Problem 1

Consider the variability of the vowel characteristics among speakers. If humans can create a phoneme from similar vocal tract shapes, but with different speaker identifiability, then the source and prosody must play a large role in speaker characteristics, as well as the vocal tract. A ventriloquist, for example, can keep his/her vocal tract in roughly the same shape (i.e. for a particular phoneme), yet mimic different voices. Offer an explanation in terms of source and prosodic characteristics.

Solution.
The Question asks to explain why vowels can sound quite different due to source and prosody variability even though vocal tract shapes might be similar.
The shape of the vocal tract decides the formants of the syllables/vowels, but the envelope of the spectrum is decided by several other factors given below, all leading to variability in the characterisation of syllables spoken.
Prosody is a suprasegmental property of speech which largely depends upon:

- Fundamental frequency (Pitch) - Pitch plays an immense role in stress and prominence of phonemes, which thereby shows the mood and salient features of speech that can't be portrayed by its meaning alone. Thus even though syllables are the same, more meaning is supplied by varying the pitch.
- Duration - Duration/Timing that a particular phoneme is stretched also shows the stress the speaker wants to provide to that part of speech.
- Intensity - Loudness of a phoneme also shows the stress and mood of the speaker.
- Spectral characteristics - The distribution of energy on different parts of audible frequency range.

Prosody therefore decides the tones and extra information carried by a syllable. Examples are, in Mandarin, four different tones can be used in each syllable, represented by four different pitch patterns. In European languages, prosodic features don't normally effect the identities of words, but provide useful additional information about what is being said.
Source depends upon the shape of vocal tract, but other factors include:

- Glottis - Slit-like opening between vocal cords, through which air flow makes folds vibrate, thus modulating the air flow (phonation), which is decides the $h[n]$ in source-system model.
- Vocal Folds - Build up of vocal folds oscillation takes place upon phonation, which is mainly determined by the mass and tension of folds.
- Lungs - Phonation is determined by air pressure in lungs, which decides $e[n]$.

■

Problem 2

A voiced fricative is generated with both a periodic and noise source. In a simplified model of a voiced fricative, the periodic signal component $u[n]$ is passed through a linear time-invariant vocal tract with impulse response $h[n]$. In the model of the noise source component of the voiced fricative, the vocal tract is constricted along the oral tract and air flows through the constriction, resulting in a turbulent airflow velocity source at the constriction that we denote by $q[n]$. In this simplified model, the glottal flow $u[n]$ modulates this noise function $q[n]$ (assumed white noise). The modulated noise then excites the front oral cavity that has impulse response $h_f[n]$. The results of two airflow sources (due to periodic glottal source and noise source) add, so that the complete output of the lips is given by,

$$x[n] = h[n] * u[n] + h_f[n] * (q[n] \cdot u[n])$$

Determine the Fourier transform of the response $x[n]$ for the voiced fricative model. Assuming you are given the vocal tract and front cavity responses $h[n]$ and $h_f[n]$, respectively. Propose a method for determining the source functions $u[n]$ and $q[n]$.

Solution.
$u[n]$ is the periodic signal component produced in source, $h[n]$ is the Time invariant IR of vocal tract, $h_f[n]$ is the time invariant IR of front oral cavity, $q[n]$ is a white noise function.

Fourier transform of complete output is

$$X(\omega) = H(\omega) \cdot U(\omega) + H_f(\omega) \cdot (Q(\omega) * U(\omega))$$

We're given $x[n]$ (and $X(\omega)$), $h[n]$, (and $H(\omega)$), $h_f[n]$ (and $H_f(\omega)$).

To obtain $u[n]$ and $q[n]$, We use the fact that $q[n]$ is white noise, i.e. its transform is a constant function in frequency domain. Also, $u[n]$ being a periodic signal has a finite support (finite non-zero locations in frequency domain).
Thus $U(\omega) * Q(\omega) = C \cdot Q(\omega) = K$, where C is some constant, and $Q(\omega)$ is a constant function.

$$X(\omega) = H(\omega) \cdot U(\omega) + H_f(\omega) \cdot C \cdot Q(\omega)$$
$$\implies U(\omega) = \frac{X(\omega) - K \cdot H_f(\omega)}{H(\omega)}$$

Also to obtain $Q(\omega)$,

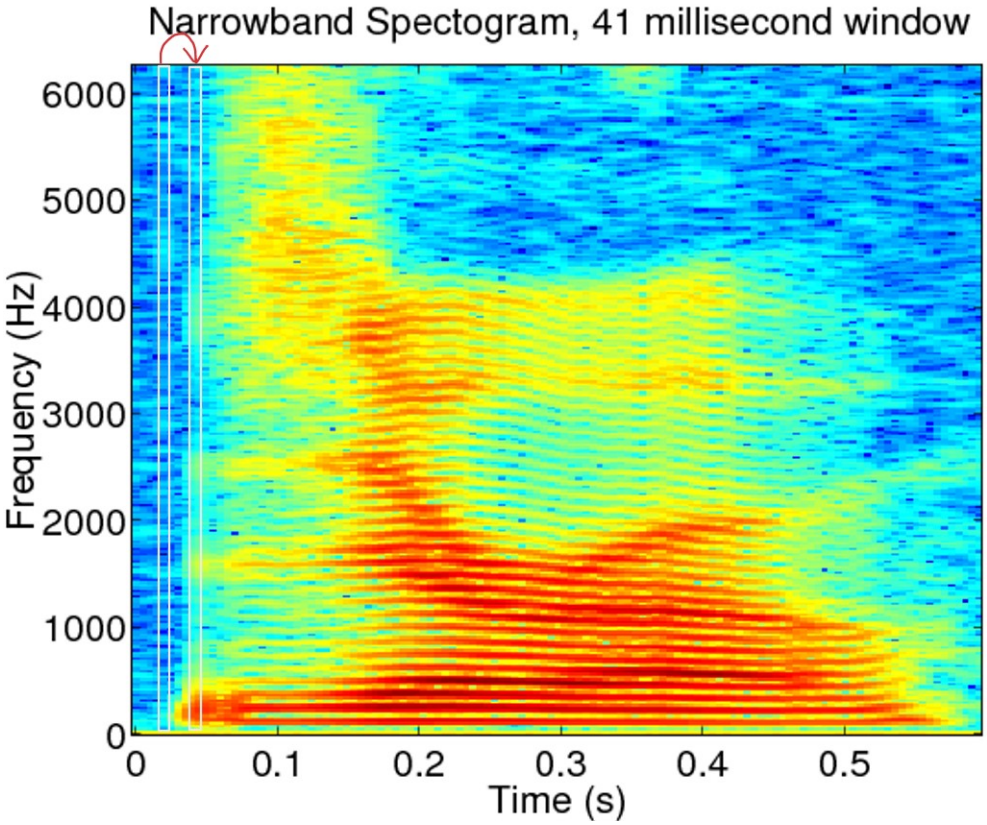
$$Q(\omega) = \mathcal{F}\{\frac{\mathcal{F}^{-1}\{K\}}{\mathcal{F}^{-1}\{\frac{X(\omega) - K \cdot H_f(\omega)}{H(\omega)}\}}\}$$

■

Problem 3

Based on the spectrogram, propose a method for measuring a speaker's rate of articulation. *Hint:* Consider the difference in spectral magnitudes in time, integrate this difference across frequency, and form a speaking rate metric.

Solution.
Rate of Articulation is a prosodic feature, defined as a measure of rate of speaking where all pauses are excluded from calculation. Articulation rates are generally measured as syllables per second.



In the above spectrogram, take a window's Fourier transform values as a vector (shown as white vertical box). Take the next shifted and adjacent window, and calculate their difference, and the magnitude of vector would define whether the syllable is the same or has changed.
Say window1 is a vector of value $[-0.012, -0.006, -0.023, -0.027, ...]$,
and window2 is $[0.034, 0.062, 0.0047, 0.024, ...]$,
thus difference being $[0.046, 0.068, 0.031, 0.089, ...]$.
Its magnitude would decide whether the current syllable has changed or not. Using this, we can calculate the number of times syllables have changed, and therefore the number of syllables throughout the speech recorded. Thus,

$$Articulation\ Rate = \frac{Total\ Number\ of\ Syllables}{Total\ Time\ of\ Speech}$$

■

Problem 4

Propose a simplified mathematical model of an unvoiced plosive, accounting for the burst after the opening of the oral tract closure and aspiration noise prior to the onset of voicing. Model the burst as an impulse $\delta[n]$ and aspiration as a noise sequence $q[n]$. Assume a linear time-varying oral tract.

Solution.
Unvoiced plosives involve a "burst", approximated as an impulse $\delta[n]$, that is generated at the release of the buildup of pressure due to complete closure of oral tract, during which no sound is radiated, no involvement of vibration in vocal cords.
This "burst" involves generation of turbulence over a very short time duration, which excites the oral cavity in front of the constriction. This aspiration, is then followed by onset of the following vowel/phoneme (not included in the fricative itself) 40-50 ms after the burst, during which $q[n]$ noise sequence can be assumed to pass through the front oral cavity.
For a simplified mathematical model, Assume the following:

$h[n, m]$: Impulse response of oral cavity that is excited due to buildup pressure and leads to the burst.

$h_v[n, m]$: Impulse response of Open Vocal folds, that leads to the turbulence and thus the Aspiration part of a fricative.

$\delta[n]$: Impulse that led to the "burst".

$q[n]$: Aspiration as a Noise sequence.

Thus the output response could be

$$x[n] = \sum_{m=-\infty}^{\infty} h[n, m] \cdot \delta[n - m] + \sum_{m=-\infty}^{\infty} h_v[n, m] \cdot q[n - m]$$

■

Problem 5

What differentiates the five unvoiced fricatives $|f|, |T|, |s|, |S|, |h|$? Why are fricatives lower in energy than vowels?

Solution.
Fricatives are produced by partial obstruction of airflow in the vocal tract, that leads to the flow getting turbulent, which is what differentiates fricatives from other phonemes. These fricative consonants differ in terms of the point of constriction in the vocal tract (i.e., place of articulation) – labiodental /f/; interdental /T/; alveolar /s/; and palatal /S/.

Energy required for vowels is higher than fricatives because of more airflow through the vocal tract occurs, while there is partial obstruction in the case of fricatives, leading to less air passage. ■

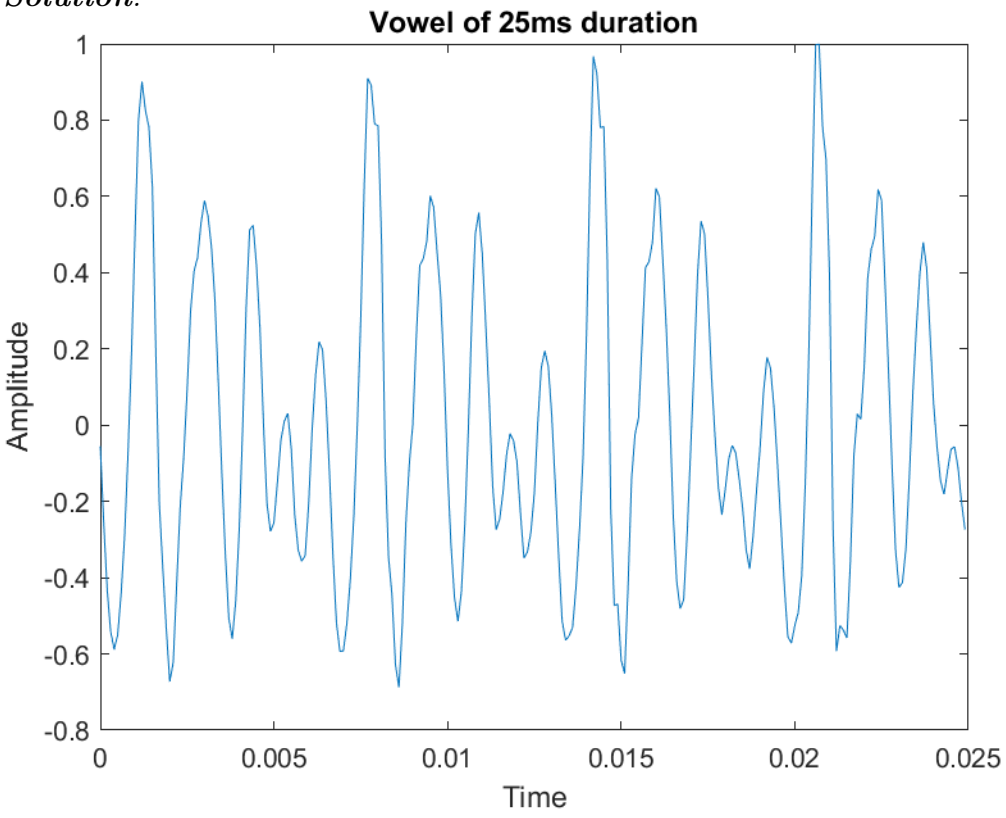
Problem 6

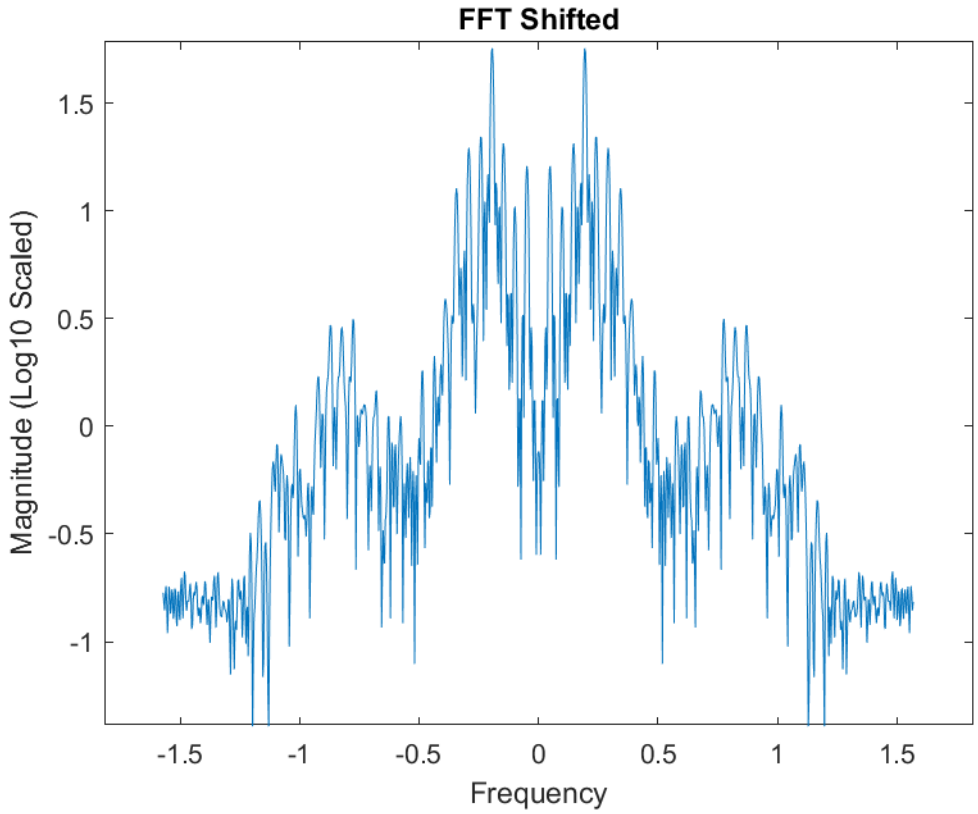
Programming Assignment: In this MATLAB exercise, you investigate some properties of a windowed speech waveform.

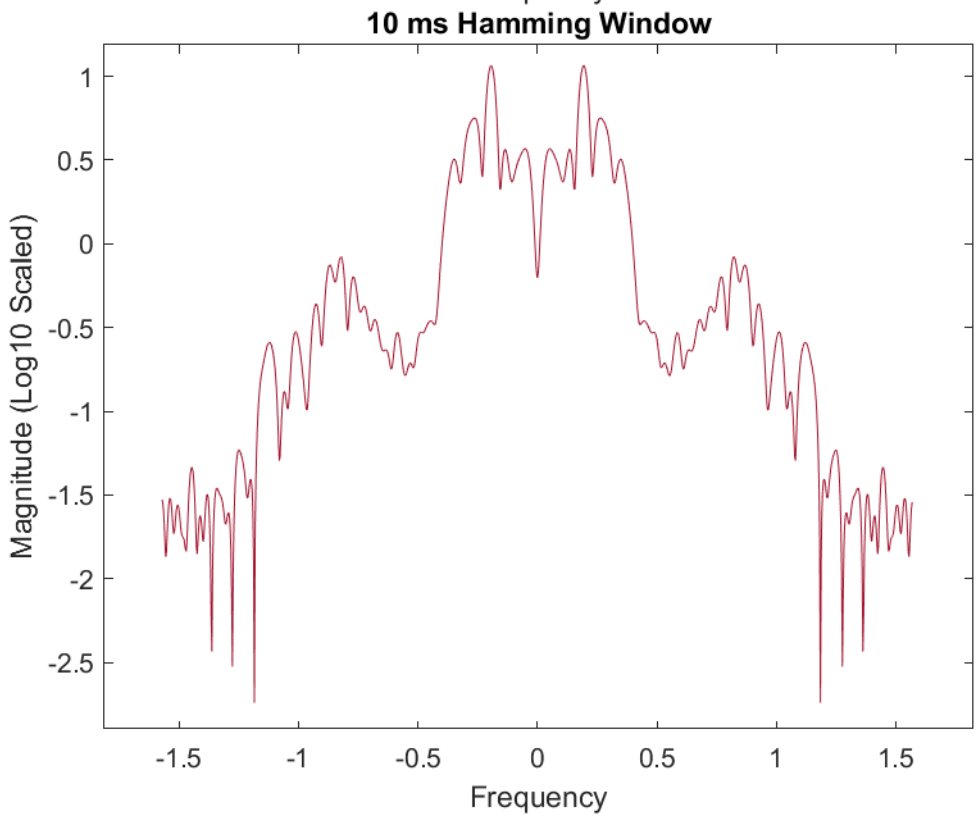
> Load the workspace *ex2Ml.mat* and plot the speech waveform labeled speech1_10k. This speech segment was taken from a vowel sound that is approximately periodic (sometimes referred to as "quasi-periodic"), is 25 ms in duration, and was sampled at $10,000$ samples/s. Plot the log-magnitude of the Fourier transform of the signal over the interval $[0, \pi]$, using a 1024-point FFT. The signal should be windowed with a Hamming window of two different durations, 25 ms and 10 ms, with the window placed, in each case, at the signal's center.

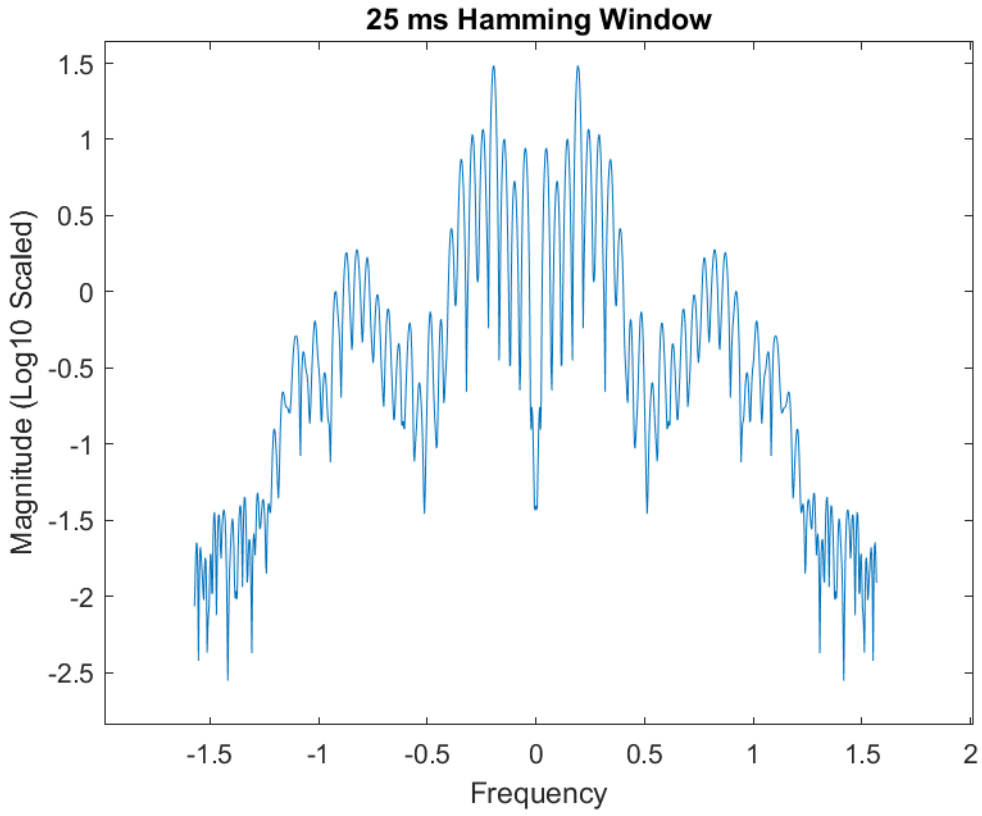
Show the log-magnitude plot for each duration. In doing this exercise, use MATLAB functions *fft.m* and *hamming.m*.

Solution.




FFT Shifted
Magnitude (Log10 Scaled)
Frequency
1.5
1
0.5
0
-0.5
-1
-1.5


10 ms Hamming Window
Magnitude (Log10 Scaled)
Frequency
1
0.5
0
-0.5
-1
-1.5
-2
-2.5



■

MATLAB Code

```
load('./ex2M1.mat');
fs = 10000;
range = 0:(1/fs):(0.025-1/fs);
figure;
plot(range, speech1_10k);
title('Vowel of 25ms duration');
xlabel('Time');
ylabel('Amplitude');

n = 1024;
Y = fft(speech1_10k, n);
Y_abs = abs(Y);
freq_range = 0:pi/n:pi-pi/n;
s_freq_range = -pi/2:pi/n:pi/2-pi/n;
s_Y = fftshift(Y);
figure;
plot(s_freq_range, log10(abs(s_Y)));
title('FFT Shifted');
xlabel('Frequency');
```

```
ylabel('Magnitude (Log10 Scaled)');

ham_10 = zeros(250);
ham_10(125-50:125+49) = hamming(100, "symmetric");
window_10 = (speech1_10k' .* ham_10);
fft_win10 = fftshift(fft(window_10, 1024));
figure;
plot(s_freq_range, log10(abs(fft_win10)));
title('10 ms Hamming Window');
xlabel('Frequency');
ylabel('Magnitude (Log10 Scaled)');

ham_25 = hamming(250, "symmetric");
window_25 = (speech1_10k' .* ham_25);
fft_win25 = fftshift(fft(window_25, 1024));
figure;
plot(s_freq_range, log10(abs(fft_win25)));
title('25 ms Hamming Window');
xlabel('Frequency');
ylabel('Magnitude (Log10 Scaled)');
```