(n) = [2, -1, \underline{1}, 3, -1, -2, -1, 1, 3, -1, -2, -1, 1, 3, -1, -2, -1, 2, 1, 1, 3, -2, 1, 2, -3, -1, 1]$. The correlation function is given by

$$R(i) = \sum_{n=i}^{N_{\mathrm{w}}-1} s^{\mathrm{w}}(n) s^{\mathrm{w}}(n-i), \quad 0 \le i \le N_{\mathrm{w}} - 1,$$

where $s^{\mathrm{w}}(n) = s(n)w(n)$, $w(n)$ is a rectangular window of length $N_{\mathrm{w}} = 15$. Let the model order be $p = 3$.

(a) (4 points) Compute $R(i)$, $i = 0, \ldots, p$

(b) (4 points) Find the 3nd-order prediction coefficients, a_1, a_2, and a_3, using the correlation method of linear prediction analysis.

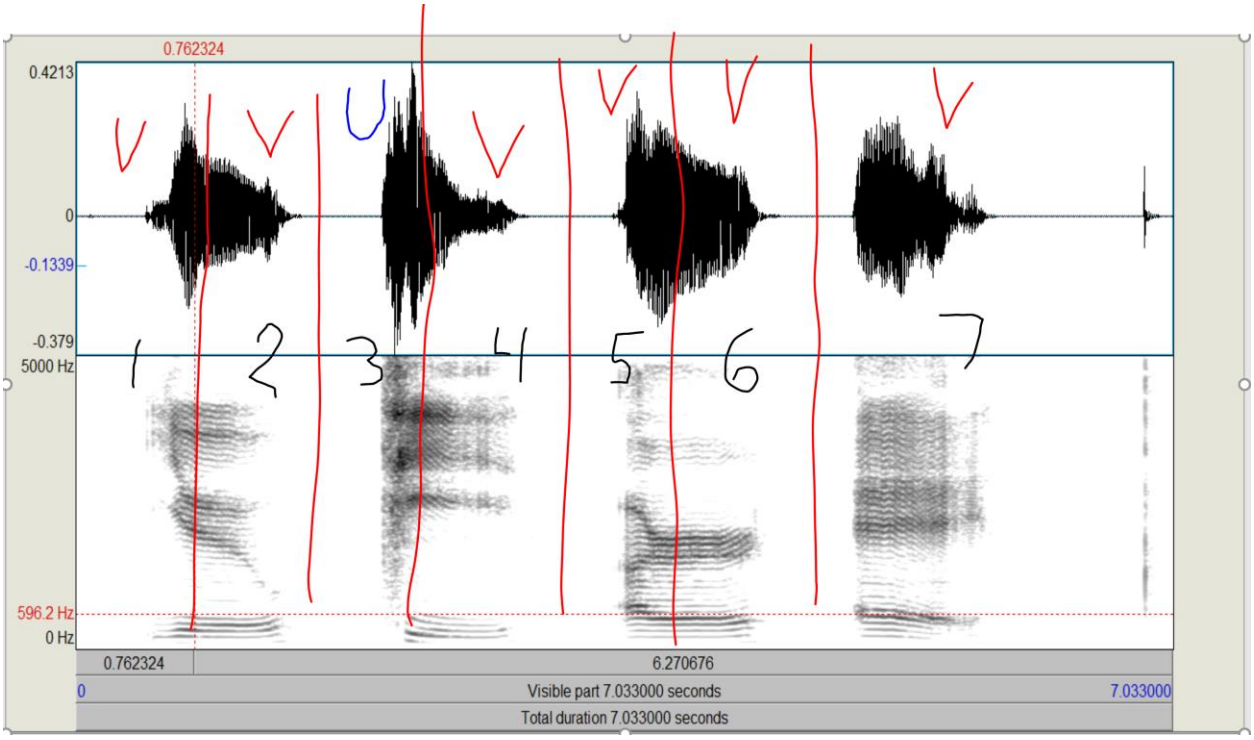
(c) (4 points) Find the corresponding error, $E_{\min}$. Write the expression for the vocal tract's transfer function, $V(z)$.

(d) (4 points) Compute the expression for the poles of the corresponding vocal tract model. Are they real or complex conjugate? How many formants are there? Explain. Let the sampling frequency be $F_{\mathrm{s}} = 8000\,\mathrm{Hz}$.

Please go to next page...

1a. UCLA": Y-UX-S-IY-EH-L-EY

1b-c.



1d. Phoneme 3 is an S. It is too short in duration and has clearer formant structure than we would expect from a fricative. The noise is also at the wrong frequency for an S sound. The short duration and formant frequency of the surrounding vowel suggests that it is instead a plosive. It is in fact a T. Phoneme 6 is not an L. F3 of the preceding vowel is much lower than it would be if the next phoneme were an L. This is in fact an R.

1e. Spectrogram a (left) is a narrowband spectrogram. You can tell by the clearly resolved horizontal striations. Spectrogram b (right) is a wideband spectrogram. You can tell by the vertical striations and dark, wider formant bands.

1f. We should use the narrowband spectrogram to identify pitch. We can count the average number of pitch harmonics over a region of the frequency axis to find the average spacing between them, F0. There are 5 horizontal lines corresponding to pitch peaks up until ~600Hz. This gives a pitch estimate of about 120Hz. The lowest line is difficult to see. If you counted four horizontal lines and gave an answer of 150Hz then that answer is also acceptable if clearly explained.

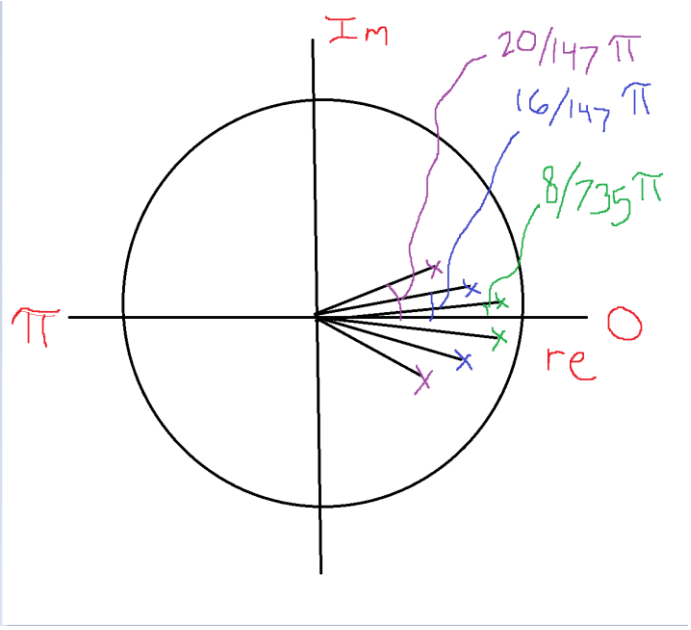
2a. The first 3 formants of an /i/ sound are approximately 240Hz, 2400Hz, and 3000Hz. The values need not be exact, as they change from speaker to speaker.

2b. At a sampling rate of $F_s = 44100Hz$, the formant locations correspond to $F_1 = \frac{240*2\pi}{F_s} = \frac{8}{735}\pi, F_2 = \frac{2400*2\pi}{F_s} = \frac{16}{147}\pi, F_3 = \frac{3000*2\pi}{F_s} = \frac{20}{147}\pi$ radians. We then need to choose magnitudes for the poles that

will result in a stable transfer function. This means that the magnitude of each pole must be less than 1. We might choose the poles to have decreasing magnitudes, ex 0.9, 0.8, and 0.7 respectively. Each formant location must be represented by a complex conjugate pair of poles. We then choose a gain of 1 for simplicity and construct a transfer function for out all pole model of the vocal tract as follows:

$$V(z) = \frac{1}{\left(1 - .9e^{\frac{j8}{735}}z^{-1}\right)\left(1 - .9e^{\frac{-j8}{735}}z^{-1}\right)\left(1 - .8e^{\frac{j16}{147}}z^{-1}\right)\left(1 - .8e^{\frac{-j16}{147}}z^{-1}\right)\left(1 - .7e^{\frac{j20}{147}}z^{-1}\right)\left(1 - .7e^{\frac{-j20}{147}}z^{-1}\right)}$$

2c.



2d.

In time: $s[n] = x[n] * v[n]$

In frequency (DFT): $S[k] = X[k]V[k]$

We need X[k] and V[k] to be taken with an N-point DFT with N large enough to resolve frequency bins spaced at a distance less than or equal to the spacing between the pitch harmonics, 150Hz. This means that we need $150Hz \geq \frac{F_s}{N}$ or $N \geq \frac{44100}{150} = 294$

3a.

Tube back tube (closest to gottis) – closed on both ends: $F_n = \frac{cn}{2l}, n = 1,2,3 \ldots$

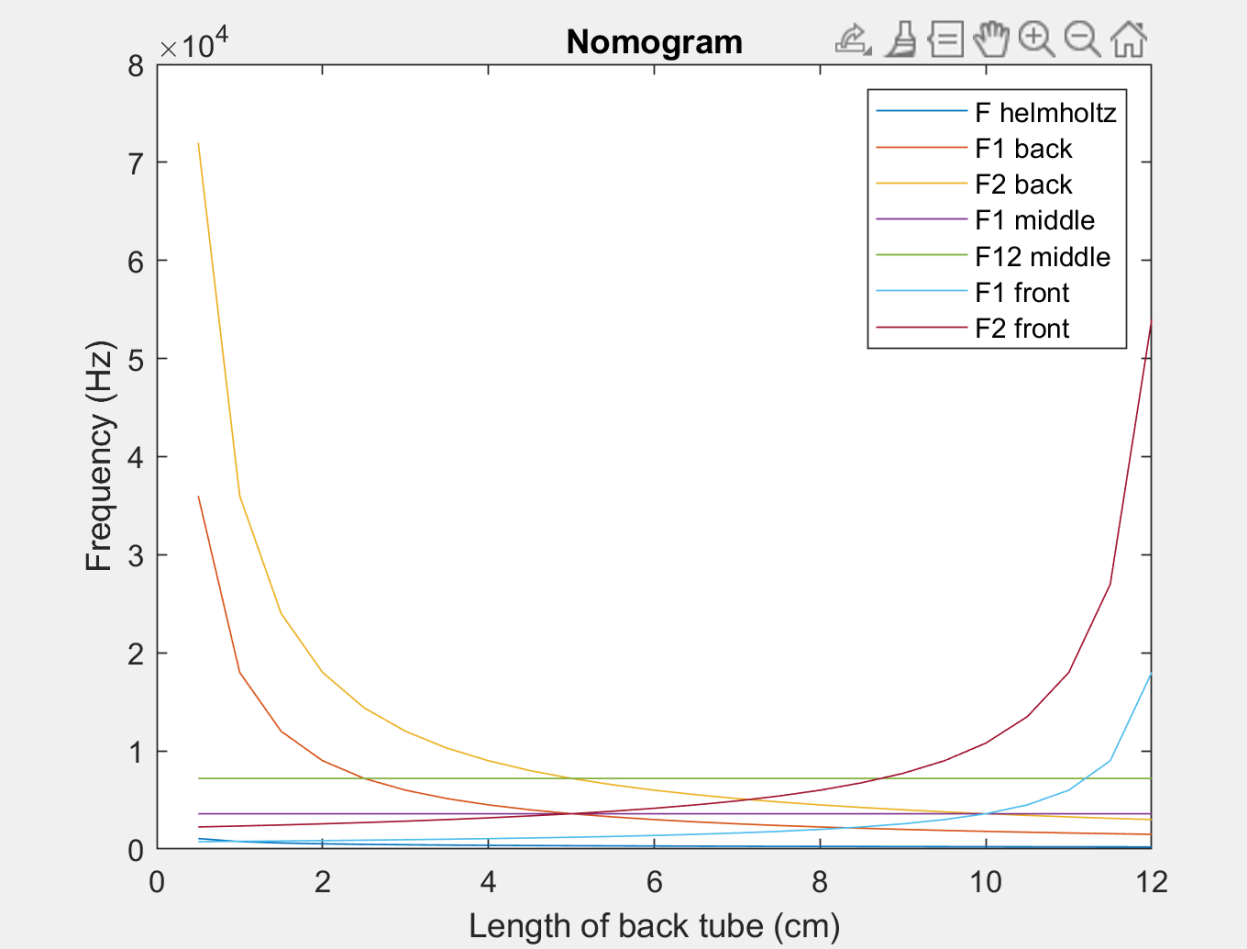
Middle tube – open on both ends: $F_n = \frac{cn}{2l}, n = 1,2,3 \ldots$

Front tube – open on one end, closed on other, $F_n = \frac{c(2n-1)}{4l}, n = 1,2,3 \ldots$

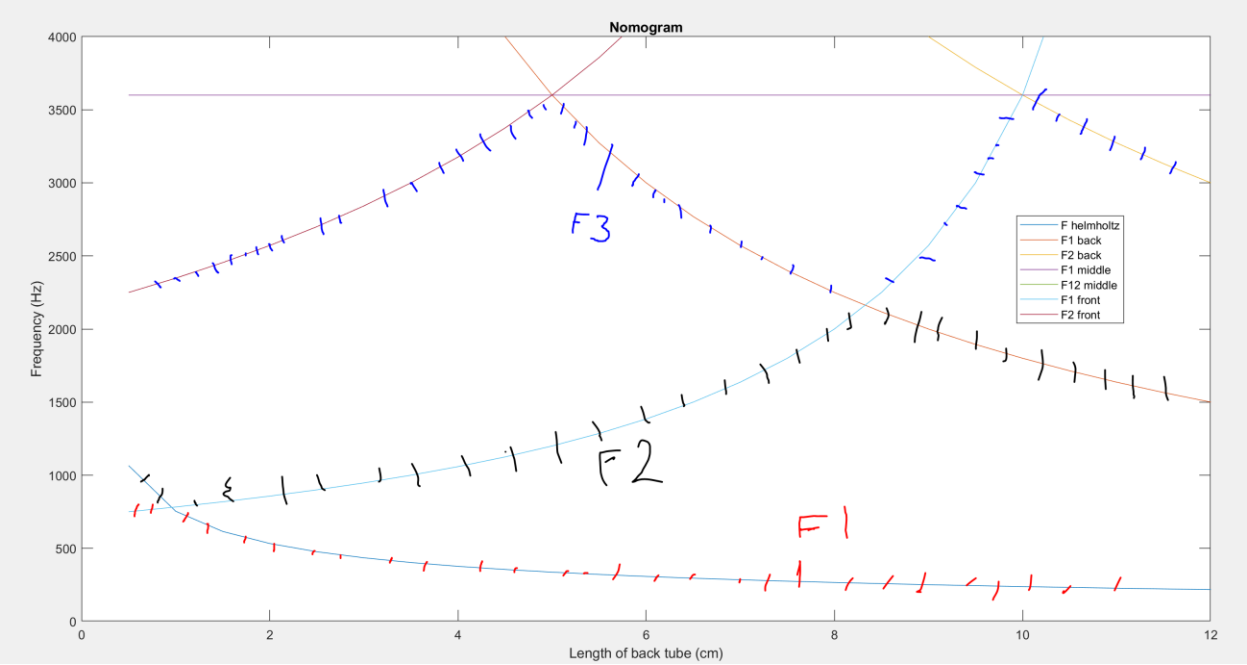
Helmholtz Resonator formed by back and middle tubes: $F_H = \frac{c}{2\pi}\sqrt{\frac{\frac{\pi d_2^2}{4}}{l_1 l_2 \frac{\pi d_1^2}{4}}}$

c	36000	cm/s									
d1	3.4										
d2	1										
	l_1	1_2	1_3		F_helmholtz	F1 back	F2 back	F1 middle	F2 middle	F1 front	F2 front
	0.5	5	12		1065.79508	36000	72000	3600	7200	750	2250
	1	5	11.5		753.6309283	18000	36000	3600	7200	782.6087	2347.826
	1.5	5	11		615.3370763	12000	24000	3600	7200	818.1818	2454.545
	2	5	10.5		532.8975399	9000	18000	3600	7200	857.1429	2571.429
	2.5	5	10		476.6380497	7200	14400	3600	7200	900	2700
	3	5	9.5		435.1090193	6000	12000	3600	7200	947.3684	2842.105
	3.5	5	9		402.8326757	5142.857	10285.71	3600	7200	1000	3000
	4	5	8.5		376.8154642	4500	9000	3600	7200	1058.824	3176.471
	4.5	5	8		355.2650266	4000	8000	3600	7200	1125	3375
	5	5	7.5		337.0339971	3600	7200	3600	7200	1200	3600
	5.5	5	7		321.3493076	3272.727	6545.455	3600	7200	1285.714	3857.143
	6	5	6.5		307.6685381	3000	6000	3600	7200	1384.615	4153.846
	6.5	5	6		295.59837	2769.231	5538.462	3600	7200	1500	4500
	7	5	5.5		284.8457167	2571.429	5142.857	3600	7200	1636.364	4909.091
	7.5	5	5		275.1871063	2400	4800	3600	7200	1800	5400
	8	5	4.5		266.44877	2250	4500	3600	7200	2000	6000

	8.5	5	4		258.4932759	2117.647	4235.294	3600	7200	2250	6750
	9	5	3.5		251.2103094	2000	4000	3600	7200	2571.429	7714.286
	9.5	5	3		244.5101604	1894.737	3789.474	3600	7200	3000	9000
	10	5	2.5		238.3190249	1800	3600	3600	7200	3600	10800
	10.5	5	2		232.5755538	1714.286	3428.571	3600	7200	4500	13500
	11	5	1.5		227.2282745	1636.364	3272.727	3600	7200	6000	18000
	11.5	5	1		222.2336366	1565.217	3130.435	3600	7200	9000	27000
	12	5	0.5		217.5545097	1500	3000	3600	7200	18000	54000



Zoom in to show system level formants F1, F2, and F3



3b. The overlapping values should be pushed apart by acoustic coupling

3c. Only the Helmholtz frequency changes. It becomes lower.

3d. Values should be any reasonable estimate around:

	F1	F2
a	850	1600
i	240	2400
u	250	600
e	400	2300
o	350	650

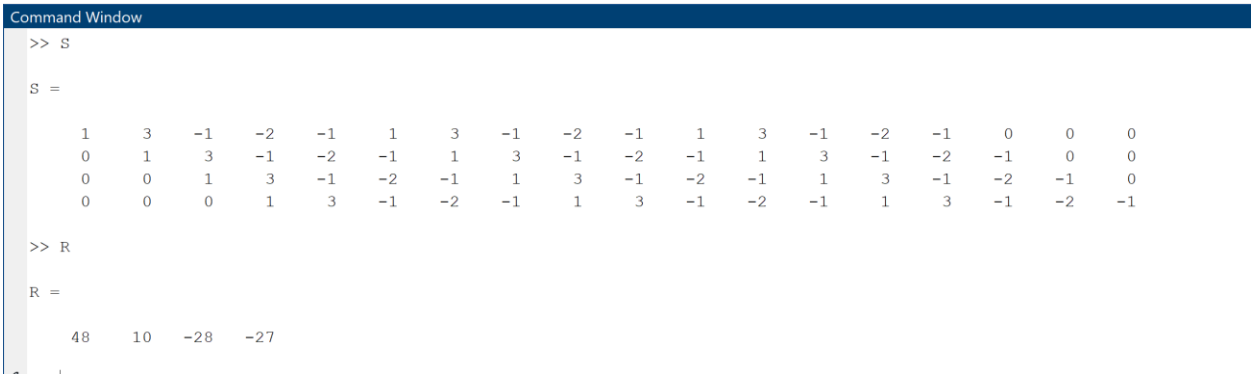
4.

```
%% part a

%calculate s[n] after being windowed with the
s_w=[1, 3, -1, -2, -1, 1, 3, -1, -2, -1, 1, 3, -1, -2, -1];

%calculate all relevant shifts of s^w [n].  Ie. s^w[n-1], s^w[n-2], and
s^w[n-3]
S = [s_w 0 0 0; 0 s_w 0 0; 0 0 s_w 0; 0 0 0 s_w];

%Then calculate the needed autocorrelation values
R0=sum(S(1,:).*S(1,:));
R1=sum(S(1,:).*S(2,:));
```

```
R2=sum(S(1,:).*S(3,:));
R3=sum(S(1,:).*S(4,:));

R = [R0, R1, R2, R3];
%or equivalently, R=conv(s_w,s_w(end:-1:1))

%% part b
%remember that Matlab indexing starts at 1, not 0
R_mat = [R(1) R(2) R(3);...
         R(2) R(1) R(2);...
         R(3) R(2) R(1)];

target_vec =[R(2) R(3) R(4)]';

LPCs = R_mat\target_vec;

%% part c

E_min = R(1) - sum(LPCs.*target_vec);
G=sqrt(E_min);

%% part d

denom = [1; -1*LPCs];
poles = roots(denom);
zplane(1,denom)

formant_locations = angle(poles)*8000/2/pi;
```

4a. We first find $s^w[n] = [1, 3, -1, -2, -1, 1, 3, -1, -2, -1, 1, 3, -1, -2, -1]$

Then plugging into the equation for R[i], we get



4b. We need to set up the 3rd order autocorrelation method LPC normal equations:

$$\begin{bmatrix} R[0] & R[1] & R[2] \\ R[1] & R[0] & R[1] \\ R[2] & R[1] & R[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} R[1] \\ R[2] \\ R[3] \end{bmatrix}$$

and solve for the vector of coefficients plugging in the autocorrelation values found in part a

```
Command Window
>> LPCs

LPCs =

    0.0750
   -0.5132
   -0.4118
```

4c. We calculate the prediction error as $E_{min} = R[0] - \sum_1^p a_p R[p]$

```
E_min =

   21.7618
```

The gain is then G = sqrt(E_min)

The transfer function is then:

$$V(Z) = \frac{4.665}{1 - (0.075)z^{-1} + 0.5132z^{-2} + 0.4118z^{-3}}$$

4d.

```
poles =

   0.2918 + 0.8513i
   0.2918 - 0.8513i
  -0.5085 + 0.0000i
```

Two poles are at complex conjugate pairs and one is purely real. This corresponds to two formants. With a sampling frequency of 8kHz, this means that the formants occur at

```
formant_locations =

   1.0e+03 *

    1.5796
   -1.5796
    4.0000
```

1580 Hz and 4000 Hz