EE679 Speech Processing - Assignment 1

Shreya Laddha - 180070054

2/9/2021

1 Question 1

Given the following specification for a single-formant resonator, obtain the transfer function of the filter H(z) from the relation between resonance frequency / bandwidth, and the pole angle /radius. Plot filter magnitude response (dB magnitude versus frequency) and impulse response.
F1 (formant) = 900 Hz
B1(bandwidth) = 200 Hz
Fs (sampling freq) = 16 kHz

The vocal tract transfer function for a single formant resonator can be specified in terms of formant frequency, bandwidth and sampling frequency as follows -

$$H(z) = \frac{1}{1 - 2rcos\theta z^{-1} + r^2 z^{-2}}$$

$$r = e^{-\pi B_i T}$$

$$\theta = 2\pi F_i T$$

$$T = \frac{1}{F_s}$$

We can also write it as

$$H(z) = \frac{z^2}{z^2 - 2rcos\theta z + r^2}$$

which gives zeroes at [0,0] and poles at $[re^{j\theta}, re^{-j\theta}]$. We can obtain frequency response using this.

Impulse response can be obtained by converting H(z) to the difference equation. Consider,

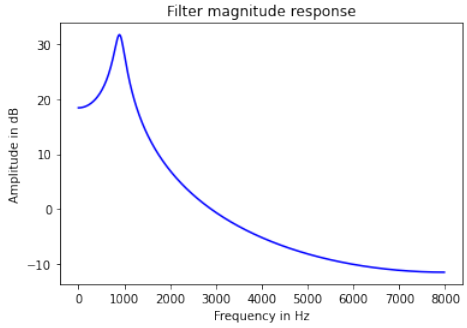
$$H(z) = \frac{b0}{a0 + a1 * z^{-1} + a2 * z^{-2}}$$

Now, x[n] is the input Impulse signal and y[n] is the Impulse Response

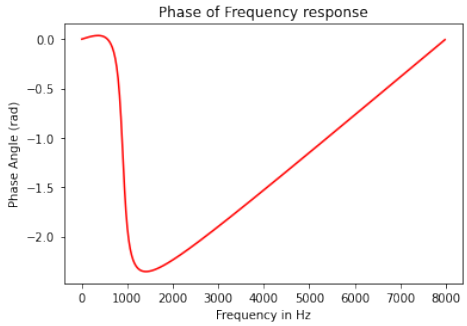
$$b0 * x[n] = a0 * y[n] + a1 * y[n-1] + a2 * y[n-2]$$

Since b0=1, a0 = 1, a1 = -2rcosθ, a2 = r^2 in our case, we can obtain y[n] as follows -

$$y[n] = x[n] - a1 * y[n-1] - a2 * y[n-2]$$



(a) Magnitude response



(b) Phase response

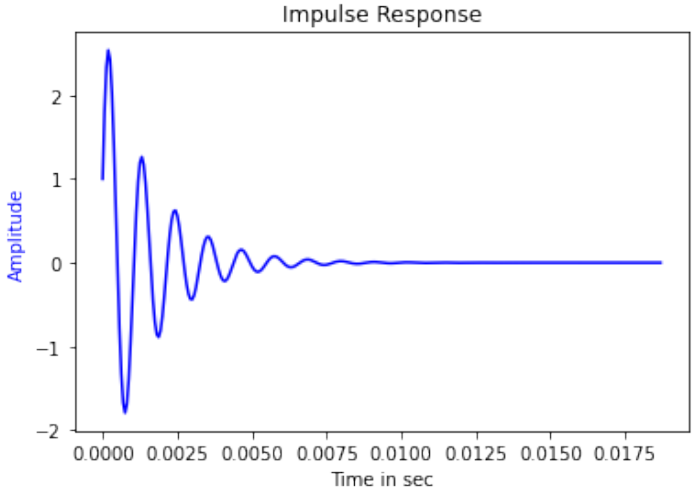


Figure 2: Impulse response

2 Question 2

Excite the above resonator ("filter") with a periodic source excitation of F0 = 140 Hz. You can approximate the source signal by narrow-triangular pulse train. Compute the output of the source-filter system over the duration of 0.5 second using the difference equation implementation of the LTI system. Plot the time domain waveform over a few pitch periods so that you can observe waveform characteristics. Play out the 0.5 sec duration sound and comment on the sound quality

Total no of samples = duration*F0 = 0.5*16k = 8000

The input waveform was approximated as a impulse train of 8000 samples, with amplitudes at multiples of int(floor(Fs/F0)) being 1. If we were to see in continuous range, the impulse will be a narrow-triangular pulse with having amplitude 1 at just one point and with symmetric rise and fall arms. Hence, it suffices to make this approximation.

Using the same formula as in previous question for difference equation implementation of the LTI system, we obtain results as follows -

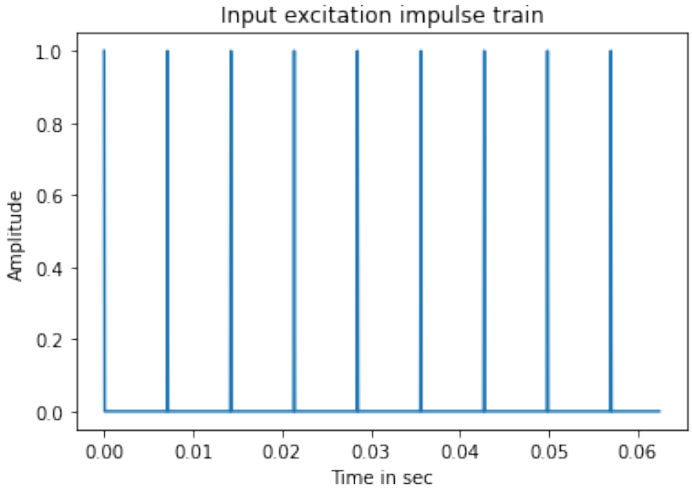
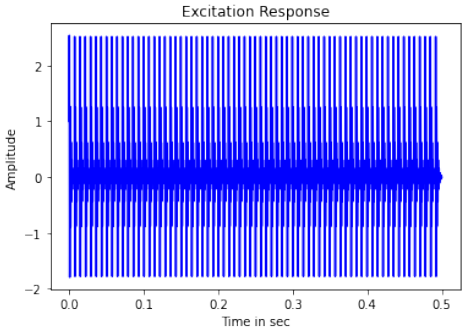
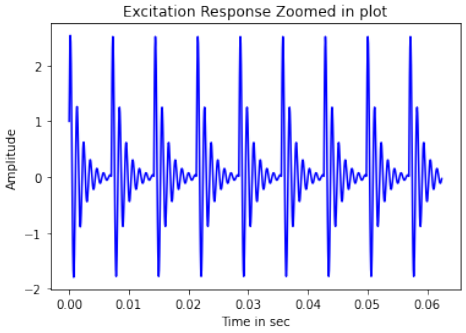


Figure 3: Excitation signal



(a) Output response for the impulse train



(b) Zoomed excitation response

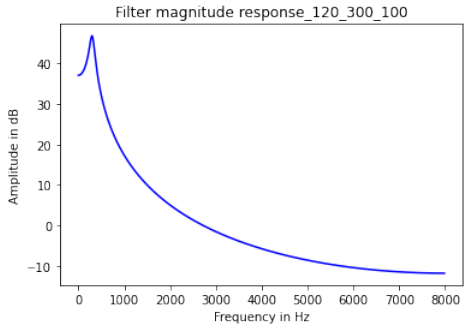
Observations - The audio heard was a constant low pitched sound and was a little rough in quality. It is probably an approximation of a vowel but the quality is very poor and is noisy.

The periodic excitation response is basically the impulse response repeating at every period of impulse, much like how convolution works of the input and the impulse response.

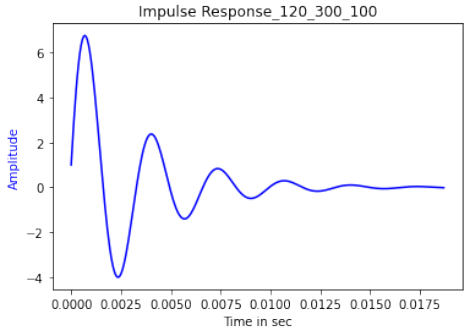
3 Question 3

Vary the parameters as indicated below; plot and comment on the differences in waveform and in sound quality for the different parameter combinations.

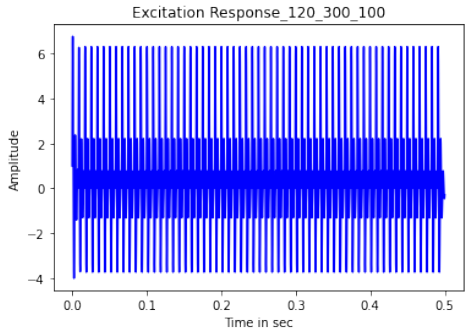
(a) F0 = 120 Hz, F1 = 300 Hz, B1 = 100 Hz



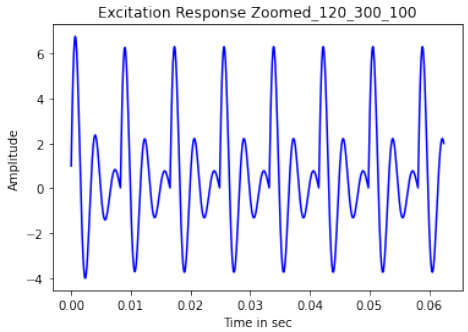
(a) Magnitude response



(b) Impulse response

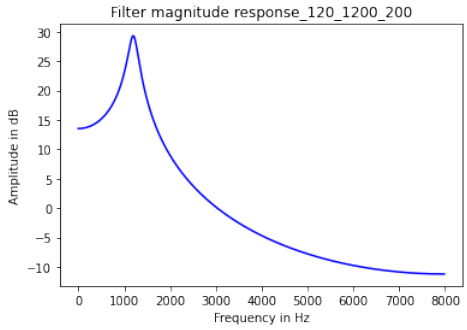


(a) Output response for the impulse train

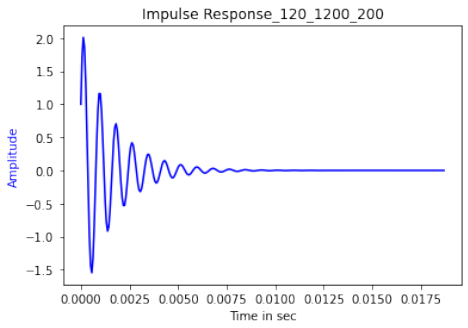


(b) Zoomed excitation response

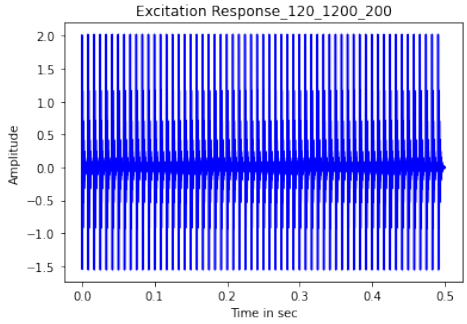
(b) F0 = 120 Hz, F1=1200 Hz, B1 = 200 Hz



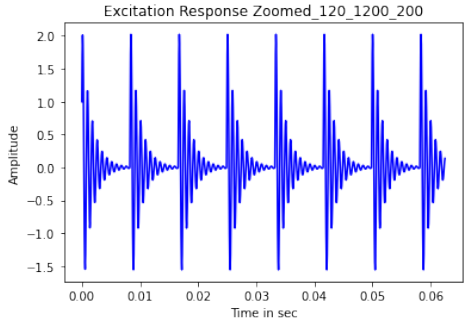
(a) Magnitude response



(b) Impulse response

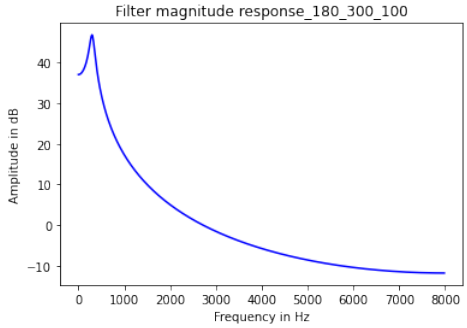


(a) Output response for the impulse train

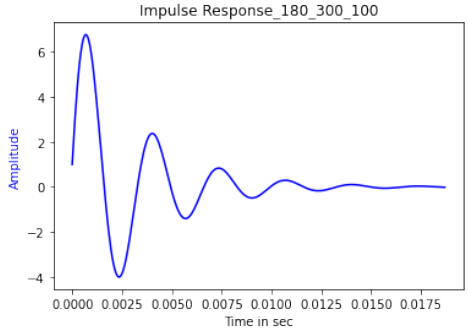


(b) Zoomed excitation response

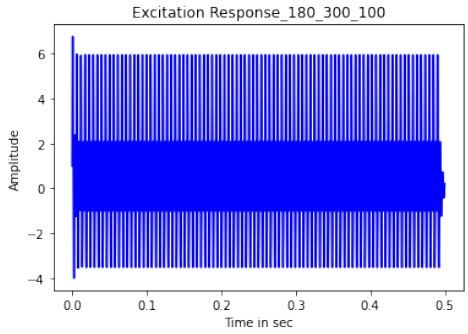
(c) F0 = 180 Hz, F1 = 300 Hz, B1 = 100 Hz



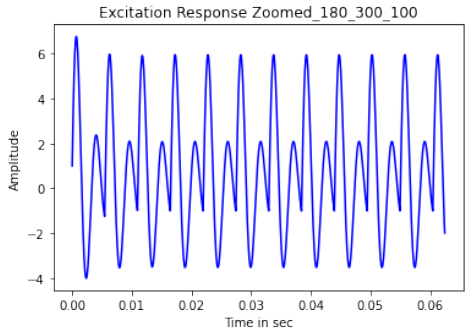
(a) Magnitude response



(b) Impulse response



(a) Output response for the impulse train



(b) Zoomed excitation response

Observations

- Since bandwidth in (b) is more, the decay rate of the impulse response waveform is more as compared to other two and hence there is far lesser interference in (b) as compared to the other 2 parts
- The 1st and the 3rd audios appear to sound alike whereas the 2nd waveform is totally distinct from the other two. The second one sounds like 'aa' sound and the 1st sounds like /u/.

- The 1st audio appears to be a little rougher as compared to the 3rd audio. Pitch is higher for 3rd one which is expected due to increase in F0. They sound alike since the formant frequency and the bandwidth do not change which are the only parameters of vocal tract and responsible for articulation. Only the glottal vibrations are changing. This is also reflected in their waveforms.
- The waveforms of 1st and 2nd look different as here the formant frequency changes and hence the sound also changes. The second waveform shows a sharp decline to the pulse locally which is not exhibited by the 1st waveform which rather has a more or less smooth waveform. In the wavfile, this shows up as abruptness in sounds vs a smooth sound. We can also make out from audios that the pitch of these two sounds are very same as both the F0 are same.

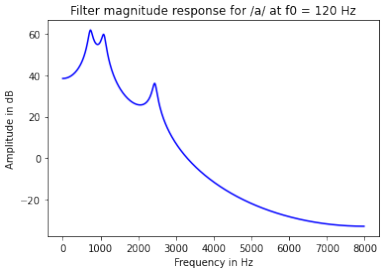
4 Question 4

In place of the simple single-resonance signal, synthesize the following more realistic vowel sounds at two distinct pitches (F0 = 120 Hz, F0 = 220 Hz). Keep the bandwidths constant at 100 Hz for all formants. Duration of sound: 0.5 sec. Comment on the sound quality across the different sounds. Plot a few periods of any 2 examples.
Vowel F1, F2, F3
/a/ 730, 1090, 2440
/i/ 270, 2290, 3010
/u/ 300, 870, 2240

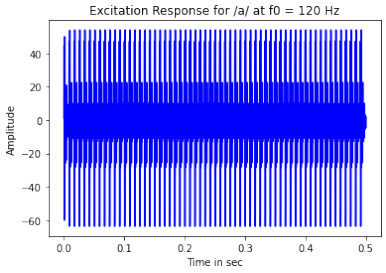
With three formant frequencies, we can consider the whole system as a cascade of three single formant resonators. So, we can calculate the poles and the zeros for the individual single-formant resonators. The system transfer function will be a product of the three second order transfer functions, thus the system transfer function will be having six poles. Magnitude response is plotted in every case.

For calculating output response, we can apply the difference equations to each of the three cascade systems in the order - F1,F2,F3. This can be clearly seen in the code. So, given x as input to system with F1 formant, we get a response state1. Next we give state1 as a input to system with F2 formant, we get a response state2. Finally we give state2 as input to system with F3 formant, we finally get the output of the entire system.

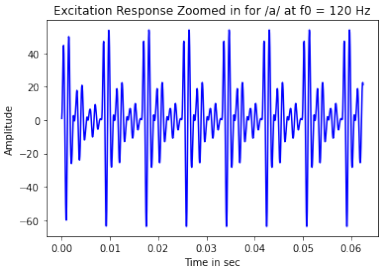
(a) **F0 = 120 for /a/**



(a) Magnitude response

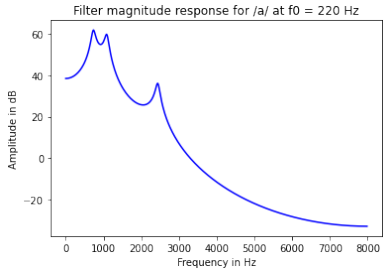


(b) Output response for pulse train

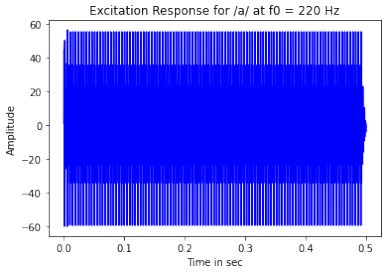


(c) Zoomed excitation response

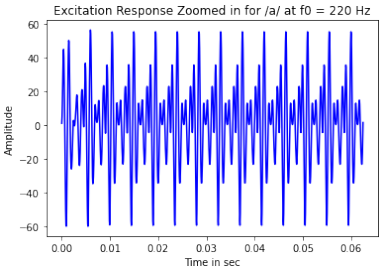
(b) **F0 = 220 for /a/**



(a) Magnitude response

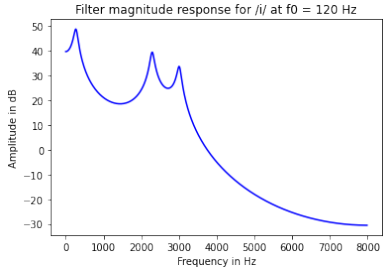


(b) Output response for pulse train

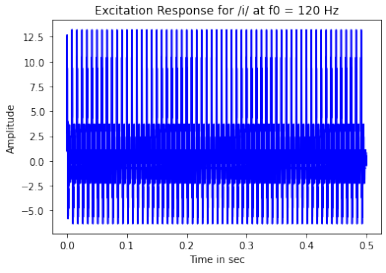


(c) Zoomed excitation response

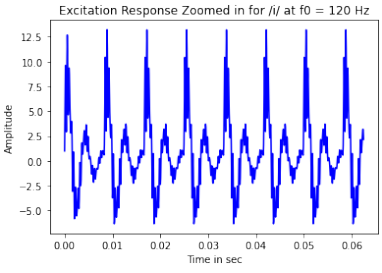
(c) **F0 = 120 for /i/**



(a) Magnitude response

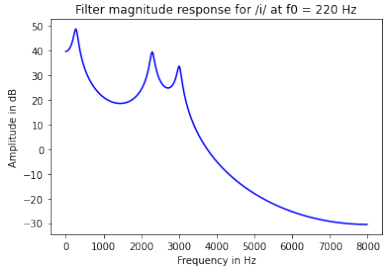


(b) Output response for pulse train

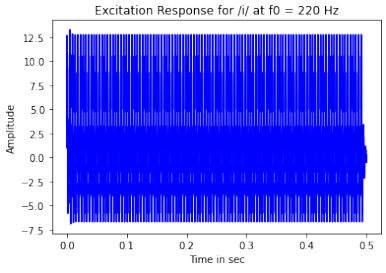


(c) Zoomed excitation response

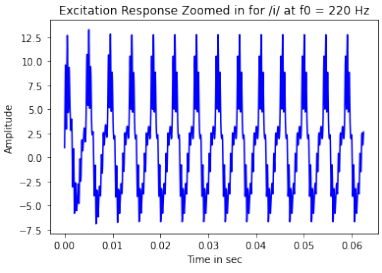
(d) **F0 = 220 for /i/**



(a) Magnitude response

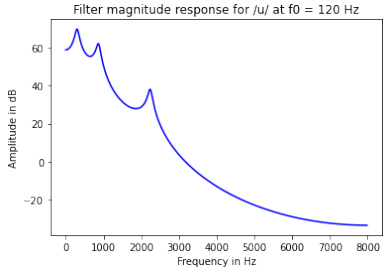


(b) Output response for pulse train

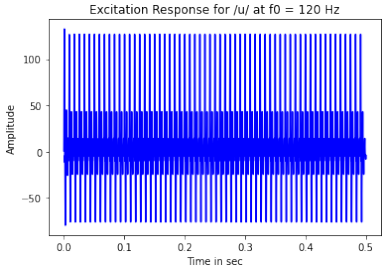


(c) Zoomed excitation response

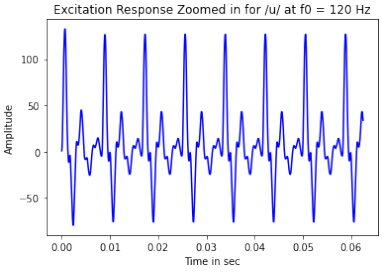
(e) **F0 = 120 for /u/**



(a) Magnitude response

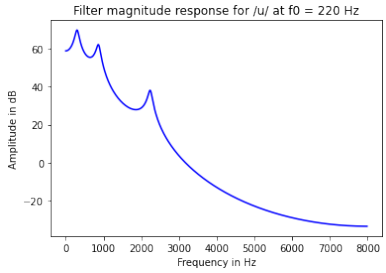


(b) Output response for pulse train

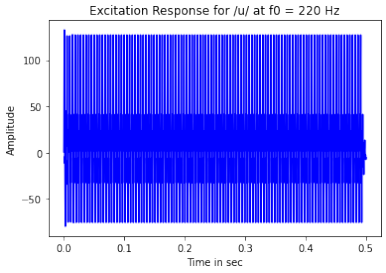


(c) Zoomed excitation response

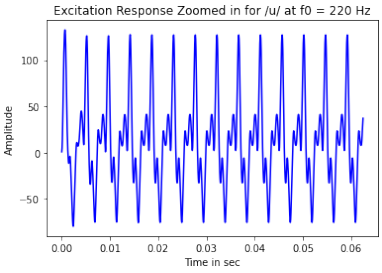
(f) **F0 = 220 for /u/**



(a) Magnitude response



(b) Output response for pulse train



(c) Zoomed excitation response

Observations

- On listening to the audios, we can distinguish the vowels, however the sounds produced are still noisy. The quality is not very good.
- On going from F0=120 to 220 Hz, the pitch increases for all as expected. Although they are like mechanical voice, but one can say that it feels like male voice for F0=120Hz and female voice for F0=220 Hz.
- One can also observe that there is not much change in shape of waveform for the same vowel at different F0, as the vocal tract parameters remain same. However, waveforms of two different vowels do not look alike due to the change in the formant frequencies and hence in vocal tract parameters responsible for articulation.

Note - Codes in .ipynb file