Linear Predictive Coding speech synthesis performance in different scenarios

# Intuition

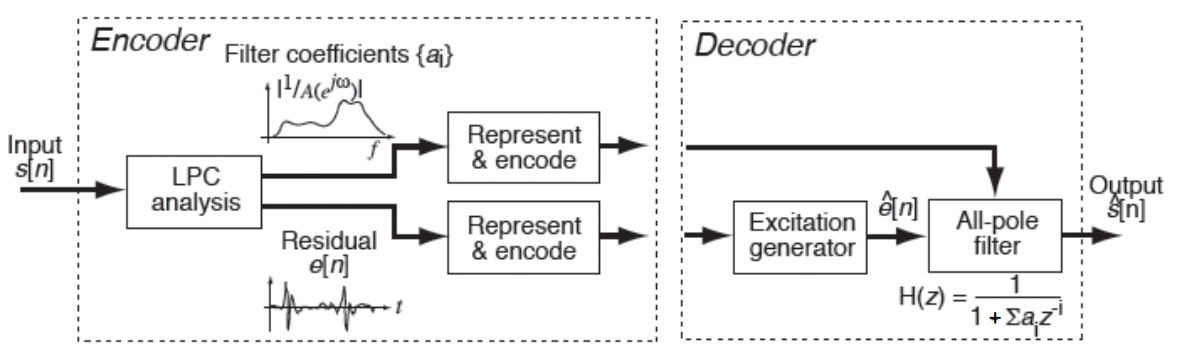
 (Ellis, 2013)

Figure 1: The overall process of speech synthesis of this course with 1 adjustment: Residual will not be computed. Linear Predictive Coding (LPC) approximation gives denominator coefficients of the vocal tract transfer function H(z) in polynomial form. Solving denominator coefficients gives the poles and consequently their angle and finally formant frequency and formant bandwidth, which are used to eliminate “false formant,” this entire process is “Represent & encode”. Finally, as vocal tract is treated as filter, so we apply it on the excitation pulses from “Excitation generator” to get synthesised speech

Throughout the coursework, we adjust segment duration, LPC pole order to evaluate how each affects the formant frequencies. In addition, we change excitation waveform to see how it affects the quality of synthesised speech.

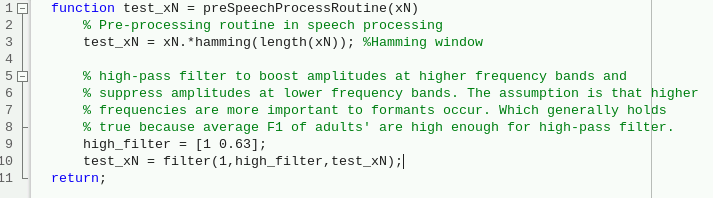
# Implementations

## Speech pre-processing routines

Pre-processing routines include 2 actions:

* Apply FIR filter window such as Hamming to suppress noises in the speech sample
* Apply high-pass filter to the speech to amplify magnitude at higher frequencies as they are usually where the formants occur

Though these 2 routines are not necessity, we think that at least it enhances readability of the speech spectrum and its spectral envelop, which is crucial whenever we need look at the plot to find ground truth formant frequencies for example.



## H(z) denominator coefficients

As vocal tract H(z) is commonly modelled as autoregressive (AR) model, thus we concern about the coefficients of the denominator H(z). There are 2 common methods to approximate coefficients of H(z):

1. Yule-Walker method which is available on MATLAB as lpc() function.
2. Covariance method which is available on MATLAB as arcov() function.

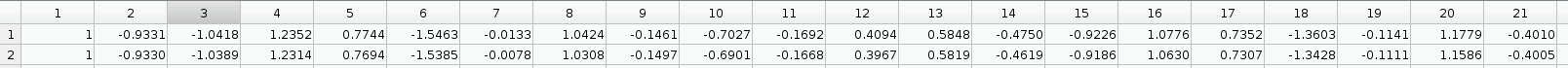
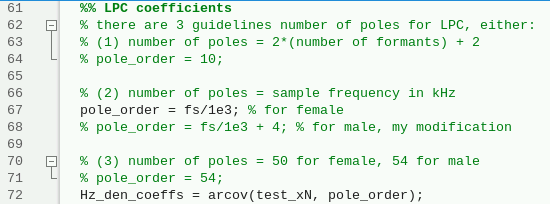


Figure : 1st row are the coefficients of denominator polynomial of H(z) approximated by Covariance method; 2nd row are the coefficients of denominator polynomial of H(z) approximated by Yule-Walker method. As coefficients on 2 rows are almost identical, using either method does not affect the performance of LPC speech synthesis.

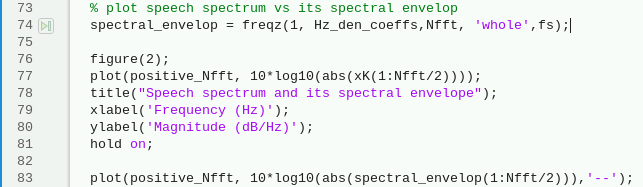
Whether readers use lpc() or arcov(), MATLAB will demand users to input the second argument: pole order (number of poles) for H(z). There are 3 school of thoughts for choosing pole order, either:

1. Number of poles = 2 × number of significant formants to find + 2
2. Number of poles = sampling frequency in kHz (for male, plus 4).
3. Number of poles for adult female = 50, and for adult male = 54 according to table A4 of this study (Derdemezis, et al., 2016)



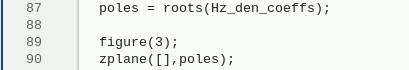
## Spectral envelop

To model spectral envelope, we compute the frequency response of H(z) with freqz() function in MATLAB.



## Formants

Conveniently, because the output of lpc() and arcov() is the denominator coefficients of H(z), we can immediately find the poles as root of denominator polynomial

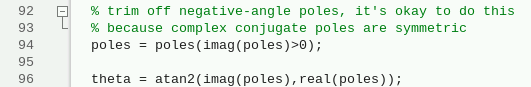


Next, we find every angle for every pole in pole-zero plot in Z-domain according to the following equation:

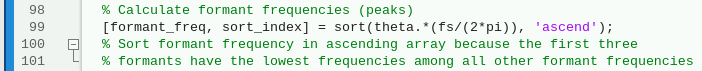
A drawing on a piece of paper

Description automatically generated with medium confidence

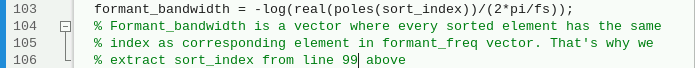
Figure 3: Pole-zero plot in z-domain with quadrant indices marked on different segments of the unit circle.



Once we have obtained for every pole, formant frequency can be computed as:

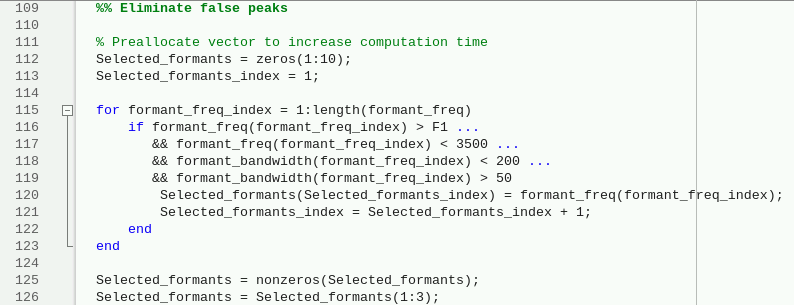


Correspondingly, formant bandwidth can be computed as:



With formant frequencies and formant bandwidths, we can eliminate false formants using institutional common sense about most significant formant frequencies and their formant bandwidths for vowels.

1. Most significant formants occur after fundamental frequency but not too high (<3500Hz) – table 1 of (Kent & Vorperian, 2018)
2. Most significant formants in adults have bandwidths about 50-200Hz - table 5 of (Kent & Vorperian, 2018)

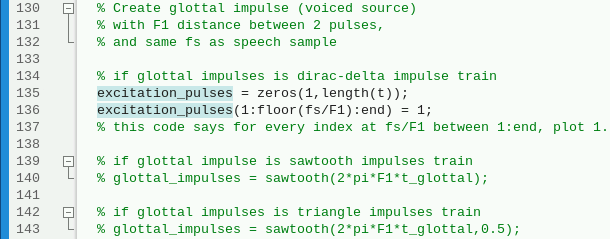


Let F1 = ground truth fundamental frequency (we name F1 to keep consistent the term 1st harmonic). From looking at dozens of speech segments with different segment duration and start times, we find F1 = 199.219Hz for female ‘whod’ speech and F1 = 123.047 Hz for male ‘whod’ speech.

## Speech synthesis

To generate excitation pulses, we have prepared 3 different waveforms to represent excitation pulses, either:

1. Kronecker delta impulses train
2. Sawtooth wave pulses train
3. Triangle wave pulses train



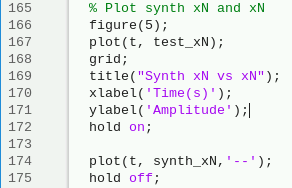
Then we add some noise to the excitation pulses, rather than creating random signal and convolve it to the excitation pulses. We use MATLAB built-in function awgn() to add noise to the pulses with respect to SNR, this allows to have another parameter to experiment with during evaluation.

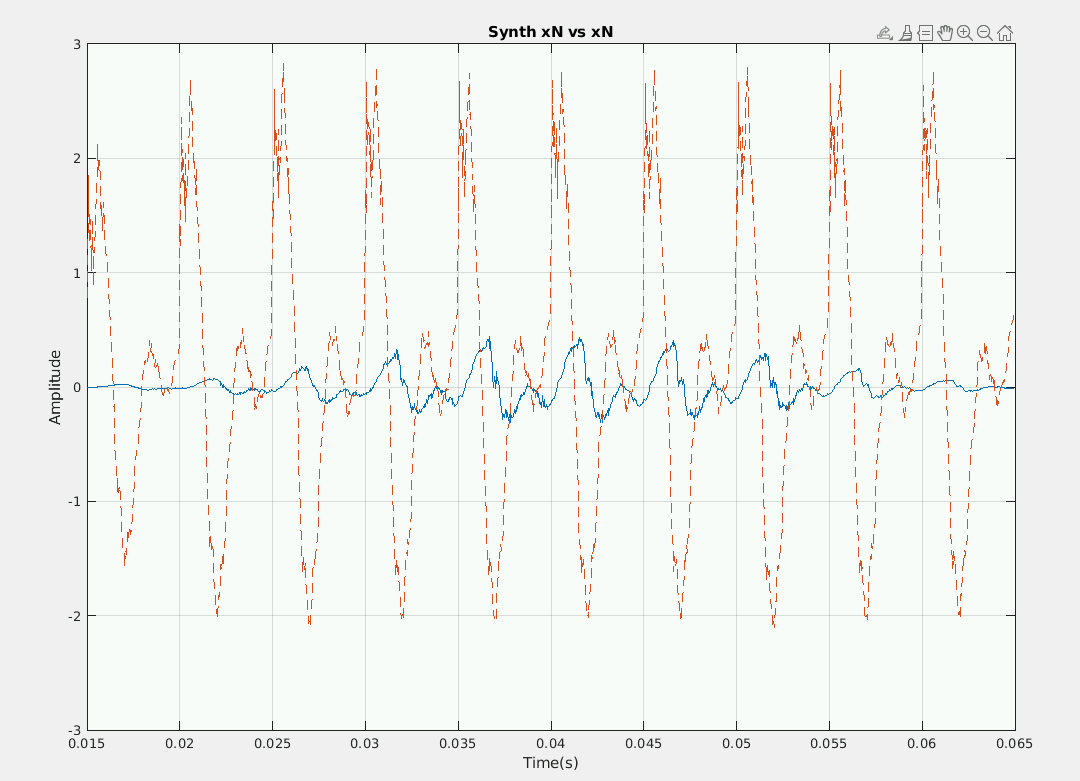


Finally, to create synthesised speech, we filter noisy excitation pulses into H(z) that we have just found earlier with either arcov() or lpc().

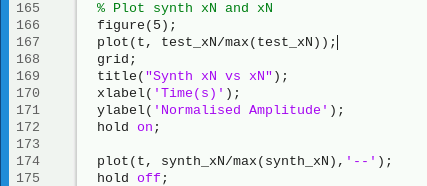


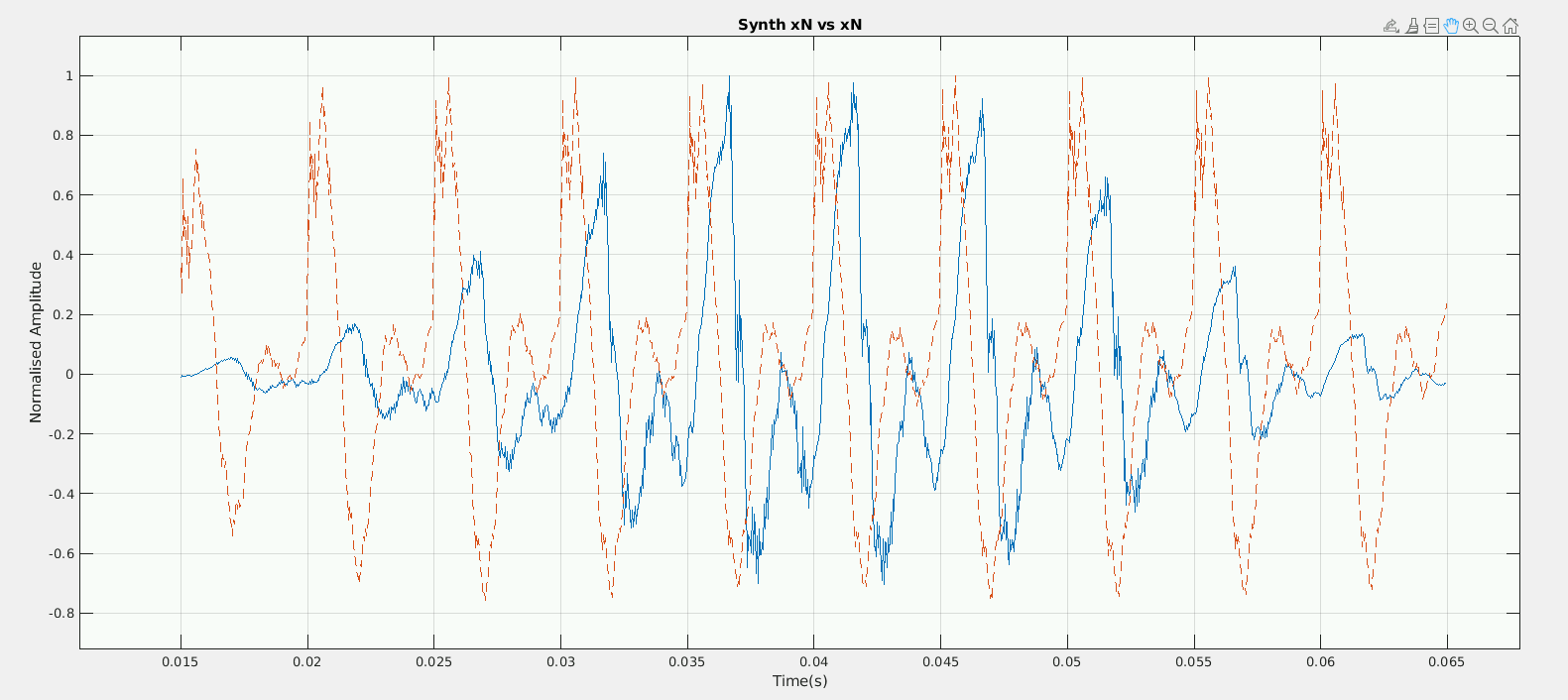
As we compare synthesised speech against original speech, we have realised that because synthesised speech has significantly greater amplitude, it is hard to compare both speeches on the same graph.





But if we normalise the amplitudes of 2 speeches, then they become much easier to compare on the same graph





However, on frequency domain, we do not normalise the magnitude of the two speeches’ spectrums as we feel they might be crucial information should

# Evaluation

## Segment duration and formant frequency

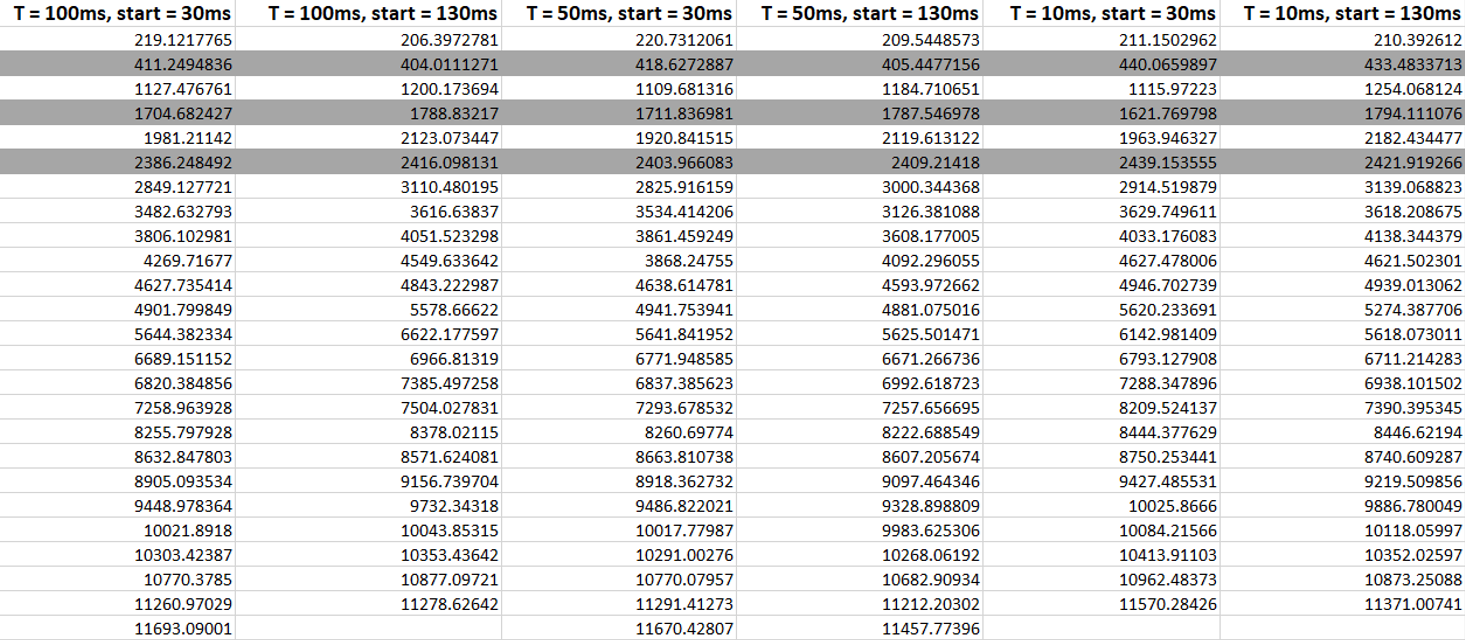


Table 1: Formant frequencies of female ‘whod’ speech with AR pole order = 50 at different starting time and segment duration, shaded rows are ground truth most three significant formants (see figure 4 below). Ground truth 2nd formant is most affected by different segment duration. However, when speech segment starts at = 130ms, 2nd formant is significantly less affected by different segment duration.

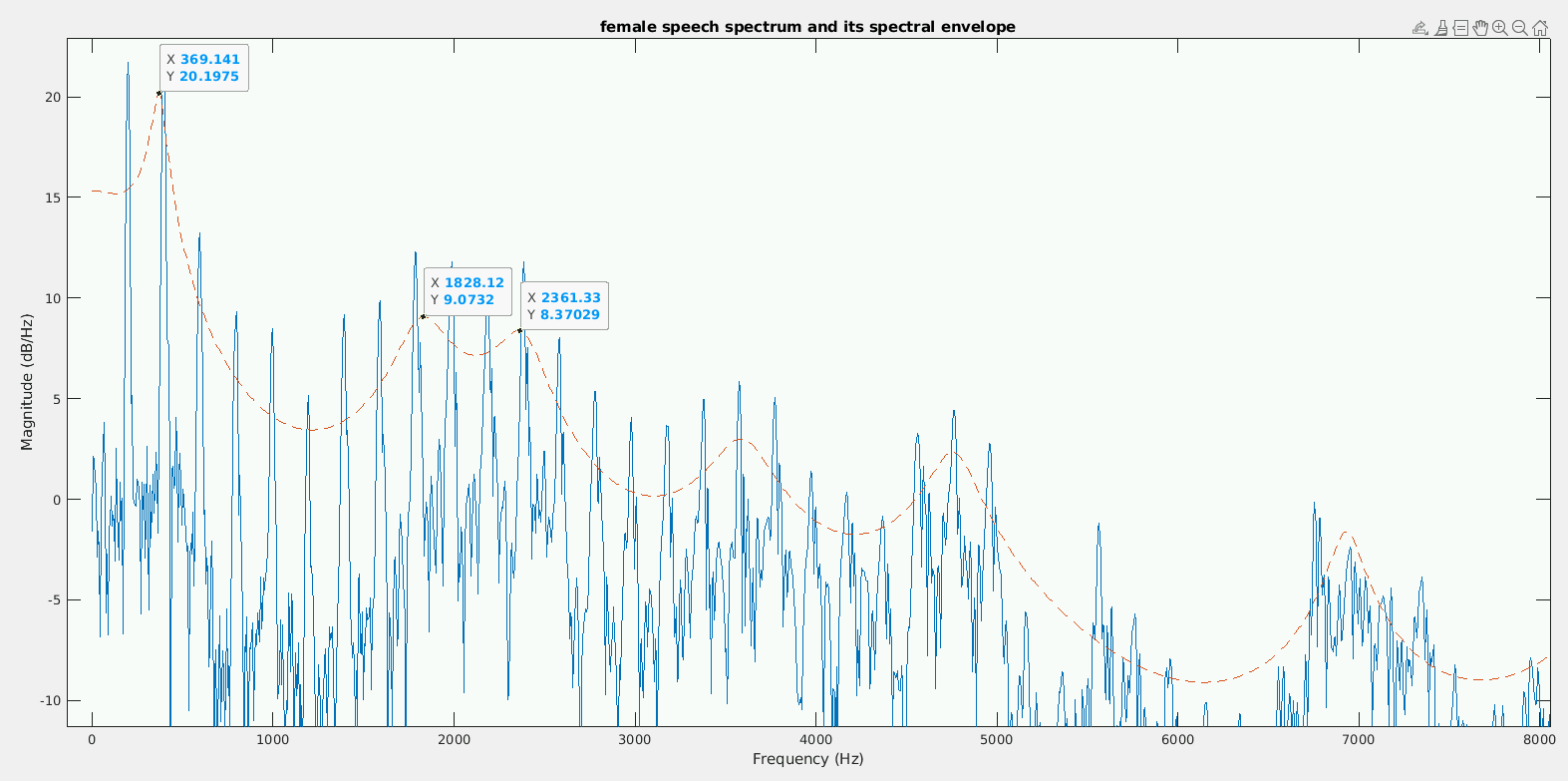


Figure 4: Female ‘whod’ speech spectrum and its spectral envelop when AR pole order = 24. After we have looked at over dozens spectral envelops of different pole order, start time and segment duration; we notice a consistent pattern that ground truth 1st, 2nd, and 3rd formant always occur at roughly (± 64Hz) 369Hz, 1828Hz, 2361Hz. However, when AR pole order is small, 1st formant is shifted dramatically while 2nd formant disappear. Therefore, we advise against using small AR pole order (<20)

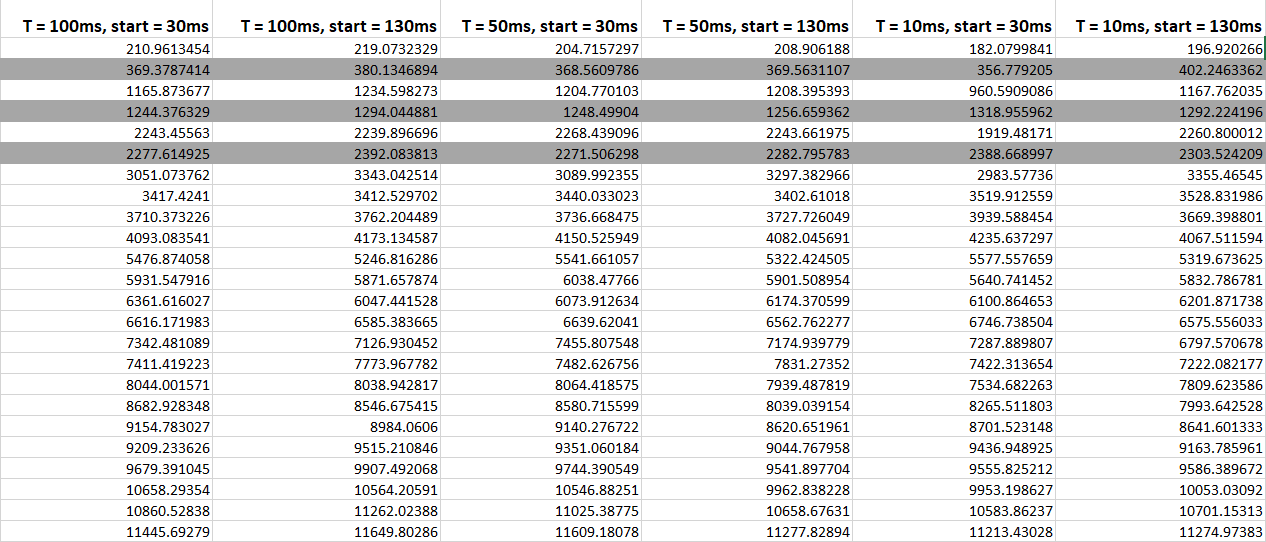


Table : Formant frequencies of male ‘whod’ speech at AR pole order = 50 with different starting time and segment duration. Shaded rows are ground truth three most significant formants (see figure 5 below). Ground truth 2nd formant is still most affected by different segment duration. However, 1st and 3rd ground truth formants are now also more marginally affected by different segment duration now as opposed to back in table 1. Again, we observe that when start time = 130ms, formant frequency is less affected by different segment duration.

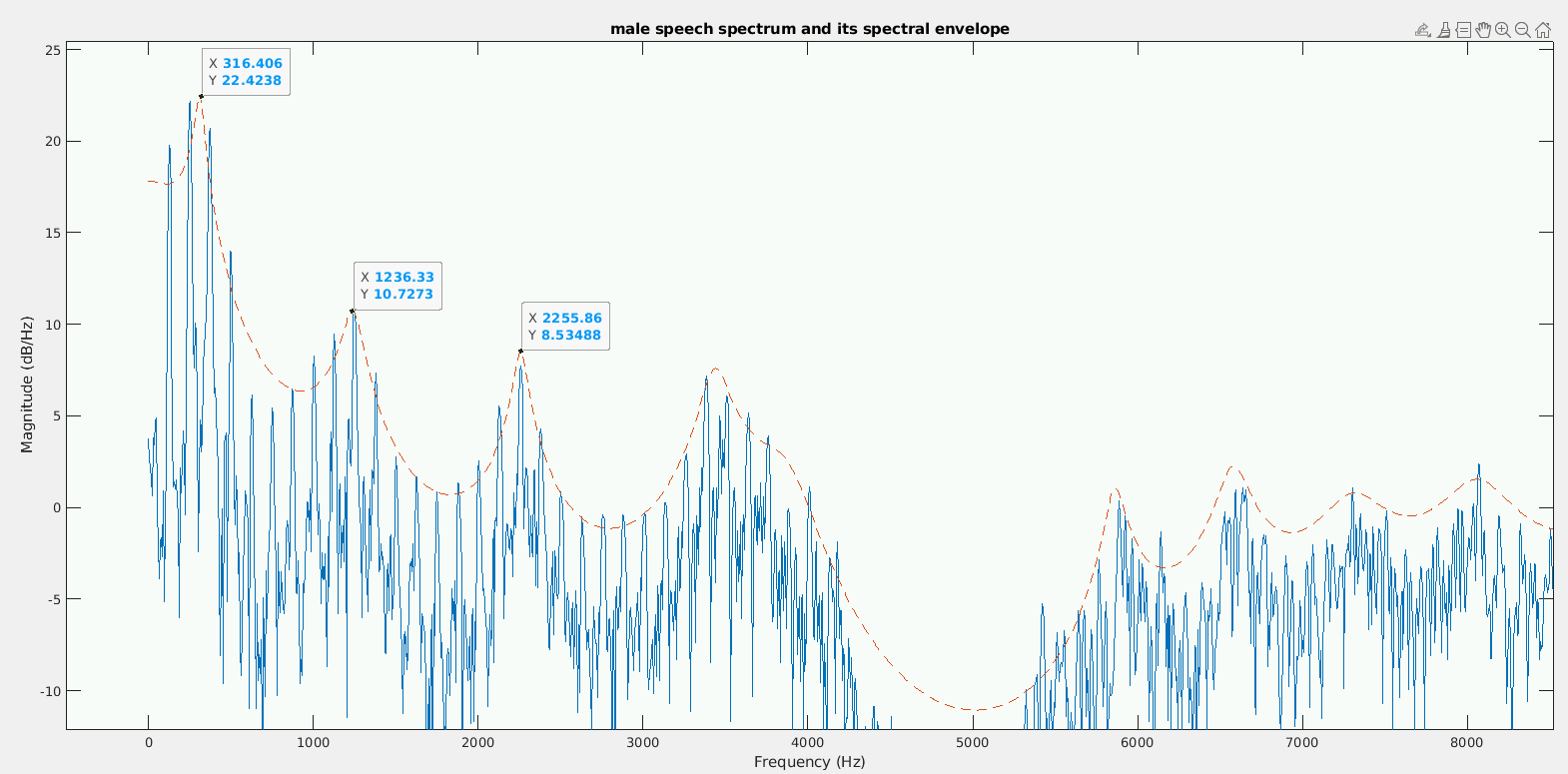


Figure 5: Male ‘whod’ speech spectrum and its spectral envelop when AR pole order = 28. After we have looked at over dozens spectral envelops of different pole order, start time and segment duration, we notice a consistent pattern that ground truth 1st, 2nd, and 3rd formant always occur at roughly (±64 Hz) 298Hz, 1236Hz, 2279Hz. However, having high AR pole order or low AR pole order change 1st formant dramatically and potentially vanish the other twos. Therefore, we advise using the pole-order within our recommended sweet spot (25-30)

## AR pole order and formant frequency

Both male and female speech samples suffer the same issues at different pole order: When we go by the 1st school of thought for pole order: Number of poles = 2 × number of significant formants to find + 2. We have made the following discoveries in figure 6 and figure 7:

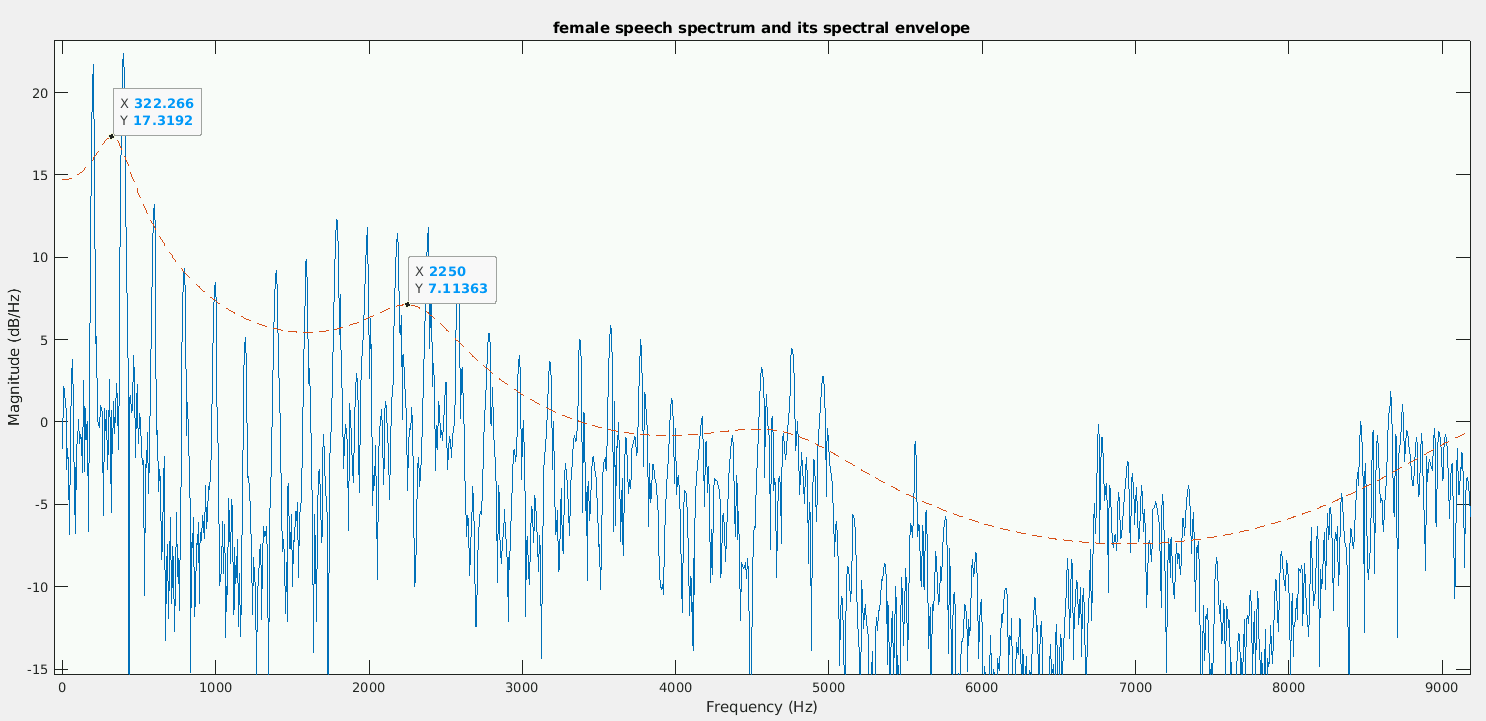


Figure 6: Female ‘whod’ speech spectrum in comparison to its spectral envelop for AR order = 2 × number of significant formants to find + 2. In this AR setting, the ground truth 2nd formant vanishes.

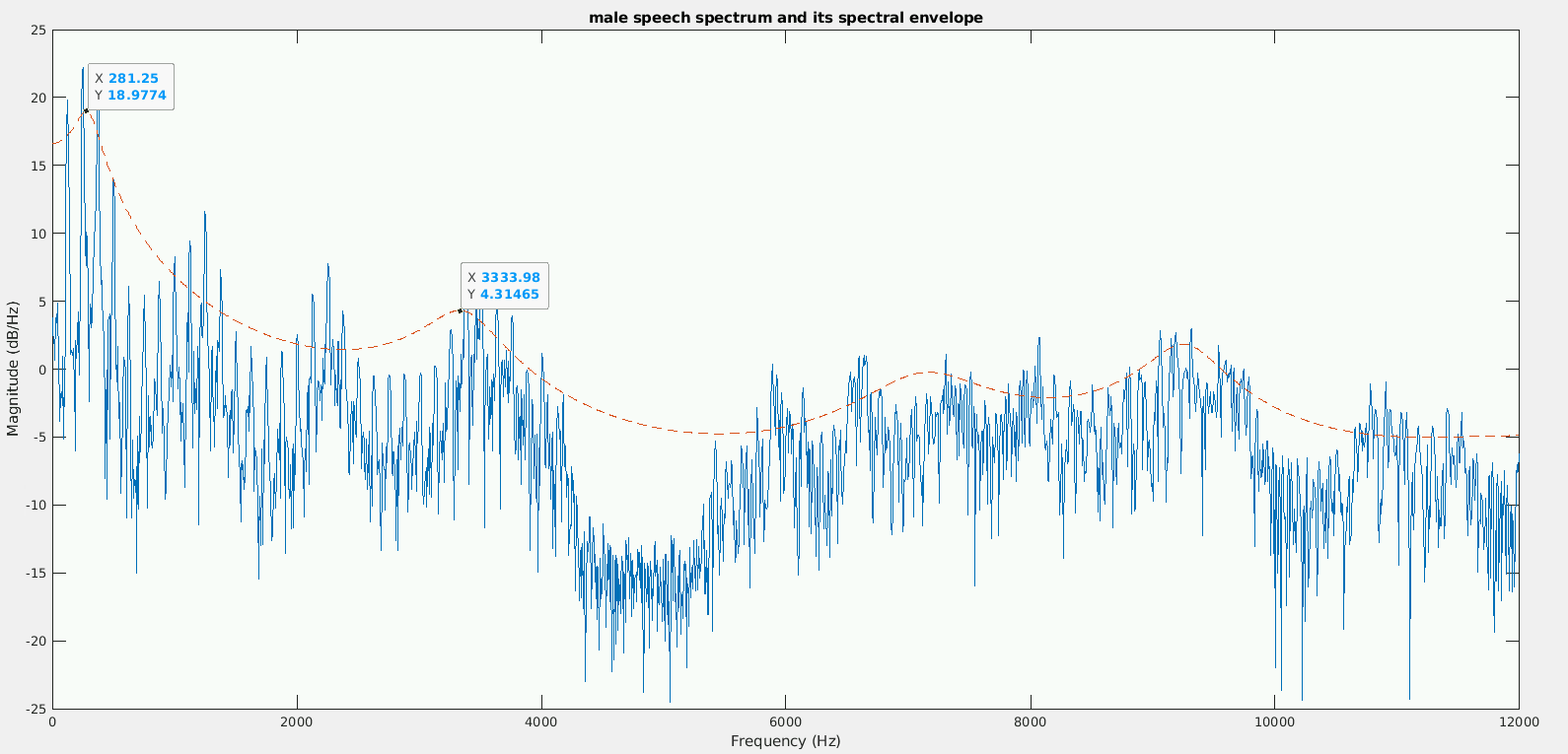


Figure 7: Male ‘whod’ speech spectrum in comparison to its spectral envelop for AR order = 2 × number of significant formants to find + 2. In this AR setting, the ground truth 2nd and 3rd formant vanish, which says that male’s spectral envelop is more adversely affected by low AR pole order than female’s spectral envelop.

On the other hand, when we go by the 3rd school of thought: number of poles = 50 for female and 54 for male speech, so many formant frequencies were generated by the LPC model that our “false formant elimination” scheme that we have implemented above fails to acquire ground truth formants from figure 4 and 5. This notion is reflected in table 3 and table 4

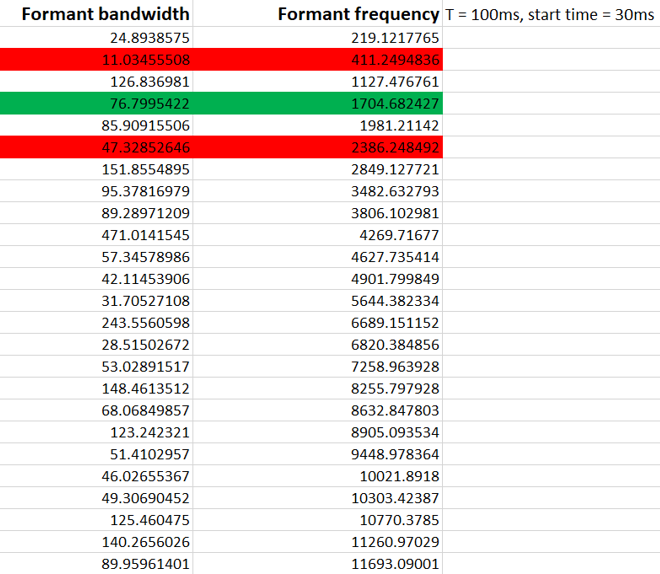


Table 3: Formant bandwidth and formant frequency of female ‘whod’. Formant frequency of this table is extracted from the 1st column of table 1. Red rows are ground truth formant rejected by “false formant elimination” scheme (False Negative). Green rows are ground truth formants accepted by the scheme (True Positive)

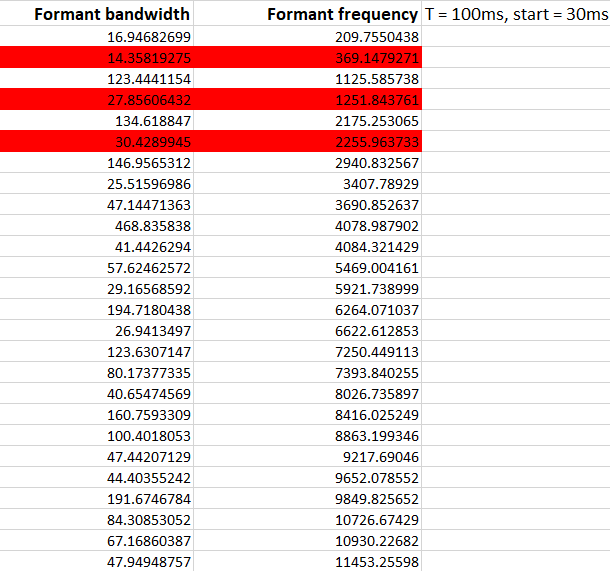


Table : Formant bandwidth and formant frequency of male ‘whod’. Formant frequency of this table is extracted from the 1st column of table 2. Red rows are ground truth formants rejected by “false formant elimination” scheme (False Negative).

Finally, when we go by the 2nd school of thought: Number of poles = sampling frequency in kHz (with some modification for male’s formants), we notice that we can completely and consistently ignore the “false formant elimination” scheme altogether and just pick the first 3 formants in the formant\_freq array to acquire the ground truth significant formants, as demonstrated in table 5 and table 6

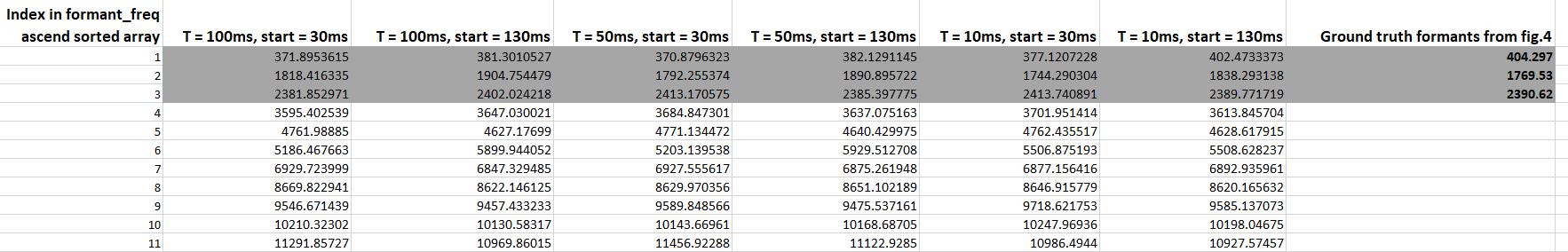


Table 5: Formant frequencies of female ‘whod’ speech at different segment duration and start time in comparison to ground truth formants from figure 4. The first three formant in formant\_freq array are also the ground truth formants consistently

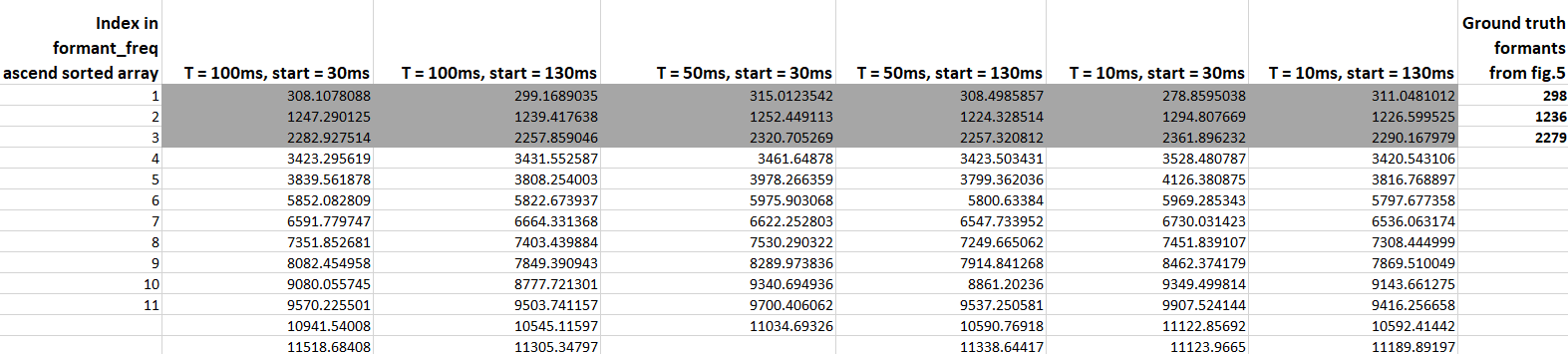


Table : Formant frequencies of male ‘whod’ speech at different segment duration and start time in comparison to ground truth formants from figure 5. The first three formant in formant\_freq array are also the ground truth formants consistently

## Conclusions about duration and AR order

In overall, male’s formant is more affected by different segment duration than female’s counterparts. However, both are less affected when start time is roughly at the middle of the original speech sample (130ms, as speech sample lasts about 250ms).

For female’s speech, we recommend using AR order at the sweet spot (24-27), and for male’s speech the sweet spot is (25-30).

## Excitation waveform and synthesised speech quality

At this stage, we use long segments (200ms) as they are easier to evaluate by listening while short segments are easier (50ms) to evaluate by looking at the plots.

For Kronecker-delta impulses, we discover that synthesised speeches are consistently out of phase in both genders when comparing to original speeches (figure 8 and 9), even though there is little noise within the excitation pulses. As a result, though the word ‘whod’ is still distinguishable in hearing synthesised speech, it is nonetheless no longer sounds natural and feels noisy despite there is little noise in the excitation pulses.

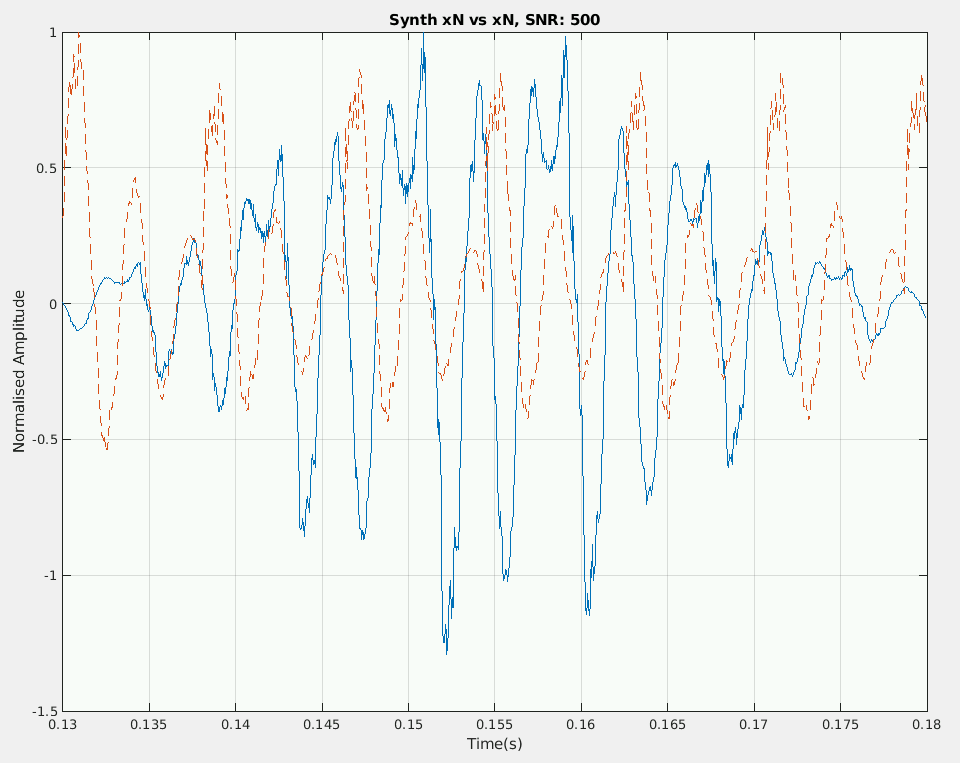


Figure : Synthesised ‘whod’ comparing to original ‘whod’ of male speech. Synthesised speech is created by filtering noisy Kronecker delta impulses train with AR vocal tract. Observation of the plot uncovers that the synthesised speech is out of phase and misrepresent the natural build-up amplitude of the speech at the beginning.

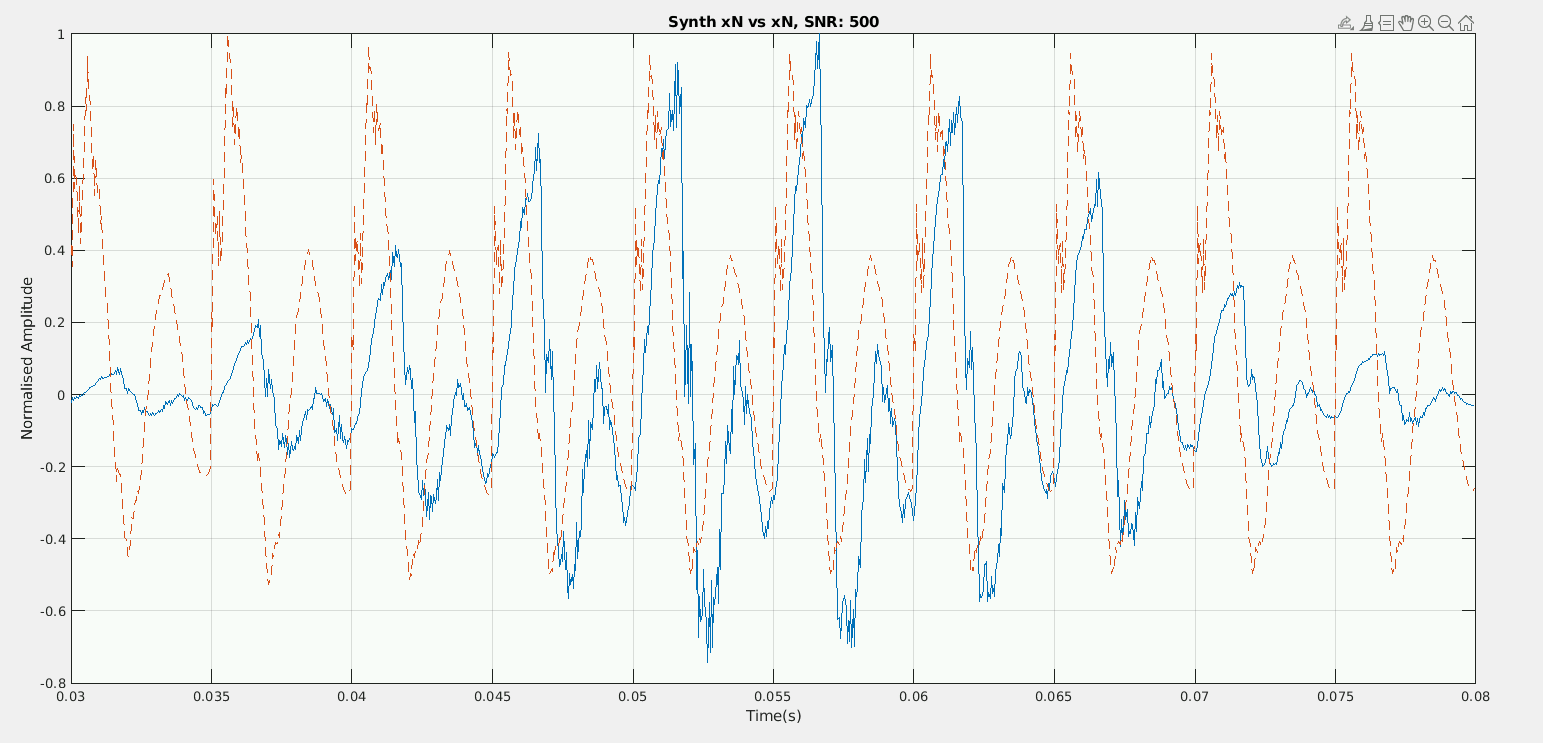


Figure : Synthesised ‘whod’ vs original ‘whod’ of female speech. Synthesised speech is created by filtering noisy Kronecker delta impulses train with AR vocal tract. Same as male’s synthesised speech, the synthesised female speech is also out of phase and misrepresent the natural build-up amplitude at the beginning; but comparing to male synthesised speech, female synthesised speech is more in phase at the mid-segment.

For sawtooth wave, the out of phase mismatch is at even greater extent. In addition, there are 2 new mismatches: (1) negative amplitude is surprising large in middle segment; (2) the synthetic speech smooths out the rugged pattern of the original speech (see figure 10 and 11). Therefore, despite common sense tells us that sawtooth pulses should output more accurate data and thus better sound quality, Kronecker delta impulse delivers better quality than sawtooth pulses

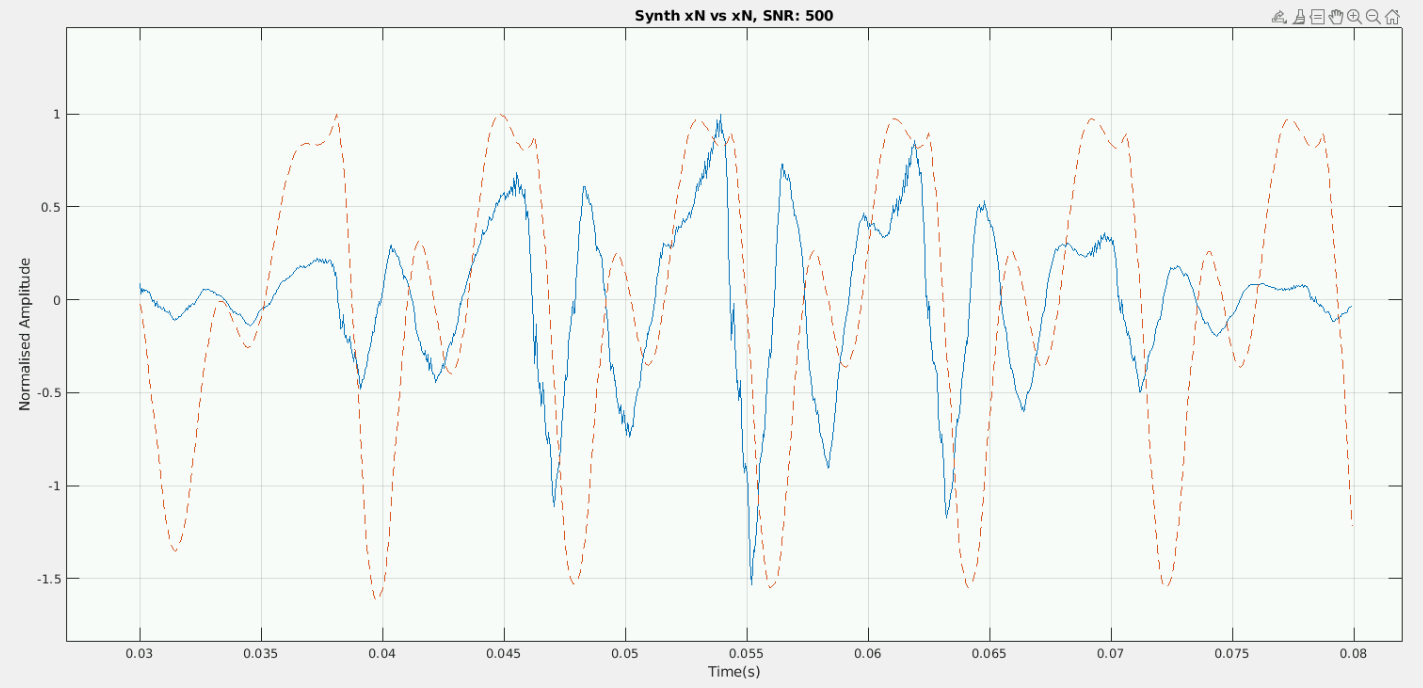


Figure 10: Synthesised ‘whod’ comparing to original ‘whod’ of male speech. Synthesised speech is created by filtering noisy sawtooth pulses train with AR vocal tract. Comparing to synthesised speech from figure 8, this synthesised speech is marginally more out of phase and had unusually large negative amplitudes.

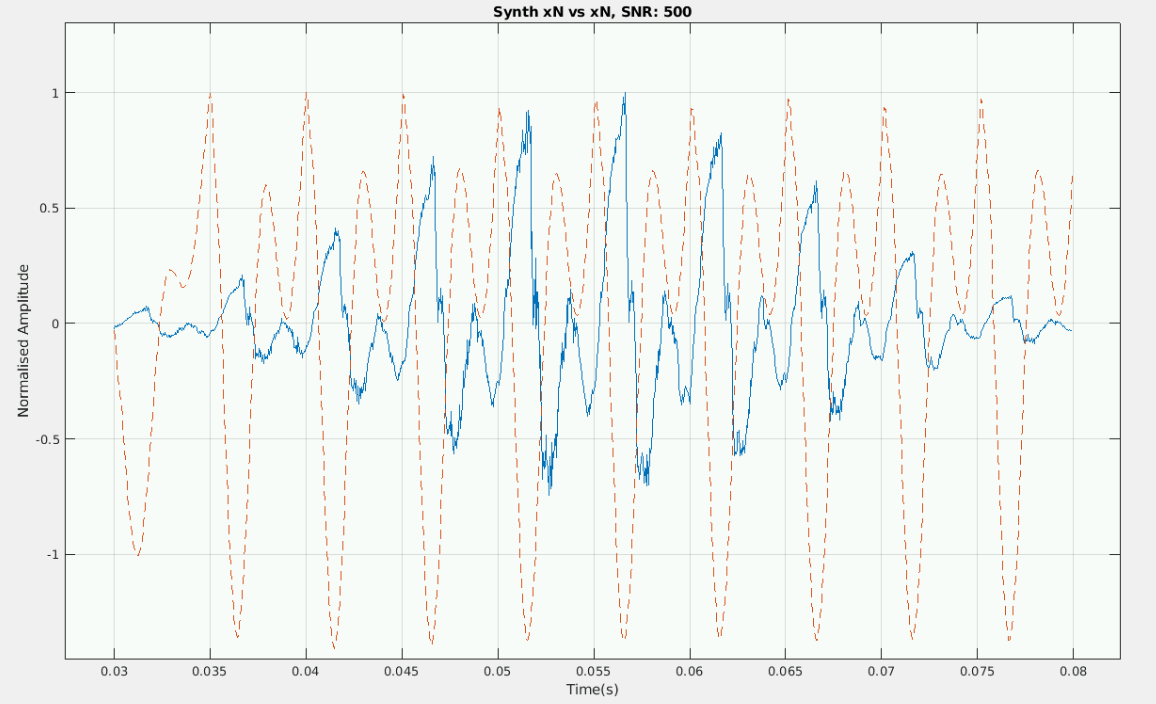


Figure 11: Synthesised ‘whod’ comparing to original ‘whod’ of female speech. Synthesised speech is created by filtering noisy sawtooth pulses train with AR vocal tract. Comparing to synthesised speech from figure 9, this synthesised speech is marginally more out of phase and had unusually large negative amplitudes.

Similarly, synthesised speeches created from triangle pulses inherit most of the same issues as sawtooth pulses (figure 12 and 13) thus Kronecker delta impulses still outputs better audio quality for synthesised speech than triangle pulses

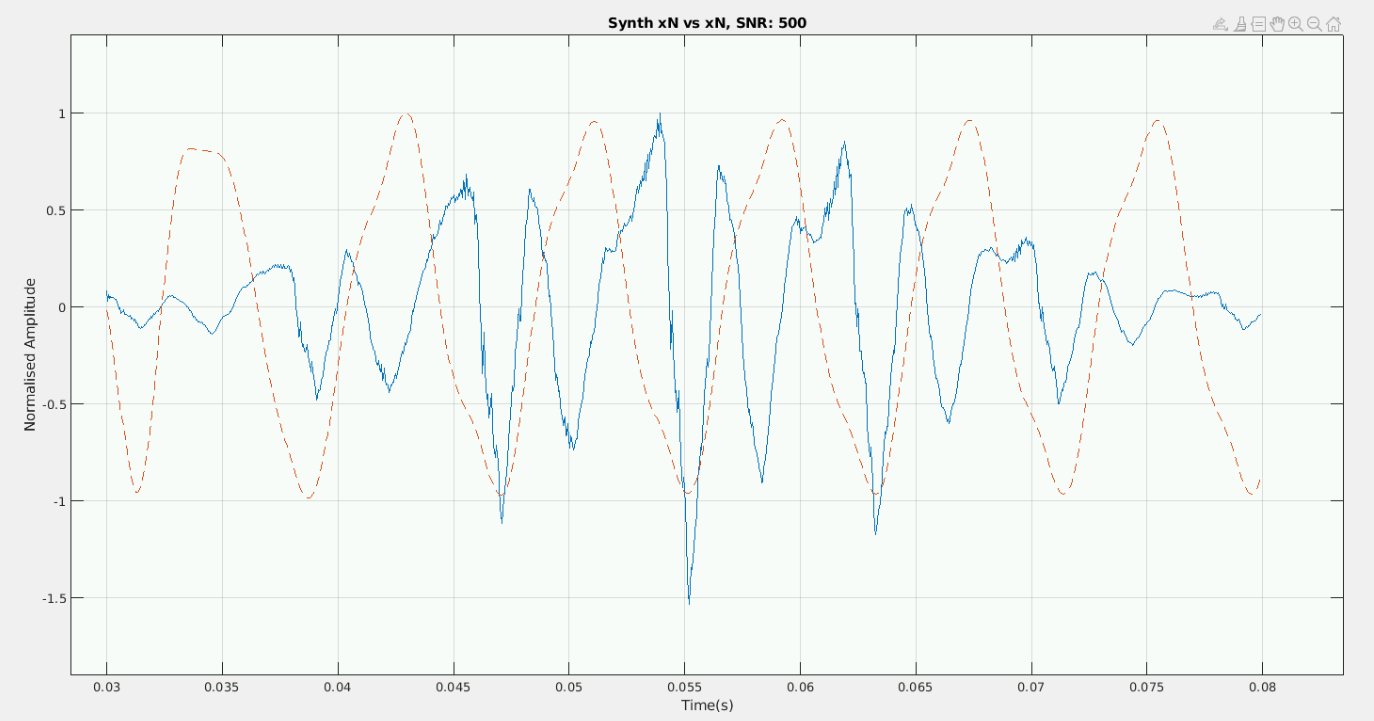


Figure 12: Synthesised ‘whod’ comparing to original ‘whod’ of male speech. Synthesised speech is created by filtering noisy triangle pulses train with AR vocal tract.

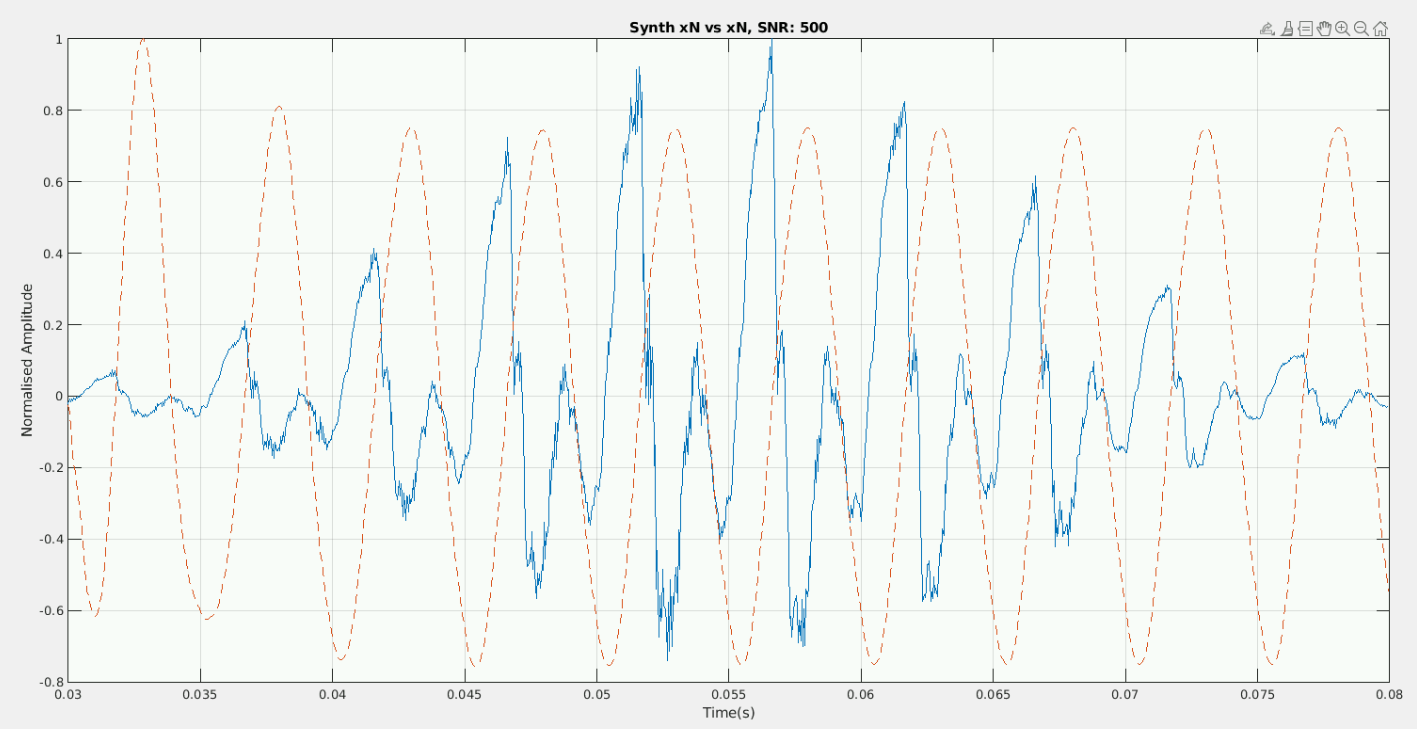


Figure 13: Synthesised ‘whod’ comparing to original ‘whod’ of female speech. Synthesised speech is created by filtering noisy triangle pulses train with AR vocal tract.

# Conclusion

We discover that male’s formant is more affected by different segment duration than female’s counterparts, but both genders’ formants are less affected when start time is roughly at the middle of the original speech sample (130ms, as speech sample lasts about 250ms) so we recommend starting the segment at near middle of the speech rather than at beginning. In term of AR pole order, we recommend using the sweet spot (24-27) when analysing female’s formants and (25-30) when analysing male’s formants. Finally, to our surprise, Kronecker impulses deliver better sound quality when being filtered through AR vocal tract filter.