1 Vowel Spectrum and Pitch Estimation

Using Matlab a script was written to read in the "vowel.wav" file. [Fig 1] In order to compute the 1024 FFT of the vowel, we need to pad the data with zeros since it contains fewer than 1024 samples. To calculate the FFT properly, we also window the function [Figure 3] with a hamming window [Figure 2] to prevent generation of frequency artifacts due to a rectangular window.

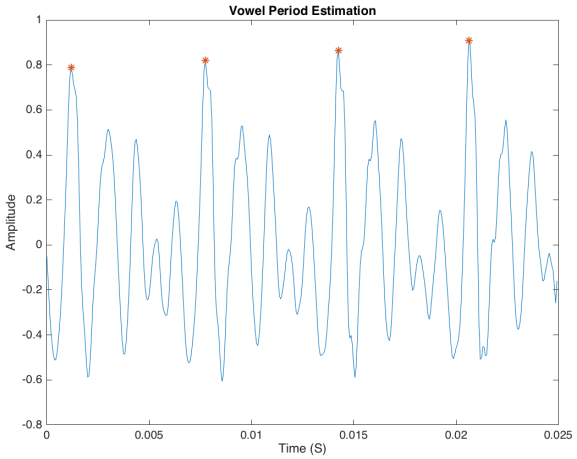

Figure 1: The original vowel sample imported into Matlab.

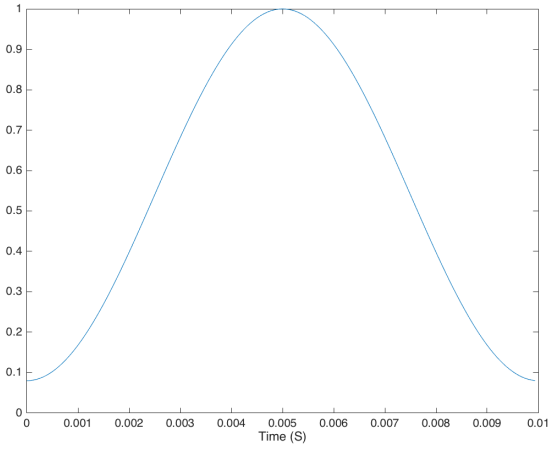

Figure 2: Hamming window with width of 10 ms.

Given this we can generate a dB spectrum of the windowed vowel sample using the Matlab FFT function. [Figs 4, 5]

1.1 Pitch Estimation

As can be seen in figures 1, 4, and 5, the relevant peaks for pitch estimation were found using the findpeaks function in matlab. The parameters for findpeaks varied between each instance since the have very different

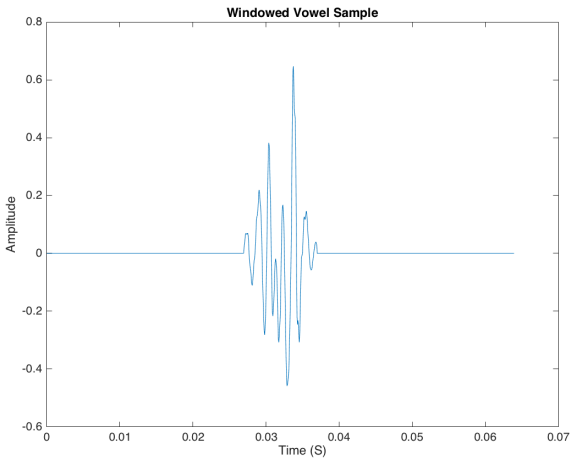

Figure 3: Vowel sample after windowing with a hamming window of 25 ms.

Source	Frequency
Waveform	154.34
10ms Window Spectrum	156.25Hz
25ms Window Spectrum	156.25Hz

peak prominences and frequency. From the time domain signal, it is obvious from observation that there are 4 periods of the waveform present in the signal. Additionally, the 25ms window spectrum was much easier to estimate the fundamental period, since there are much more pronounced harmonics. Since we know what we're looking for, the findpeaks function was tweaked manually, and this will only work for the specific case here. In order to automate the process, we would need to look more closely at the spacing of the peaks found by findpeaks. For example by specifying a minimum peak spacing, we can filter out harmonics that are too high in frequency to be glottal harmonics. In the frequency spectrum the base pitch can be estimated as the first peak, or the average distance between peaks satisfying a certain spacing requirment. Table 1.1 summarizes the results of the pitch estimation.

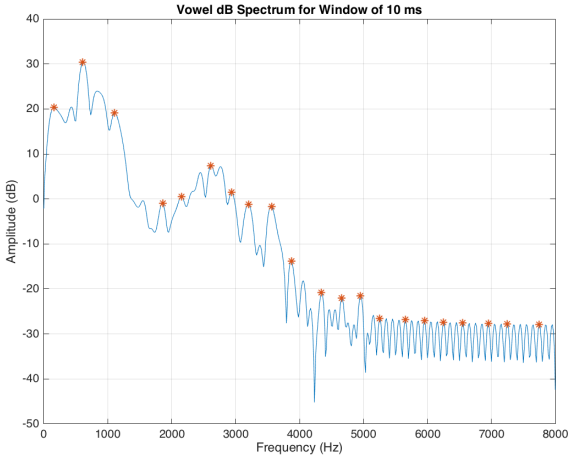

Figure 4: Spectrogram of vowel with a 10 ms hamming window.

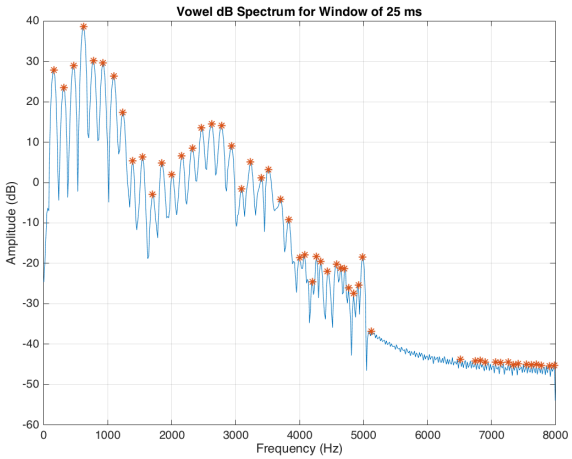

Figure 5: Spectrogram of vowel with a 25 ms hamming window. Notice the increased number of harmonics from the glottal impulses.

2 Vowel Synthesis

The vocal system can be approximated by the time convolution of a glottal wave train and an LTI filter. For our synthesis we assume that the filter whose frequency response can be described as an all pole filter described by Equation (1). This is a reasonable approximation for the synthesis of vowels since they are typically long in duration with stable formants, and the vocal tract shape has certain resonances without significant anti-resonances.

$$H(z) = \prod_{k=1}^{N} \frac{A}{(1 - c_k z^{-1})(1 - c_k^* z^{-1})} \tag{1}$$

In order to simulate a glottal pulse, we use the rosenberg glottal model which is one of the simpler methods to generate the glottal excitation: [3], [2]

$$u_g(t) = \begin{cases} \frac{\alpha}{2}(1 - cos(\frac{\pi t}{T_P})) & , 0 \leq t \leq T_P \\ \alpha cos(\frac{\pi(t - T_P)}{2T_N}) & , T_P < t < T_P + T_N \\ 0 & , \text{Rest phase (glottal closure)} \end{cases}$$

Where T_P is the time to maximal glottal amplitude (rise time), and T_N is time from maximal glottal amplitude to closure (fall time).

Instead of inventing the code from scratch, the a pre-existing matlab function was used to generate and sample this waveform. The rosenberg papers indicate that the duty cycle of the glottal pulse is slightly less than half, so we use a 45% duty cycle for the glottal wave. [Fig 6]. In order to generate the proper frequency of the pulse train, we build a vector of 'delta' functions in matlab (ones) and convolve with the single glottal pulse to generate our excitation waveform. A subset of the entire wave train is shown in Figure 7. Using the

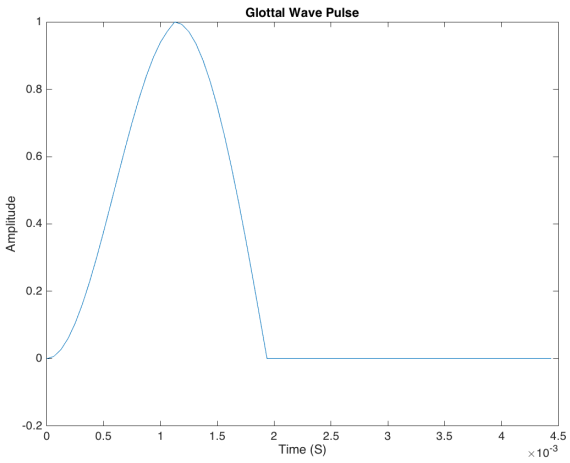

Figure 6: A single glottal pulse as generated by the Rosenberg model.

table from the Peterson and Barney '52 paper [[1]], the center frequencies for the female formants were used in the vocal model. In order to generate variability in the output vowels, we use a randomized collection of points within each vowel region. To do this, each vowel region was approximated with an ellipse with a center as defined by the table, and a rotation and sizing appropriate for the region being covered. To aid in the region selection process, the ellipses were plotted over Figure 8 As can be seen from Figure 8 the vowel regions do not necessarily match up well with the average vowel regions. However the vowels were distinctive enough that this was considered a non-issue. A 2D uniform distribution with dimensions of the major and minor ellipse axes was mapped to the proper ellipse shape, and then rotated and translated into the proper

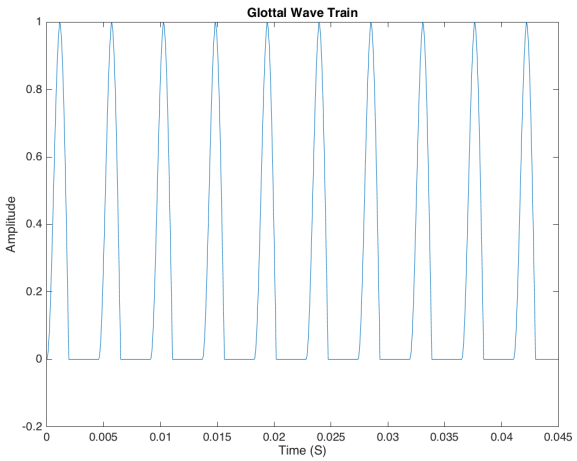

Figure 7: Glottal pulse train with ten pulses.

coordinate range using a rotation matrix. [Figure 9] The number of random points generated per vowel can be specified by the user, as well as the specific vowel for which formants are to be generated. Having satisfied the requirement for variability in the formant frequencies we generate the filter roots by computing the scaled frequency and its complex conjugate in the Z domain. This is necessary to fully describe the roots of the filter as defined in Equation (1). When generating the roots of the filter we also scale the values by 0.9 to avoid placing poles on the unit circle. Placing the poles on the unit circle would cause the filter to be unstable since there will be infinite gain ath that frequency. The roots are then expanded into filter coefficients using the poly function. Using the freqz function we can evaluate a sample set of filters generated by the matlab code. [Figure 10]

Thus having generated the two necessary elements of the system, the glottal pulse train and the filter, finally we can generate a vowel using a filtering function in matlab. The fuction filtfilt was chosen to eliminate any phase distortions that may arrise from the filter implementation. Filtfilt is phase invariant since it filters the input signal twice, once in its original orientation, and then the time reversed output of the original filtering operation. This allows us to remove the phase distortions caused by the filter which is desirable since excessive or discontinuous phase variations in speech signals can destroy the information in the waveform.

Using the matlab code written, each vowel sound was successfully synthesized using randomized formant combinations. As can be seen in figures 13, and 14 the random variations in the formant frequencies was able to generate noticable variation in the output. Listening to the vowel sounds it was also confirmed aurally that each vowel had differences that were identifiable. After listening to the sound of the formants and experimenting with adding a third formant, the third center formant frequency for each vowel was added to the filter generation step. No variability was introduced to this value since the Peterson and Barney paper did not include such information. This significantly increased sound quality and made the vowels more easily distinguishable and more natural sounding. Obviously the more information used to generate the signal the more realistic the synthesized waveform will sound. Another easy improvement would be to slightly vary the glottal waveform parameters, and add noise to the wave train to make the signal more realistic. In addition, the delta functions used to generate the final wave train should also have time variability to simulate shimmer and natural fluctuations in pitch. However, adding this information will increase the complexity of the signal construction and how much space is required to save the generated waveforms since higher formant frequencies require higher sampling rates.

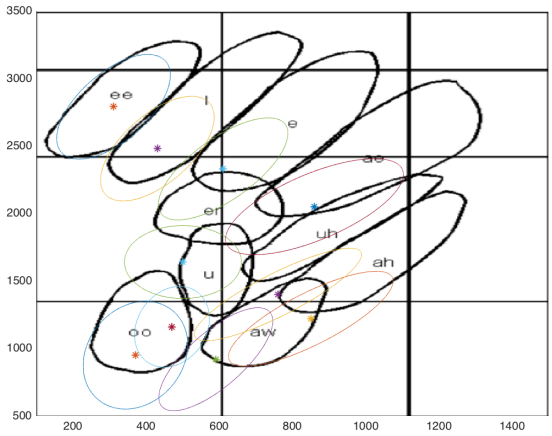

Figure 8: Vowel regions with female vowel frequency centers. As can be seen, the average vowel regions and female vowel regions do not necessarily overlap.

References

[1] Gordon E. Peterson and Harold L. Barney. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2), 1952.

[2] A. E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *The Journal of the Acoustical Society of America*, 49(2B), 1971.

[3] Yen-Liang Shue, Gang Chen, and Abeer Alwan. On the interdependencies between voice quality, glottal gaps, and voice-source related acoustic measures. In *INTERSPEECH*, pages 34–37, 2010.

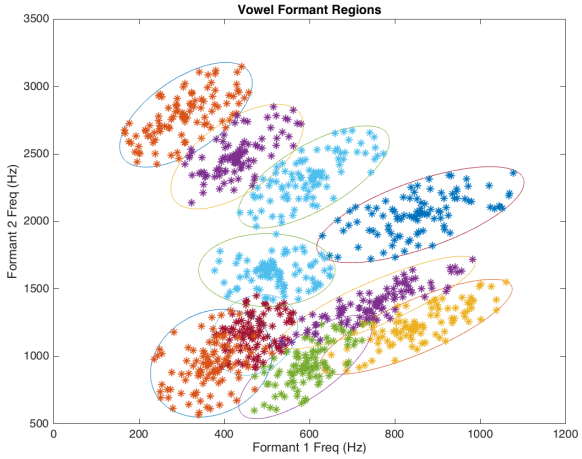

Figure 9: This figure shows each of the vowel regions with 100 randomly generated points within each region.

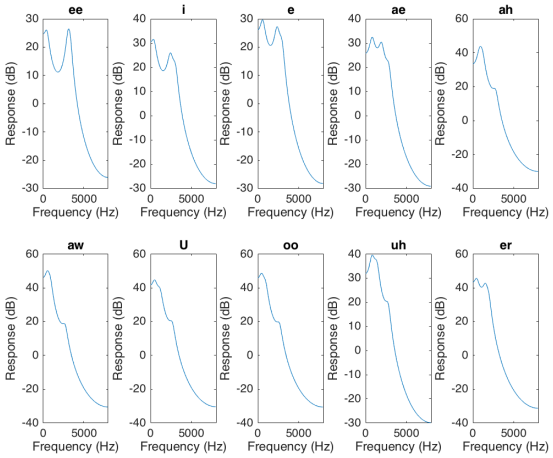

Figure 10: Frequency responses in dB a sample of each vowel.

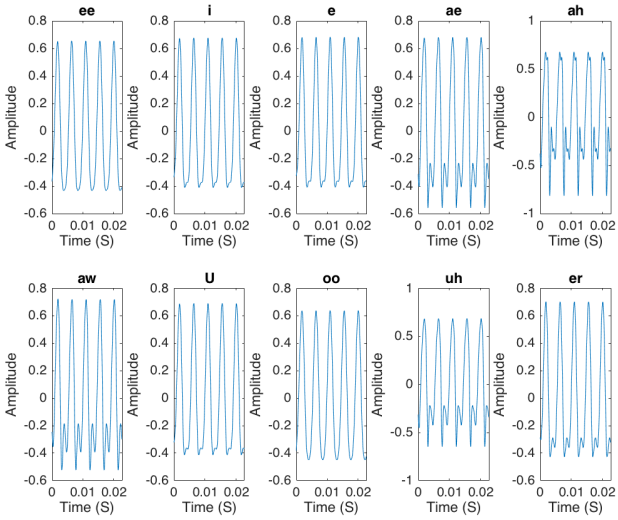

Figure 11: Time domain representation of our generated vowels.

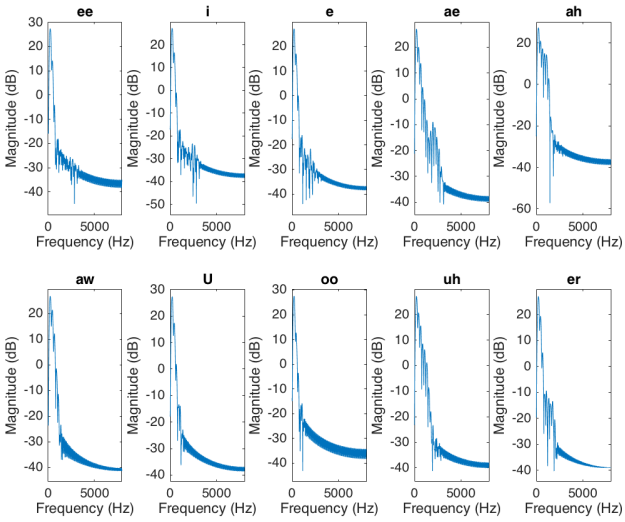

Figure 12: Frequency domain representation of generated vowels.

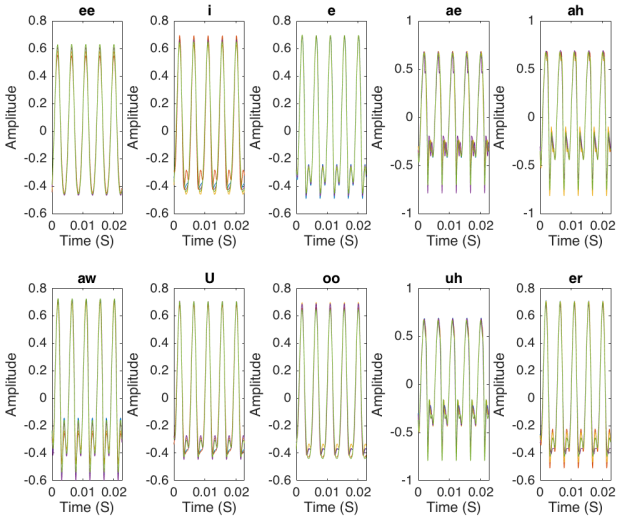

Figure 13: Time domain representation of five variations of the same vowel.

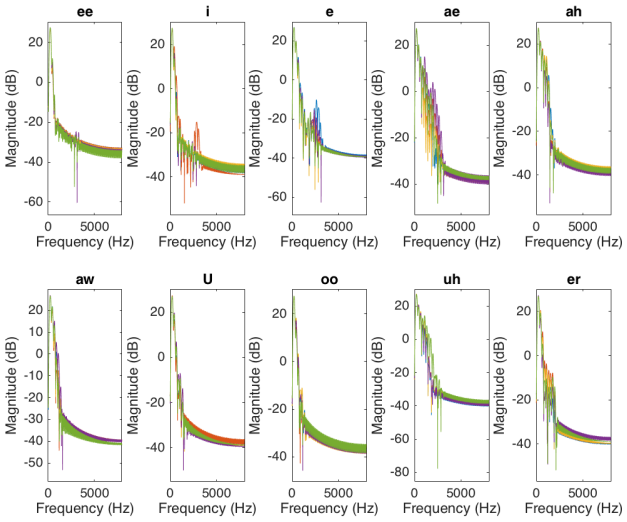

Figure 14: Frequency domain representation of five variations of the same vowel.