EEN 540 Homework 3

Connor McCullough

1: List the acoustic features (source, articulatory, temporal and spectral) that distinguish the sounds in the following pairs.

a) The acoustic features in vowels are determined mainly by the position of the tounge. The source for both vowels is a periodic puff through the vocal folds, which is not dependent on pitch. In /o/, the tounge is at the back of the oral cavity and in /e/, it is at the front.

b) For both /h/ and /f/, the glottis is relaxed and not vibrated, and noise is generated due to turbulent airflow. In /f/, constriction is created by the teeth being placed against the lips. /h/ is classified as a whisper, where the constriction is being created at the glottis.

c) /Z/ is a voiced fricative, which has both a periodic and a noise source. /s/ is unvoiced, meaning that the vocal folds are relaxed and not vibrating, and instead there is only noise generated by constriction of the oral tract. /s/ has a noisy spectrum while /Z/ has both harmonic and noisy components. In the time domain, this corresponds to a noisy signal for /Z/, but noise superimposed over periodicity for /Z/.

d) /p/ and /b/ are both plosives, generated by a buildup of pressure behind complete constriction of the oral tract, or an impulsive source. The constriction can occur in either the front, back, or center of the oral tract, and there is no vibration in the vocal folds. /b/ is a voiced plosive, meaning the vocal folds vibrate while the oral tract is constricted, with propagation through the walls of the throat. The vocal folds continue vibrating after the pressure is released. With /p/, there is no sound generated while the oral cavity is closed. For /p/, there is a significant aspiration generated by the pressure leaving the oral cavity after the burst, with this occurring to less of a degree with /b/. For both plosives, there is an onset of a vowel sound about 40-50 mS after the burst. For /p/, there is silence in the waveform and spectrogram initially, followed by energy across a wide frequencies when the burst occurs, followed by a periodic signal with formants. The same occurs with /b/, except with periodicity occurring before the burst as well.

e) Both /J/ and /tS/ are non-stationary, meaning they transition between two different states over time. More specifically, they are both affricates, which transition from plosive to fricative. /tS/ is made of the plosive /t/ and the fricative /S/, while /J/ is made up of the plosive /d/ and the fricative /Z/.

2: Provide the phonetic transcriptions of the following phrases:

a) 'University of Miami': /y//u//n//I//v//r//s//I//t//i/ /A//v/
/m//aI//@//m//i/
b) 'We were away a year ago': /w//i/ /w//r/ /x//b//aU///t/ /x/
/y//i//r/ /x//g//o/
c) 'Thieves who rob friends deserve jail': /T//i//v//z/ /h//u/ /r//a//b/
/f//r//E//n//z/ /d//I///z//r//v/ /J//e//O//l/

3. Prove the reflection coefficient, |r|≤1.

A_1 and A_2 both represent areas, which must be real and positive. Because of this A_1-$A_2 \leq A_2$+A_1, meaning the numerator is always less than the denominator, so $|r| \leq 1$.

Problem 4

4.10 Figure 4.31 represents the magnitude of the discrete-time Fourier transform of a steady-state vowel segment which has been extracted using a rectangular window. The envelope of the spectral magnitude, $|H(\omega)|$, is sketched with a dashed line. Note that four formants are assumed, and that only the main lobe of the window Fourier transform is depicted.

(a) Suppose that the sampling rate is 12000 samples/s, set to meet the Nyquist rate. What is the first formant (F_1) in Hertz? How long is the rectangular window in milliseconds? How long is the window in time samples?

(b) Suppose $F_1 = 600$ Hz, and you are not given the sampling rate. What is the pitch in Hertz? What is the pitch period in milliseconds? What is the pitch period in time samples?

(c) Suppose you are told that the speech spectrum in Figure 4.31 corresponds to a uniform concatenated tube model. How many uniform tubes make up the model? Is it possible to estimate the cross-sectional areas of each tube with the given information? If not, explain why. If yes, derive estimates of the tube cross-sections. and given a sampling rate of 12000 samples/s,

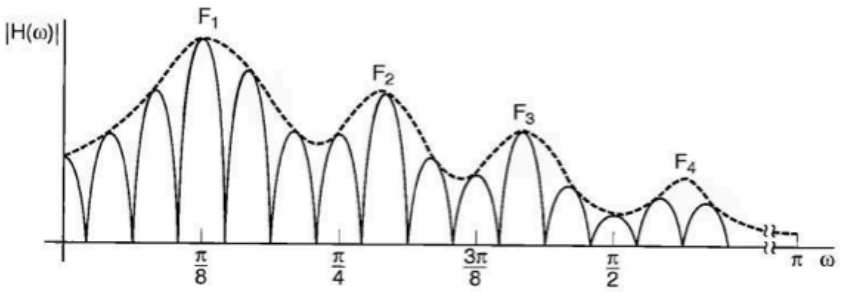


Figure 4.31 $|H(\omega)|$ of steady-state vowel segment.

(a) $F_1 = 750$ Hz

(b) F_f = fundamental frequency = $F_1/3 = 200$Hz

$T_f = 1/F_f = 5$ms

$F_s = 2*8*600 = 9600$ Hz

N = number of samples = $F_s/F_f = 48$ samples

(c) There needs to be twice as many tubes as formants, so there must be 8 tubes to make up this model. The cross sectional area of each tube can be measured based on the location of the formants in the frequency response. The formants give location of the poles. By multiplying out the entire numerator and denominator, the reflection coefficients can be obtained. From here, the reflection coefficients can be used to determine the cross sectional area of each tube.

Problem 5

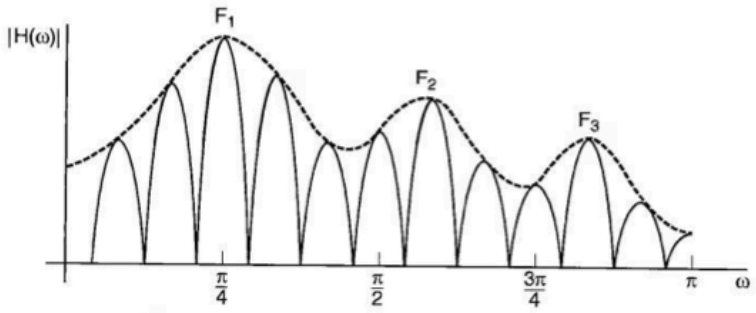


Figure 4.33 $|H(\omega)|$ of steady-state vowel segment.

4.14 Figure 4.33 represents the magnitude of the discrete-time Fourier transform of a steady-state vowel segment which has been extracted using a rectangular window. The envelope of the spectral magnitude, $|H(\omega)|$, is sketched with a dashed line. Note that three formants (resonances) are shown, and that only the main lobe of the window Fourier transform is depicted.

(a) Suppose the sampling rate is 6000 samples/s and was set to meet the Nyquist rate. What is the pitch period in milliseconds? How long is the rectangular window in milliseconds?

(b) If $F_1 = 750$ Hz and the vocal tract is considered to be a single acoustic tube, what is the length of the vocal tract? Assume zero pressure drop at the lips, an ideal volume velocity source, and speed of sound $c = 350$ m/s.

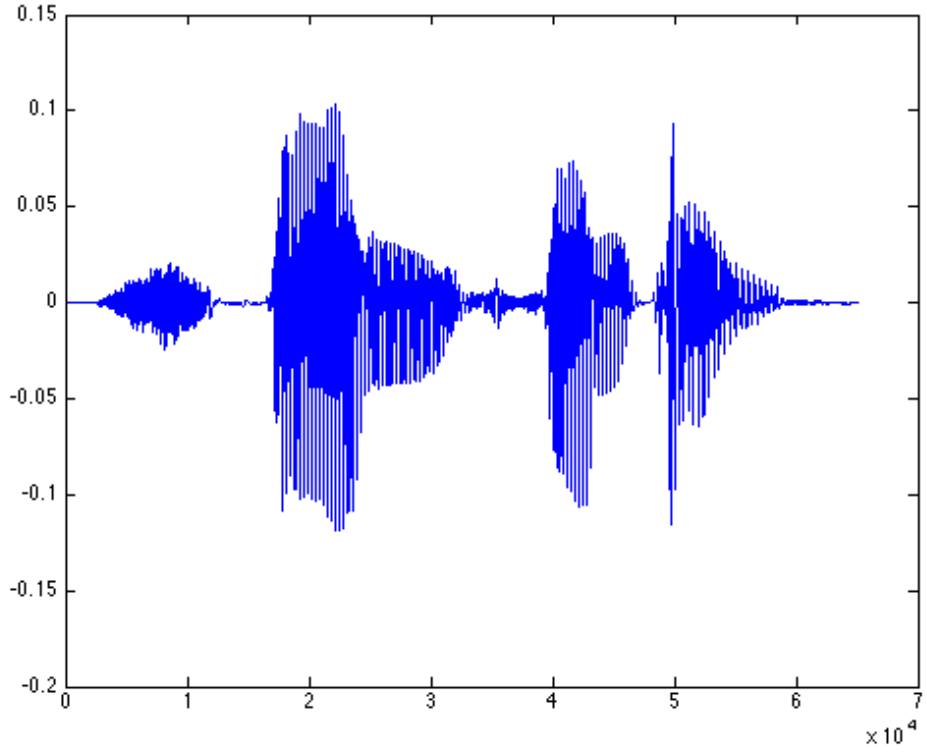
(c) If the length of the vocal tract were shortened, how would this affect the spacing of the window main lobes that make up the discrete-time magnitude spectrum of the signal? Explain your answer.

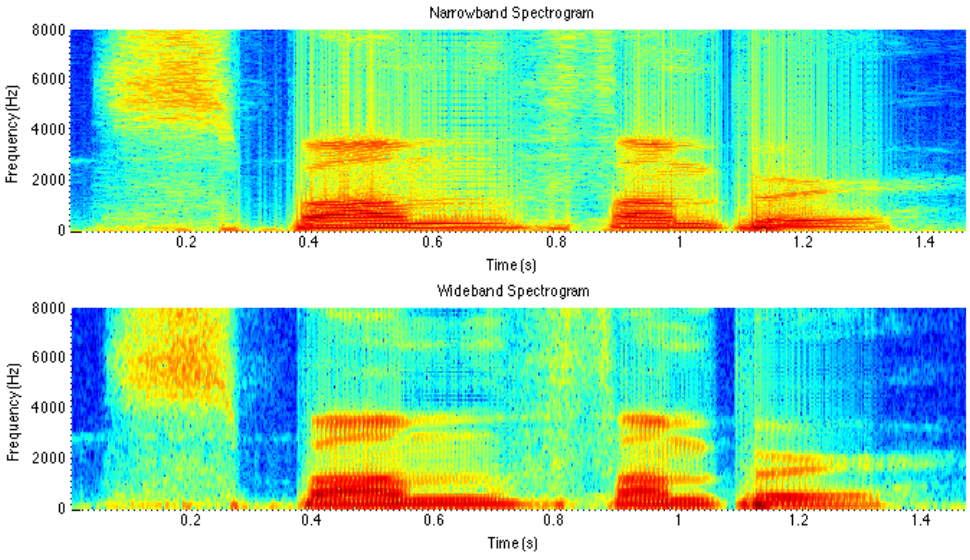
(d) Suppose that $H(\omega)$ represents the frequency response between the lip pressure and glottal volume velocity. How would the spectral magnitude change if there were no radiation load at the lips?

(a) $F_f = 6000/2/4/3 = 250$ Hz
$T_f = 1/250 = 4$ mS

(b) $F_{max} = F_1 * 4 = 750*4 = 3000$Hz
N = 3, C = 350
$L = CN/4F_{max}$
$L = (350*3)/(4*3000)$
L = 8.75 cm

Problem 6


0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
0
1
2
3
4
5
6
7
x 10^4


Narrowband Spectrogram
Frequency (Hz)
8000
6000
4000
2000
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Time (s)
Wideband Spectrogram
Frequency (Hz)
8000
6000
4000
2000
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Time (s)

Code:

```
[y,fs] = wavread('connorHW3.wav');
y = y(10000:75000,1);
plot(y);

T = 1/fs;
N = round(.025/T);
win = ones(N,1);

figure();
subplot(2,1,1);
spectrogram(y,win,round(0.97*N),[],fs,'yaxis');
title('Narrowband Spectrogram');
subplot(2,1,2);
spectrogram(y,hamming(N/4),round(0.9*N/4),[],fs,'yaxis');
title('Wideband Spectrogram');
name = [y,'spectrogram'];
print(gcf, '-djpeg',name);
```

Analysis:

Audio is read into MATLAB and the silence is removed, and only one channel is taken from the file. The time domain plot is done using the 'plot' function. The Narrowband spectrogram is created by generated by a rectangular window of fs/40. The Wideband spectrogram is generated by creating a Hamming window, with a smaller overlap size than in the narrowband spectrogram.

Vocal Tract Length:

$l = CN/4f_{max}$

C = 341 m/s

N = 5 formants observed in Spectrogram

f_{max} = 3800 (frequency of highest formant)

l = (341*5)/(4*3800) = 11.2 cm