CS578 Speech Processing
Laboratory 1
Linear Predictive Coding

George Manos
csd4333@csd.uoc.gr

Alexandros Angelakis
csd4334@csd.uoc.gr

13 December 2022

1 Analysis-Synthesis based on LP

For the LPC analysis part, in order to find the LPC coefficients and thus the gain and the excitation, we only used the positive auto-correlation values of our framed signal. Then for those values, we used the Levinson Recursion method to find the LPC coefficients and finally for the gain we used the equation:

$$Gain = \sqrt{r_n[0] - \sum_{k=1}^{p} a_k r_n[k]}$$

and for the excitation we just passed our framed signal through the filter with the LPC coefficients.

Our algorithm works really well, you can neither hear nor see any difference between the original and the synthesized signal.

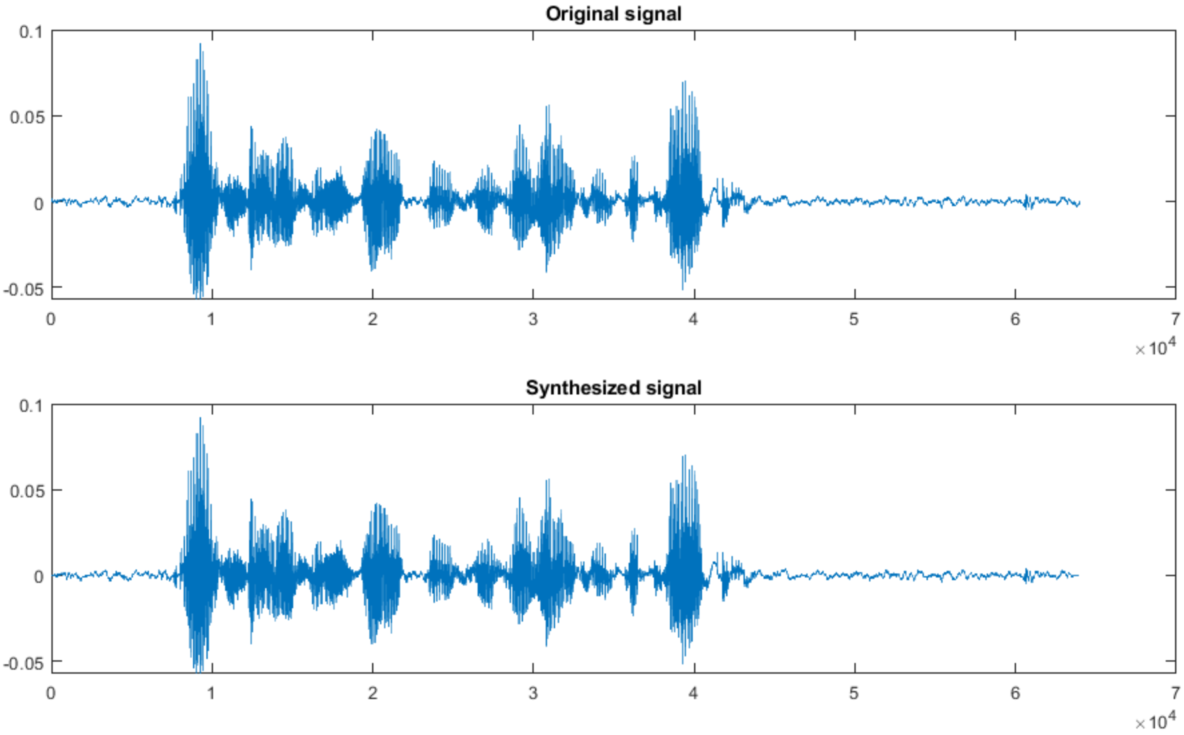

Figure 1: The original and the synthesized signal

We can see from the plot above that there are no differences between those two signals and their mean squared error (MSE) is extremely low, 2.0256e−08. All the above were tested with LPC order of 24. Regardless of the order number, the output synthesized signal and the MSE are the same.

2 Frame-by-frame the analysis procedure

We've discussed in some lectures, that the magnitude of the LP filter smoothly surrounds the magnitude of the Fourier transform of the speech frame. So in our experiments, this is exactly what we expect to see, a smooth line enveloping the Fourier transform of the speech frame, known as envelope. We've taken two frames from a speech signal, one voiced and one unvoiced, and plotted their FFT magnitude with their LP filter magnitude with three different LPC orders, our initial value (24), 12, 6, 48 and 96.

We can see that when the LPC order is very high, for example 96, almost all the peaks of the magnitude of the Fourier transform of the speech frame has been covered by the magnitude of the LP filter. For the voiced frame, all the formants have been enveloped and for the unvoiced frame, all the peaks have been covered. As a consequence, we can say that large LPC orders cover fully the magnitude spectra of all the frames. However, this is not the ideal case, we want something in between where the envelope covers the magnitude but not every peak of it. The plot with LPC order 24 is a really good example for this, thus this is the best case scenario. On the other hand, when we have a very low order, for example 6, we can see that it is not capable of covering the magnitude of the frame, neither for the voiced nor unvoiced frame. So, for smaller order numbers, the magnitude of the LP filter underfits the magnitude of the speech frame and vice versa, for higher order numbers the more it overfits.

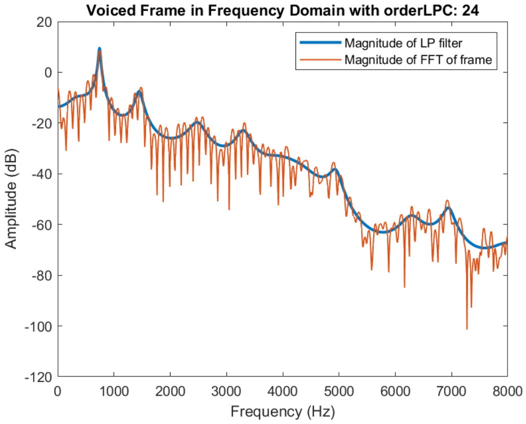

Figure 2: Voiced with LPC order = 24

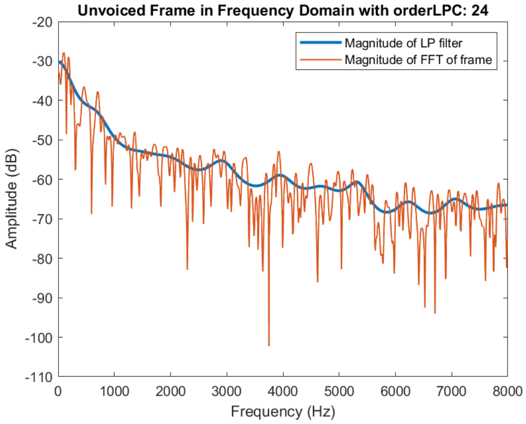

Figure 3: Unvoiced with LPC order = 24

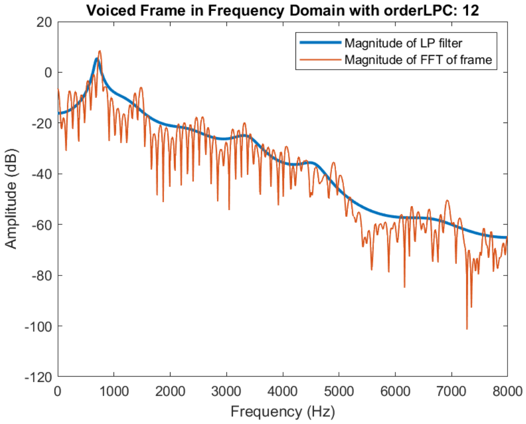

Figure 4: Voiced with LPC order = 12

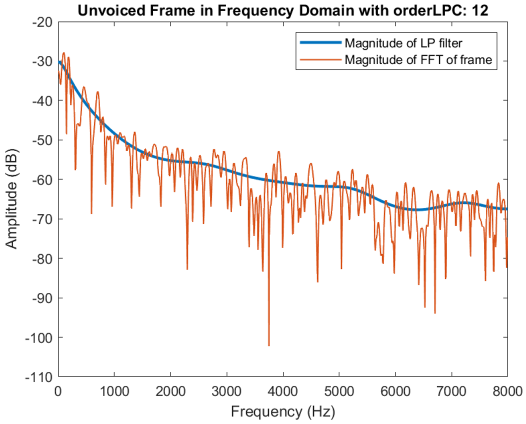

Figure 5: Unvoiced with LPC order = 12

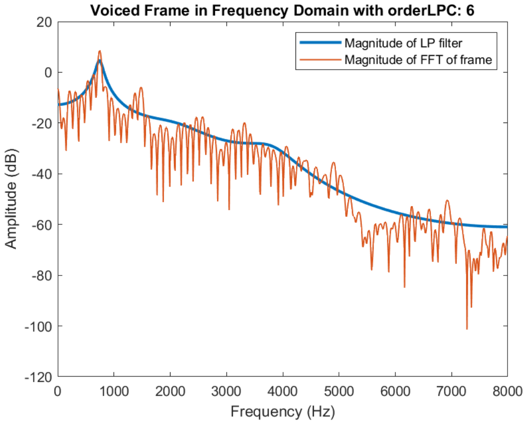

Figure 6: Voiced with LPC order = 6

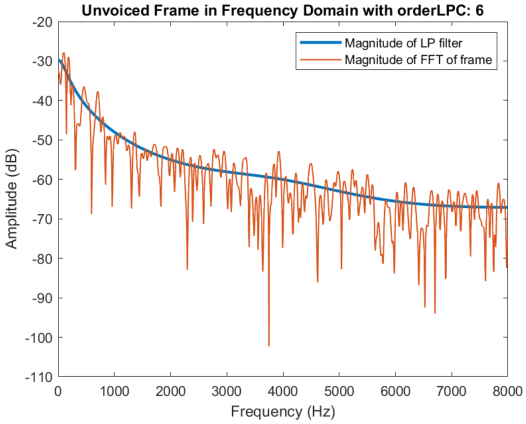

Figure 7: Unvoiced with LPC order = 6

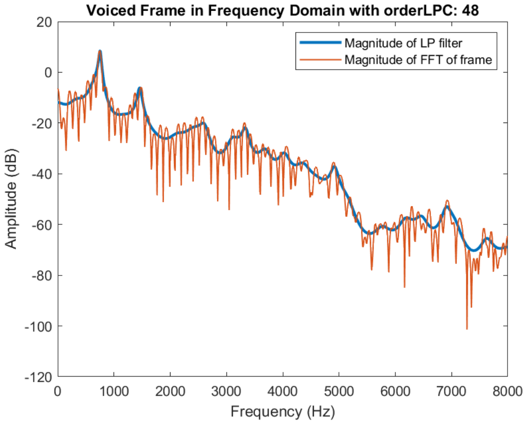

Figure 8: Voiced with LPC order = 48

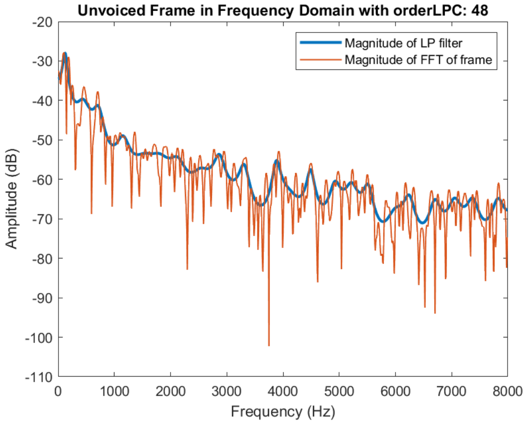

Figure 9: Unvoiced with LPC order = 48

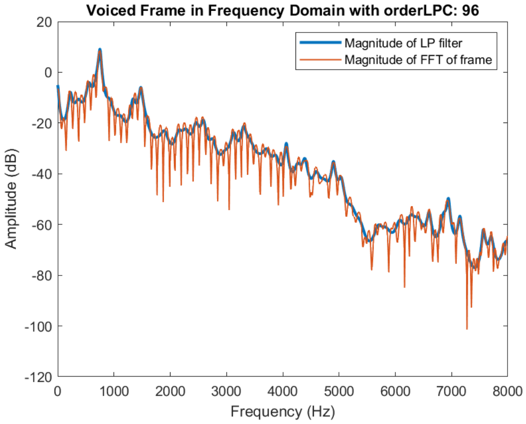

Figure 10: Voiced with LPC order = 96

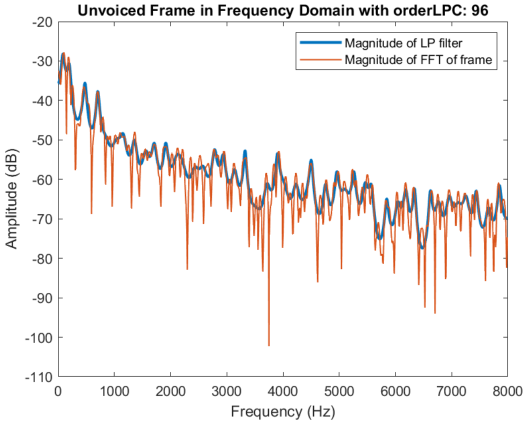

Figure 11: Unvoiced with LPC order = 96

3 Modifying the excitation signal

3.1 Whisper Voice

For the whisper voice, we took values from the standard normal distribution and consider them as our excitation signal. Since we use normally distributed random number, we expect the excitation signal to behave as random noise. We can also see that from the excitation's plot below.

Here, we can understand exactly what the excitation signal represents. We know that a speech excitation signal is produced by an excitation source (vocal chords in our case) and processed by a filter system that "modulates" the spectral characteristics of the excitation signal based on the shape of the vocal tract for the specific sound being generated*. Having already computed the filter that models the shape of the vocal tract (filter H(z)), by replacing the excitation signal with random noise and passing it on the same filter we get a noisy voice that sounds more like a whisper since the aforementioned spectral characteristics are changed.

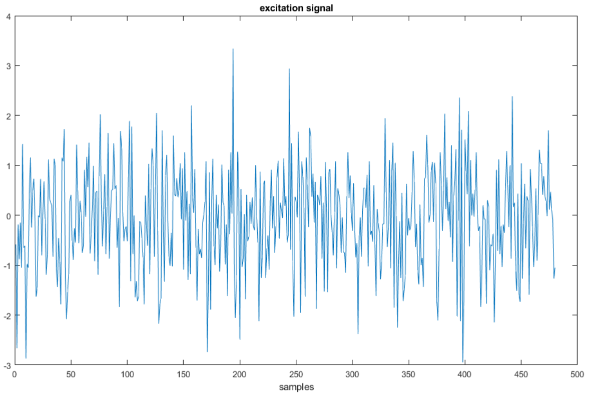

Figure 12: Excitation signal from the last frame

By using our default LPC order (24), the output signal is pretty audible, you can understand what the speaker says but it has a lot of noise (background noise that has a robotic feeling to it). The synthesized output sounds slightly like a whisper voice, it mostly sounds like a robot voice whispering.

Using lower LPC order (12), makes our synthesized output less audible with more noise, but we can still understand the speaker whispering. Moreover, by having 6 as LPC order, the noise nearly outweights the voice signal, and it sounds more synthesized than normal.

Now, by increasing the LPC order to 48, we notice that there is no big difference with the order 24, because there are no big variations in the error. Surely the synthesized output with order 48 sounds a little bit better, but it is not worth the computational cost. The same goes with LPC order that are larger than 48. Thus, we can say that our default order of 24, has the best performance over cost, and it's the one we would want to use.

*B. H. Juang et. al., "Digital Speech Processing", Bell Laboratories and AT&T Laboratories, Encyclopedia of Physical Science and Technology, 2003

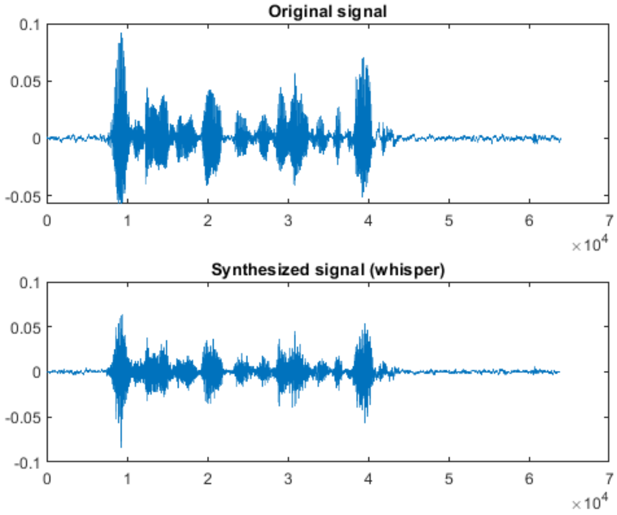

Figure 13: The original and the synthesized whisper signal

3.2 Robot Voice

This is where we modify the excitation signal, so that the synthesized output sounds like a robot voice that uses the same pitch period as the speaker. Essentially, we simply took the excitation filter and found its peaks. We then set all the locations of the peaks of the excitation signal to 1 and the rest to 0.

Considering our initial value of the LPC order (24), we can hear that the speaker's voice is now robotic, but that also affects the background noise, since now it also has a robotic feeling and it's annoying to our ears. Lowering the LPC order of our analysis (around 6 or 8), we can see now that the quality of the speech signal is worse and that robotic background noise is more perceivable. We can still understand what the speaker says tho. On the other hand, increasing the LPC order (around 30-35), we can see that there is no big difference between this order and our initial LPC order. The one difference that stands out is that when we increase the LPC order, we can slightly hear the voice of the speaker without it being robotic, but the robotic background and voice is dominant.

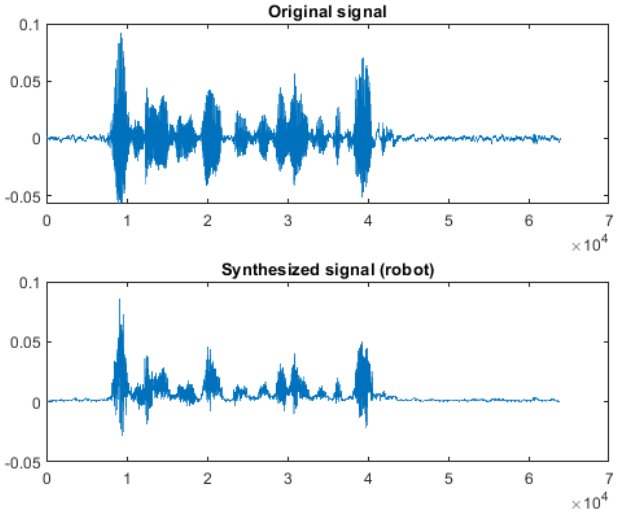

Figure 14: The original and the synthesized robotic signal

4 Modifying the vocal tract

This is where we modify the vocal tract for our speech signal. Essentially, we approximate the formant frequencies of the signal by taking the 3 largest by magnitude non real poles[†], as resulted from the Levinson's coefficients, along with their conjugate poles (6 in total). Then, we multiply those poles by a coefficient and recreate the filter. The coefficient will usually range from 0.8 to 1.2, corresponding to up to a 20% decrease and increase respectively on those formants, that will affect the speaker's age. Of course, if that coefficient is equal to 1, the speaker's age is unnaffected.

While keeping the LPC order constant, we can see that the more we reduce those formants, the more elderly the resulting voice becomes. Respectively, the more we increase those formants, the more child-like the resulting voice becomes. For high LPC orders (around 25 to 30), since we have more LPC coefficients, the voice age is affected more by those changes, but also some noise is added to the speech (sounds more ”glitchy”), and vice versa for low LPC orders (around 7 or 8), the formant changes are less impactful to the age but also less noise is added to the speech signal.

[†] Of course, a formant may not be in 0 or π frequency, so we need to discard any real poles