Speech Processing Using Linear Prediction

Connor McCullough, University of Miami

***Abstract*—Linear prediction is used in a variety of speech processing applications such as speech encoding and altering of the characteristics of the human voice. This project analyzes the use of linear prediction for analyzing and synthesizing speech signals. Linear prediction is used to isolate the glottal pulse and the formants of the vocal tract so they can be processed independently, and subsequently synthesized into edited speech.**

***Keywords*—linear prediction; speech processing; formant; analysis; synthesis;**

I. Introduction

Linear prediction is an operation where future values of a signal are estimated as a linear function of previous samples. In signal processing, linear prediction is often used in linear predictive coding, which allows the spectral envelope of a speech signal to be represented in a compressed form. While compression is one benefit of this representation, an additional use is to decompose the speech signal into its basic components: the glottal pulse and the transfer function of the vocal tract. Because the glottal pulse is a harmonic signal, it has periodic harmonics which appear in the FFT. By using linear prediction of the proper order, these harmonics are smoothed out, leaving only the envelope of the formants created by the vocal tract. Calculating the error signal of the original and linear prediction gives the frequency response of the glottal pulse. This allows various operations to be performed on the signal of the glottal pulse, and a synthesized speech signal can be created by then filtering this signal using the linear prediction coefficients.

II. The Role of Linear Prediction Residual Signal in Quality of Synthesized Speech

A. Reconstruction Using e[n] for Synthesis

The error signal e(n), as mentioned previously, is obtained by finding the error between the original frequency response and the linear prediction, and then taking the IFFT to get to the time domain. E[n] can be reconstructed into the original signal with minimal error by convolving it with the linear prediction coefficients, which represent the frequency response of the vocal tract.

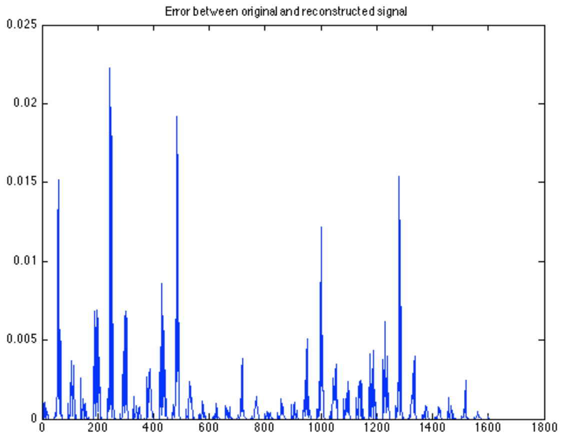


Fig. 1. Error plot for Reconstructed signal.

The error of this process can be seen above. The small periodic error is a result of mathematical error in the Hamming windowing. The total calculated rms error for the reconstruction was around 8%.

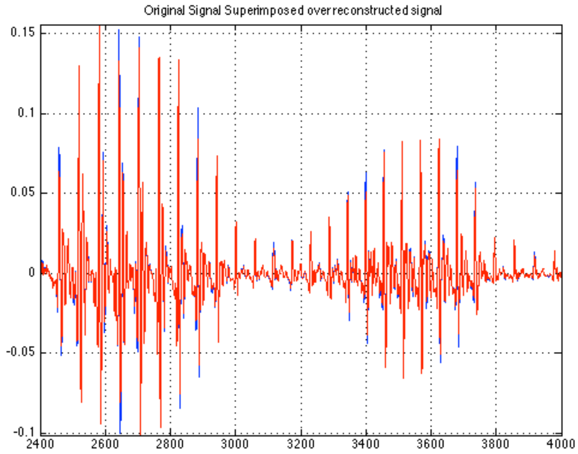


Fig. 2 Superimposing the original signal (blue) and the red signal (red) shows a very accurate reconstruction.

B. *Stylized Voicing using n Samples of e[n] Per Period*

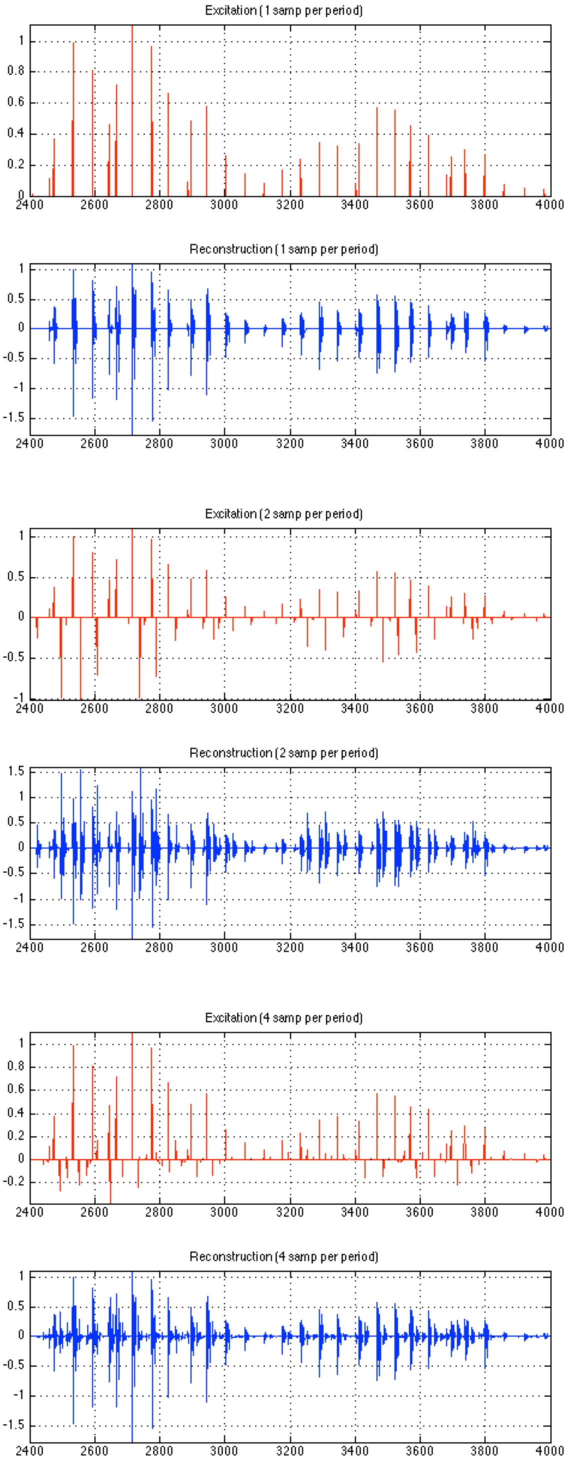


Fig. 3. Stylized Voicings for 1 sample per period (top), 2 samples per period (middle) and 4 samples per period (bottom).

Stylized syntheses using only a couple samples per period of excitation reveal the mechanism in which speech and understanding speech works. When one sample per period of excitation is used, the response to every one of these samples is seen in the reconstruction. This signal is naturally difficult to understand as speech as the excitation cannot be perceived as a pitch, but instead just a series of impulses. When only one signal is being convoluted with the response of the vocal tract, it is difficult to tell what phoneme is being pronounced, compared to a full periodic signal. As more samples are added, the signal visually and audibly sounds more like the original signal. The stylized reconstructions were created by windowing the signal at a size close to the fundamental pitch period. For the male, this fundamental frequency is around 133 Hz. The largest peak, and then n-1 other evenly spaced samples were taken from the original excitation to create the stylized signal.

III. Gender Conversion Utilizing Linear Prediction

The two main aspects which differentiate male and female voices are the fundamental frequency of the voice, and the density of formants and their distribution. The fundamental frequency of a male is around 130 Hz while for females it is 200 Hz. Males have around one resonance per 1000Hz and females have two per 3 kHz. Because both fundamental frequency and formant position change independently of each other, switching the perceived gender of speech is not as easy as speeding up or slowing down the entire recording. Instead, linear prediction must be used once again to split the speech signal into the excitation and the transfer function of the vocal tract. These must be processed separately and then synthesized.

A. *Fundamental Frequency*

The issues associated with the fundamental frequency are that (1) pitch changes with time due to natural inflections of speech and (2), not all phonemes are voiced, meaning there is no periodic excitation at all times. These factors must be taken into account when altering the fundamental frequency. First, the error signal is created using linear prediction. Next, the autocorrelation of the signal is calculated to determine if the frame is voiced, and what the fundamental frequency is. The middle frame of the autocorrelation is used to determine if the frame is voiced, and if it is, the fundamental frequency is altered. Next, the fundamental period is detected by finding the two largest peaks in the frame and finding the distance between them. A new fundamental period is then calculated based on whether the conversion is to male or female. Lastly, the major peaks of the signal are detected from the original error frame, and are placed in a new frame, spaced based on the new fundamental period. This new error signal is then synthesized with the newly calculated linear prediction coefficients.

B. *Formant Position*

The calculation of the linear prediction returns coefficients which approximate the frequency response. Another common way to represent filters is with a pole-zero plot. This is done by finding the roots of all the coefficients. In the case of the linear prediction algorithm used in this project, the filter is represent only with resonances, or poles. In order to change the spacing of the poles, their positions must be decomposed into polar form, and their angle must be changed accordingly. The magnitude however, is maintained. A set of new poles is then constructed using the new angles and old magnitudes, and these are multiplied together to create a set of new linear prediction coefficients.

C. Results

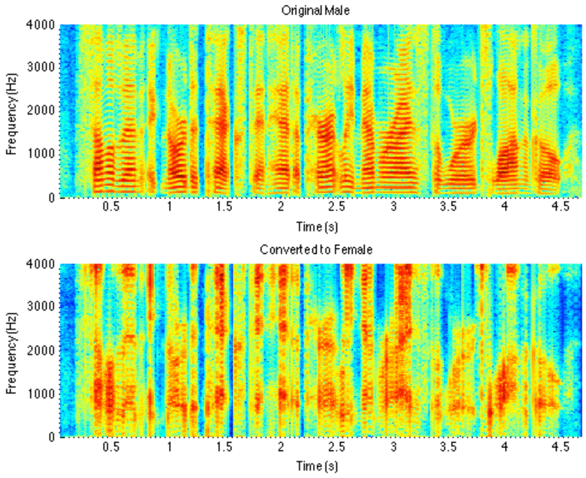
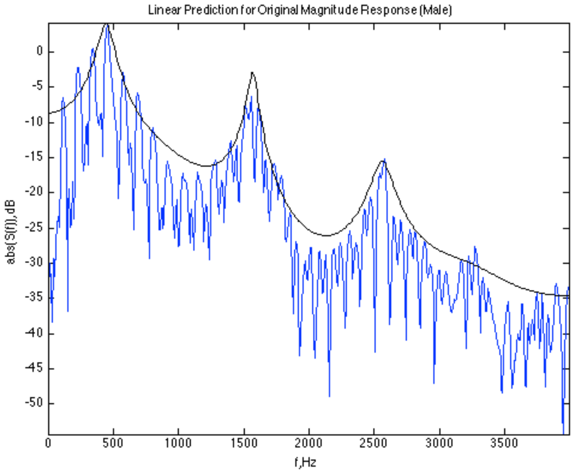


Fig. 4. Spectogram comparing male speech before processing to converted female speech.



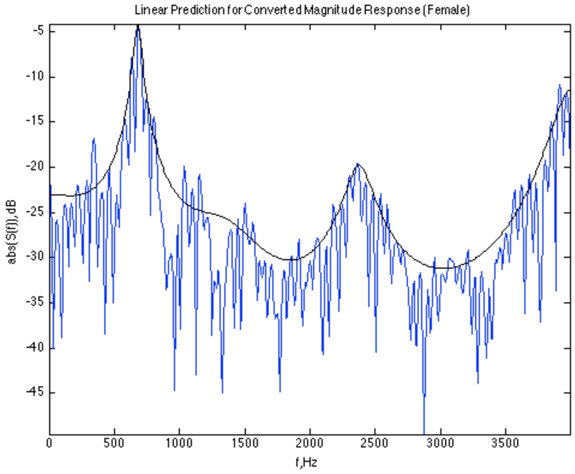


Fig. 5. Linear prediction and magnitude plots for original male speech sample (first) and converted female speech sample (second).

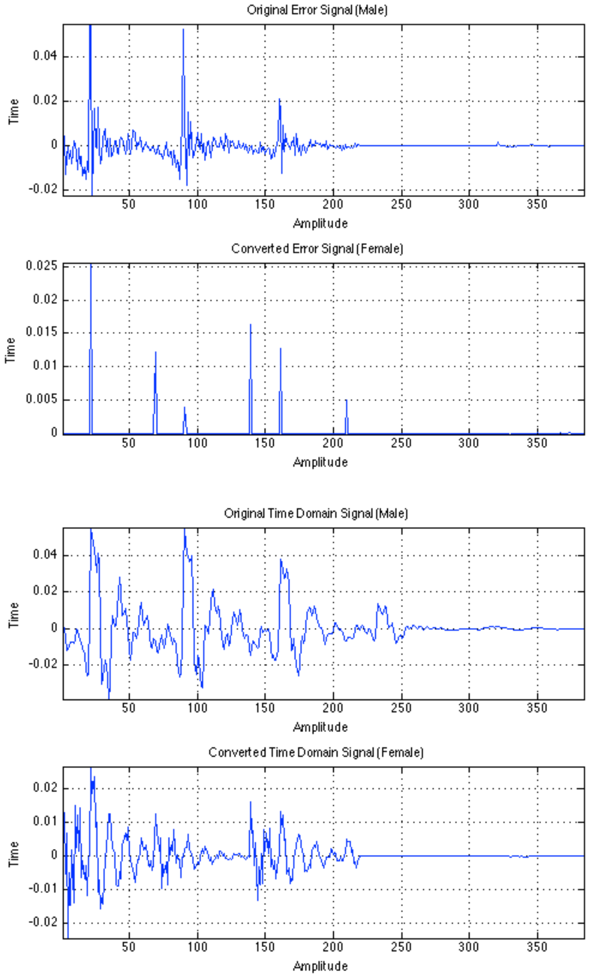


Fig. 5. Plots for unprocessed and processed error signal and reconstructed signals.

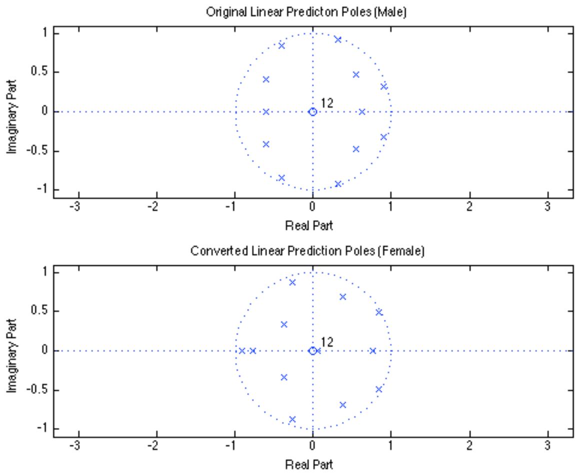


Fig. 6. Pole plot for original male speech sample (top) and female speech sample (bottom).

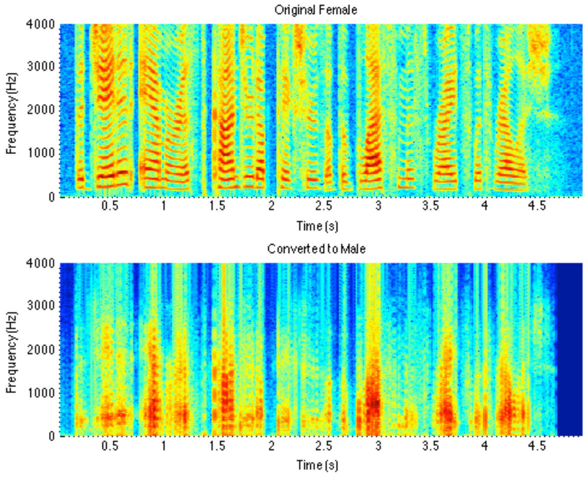


Fig. 7. Spectogram comparing female speech before processing to converted male speech.

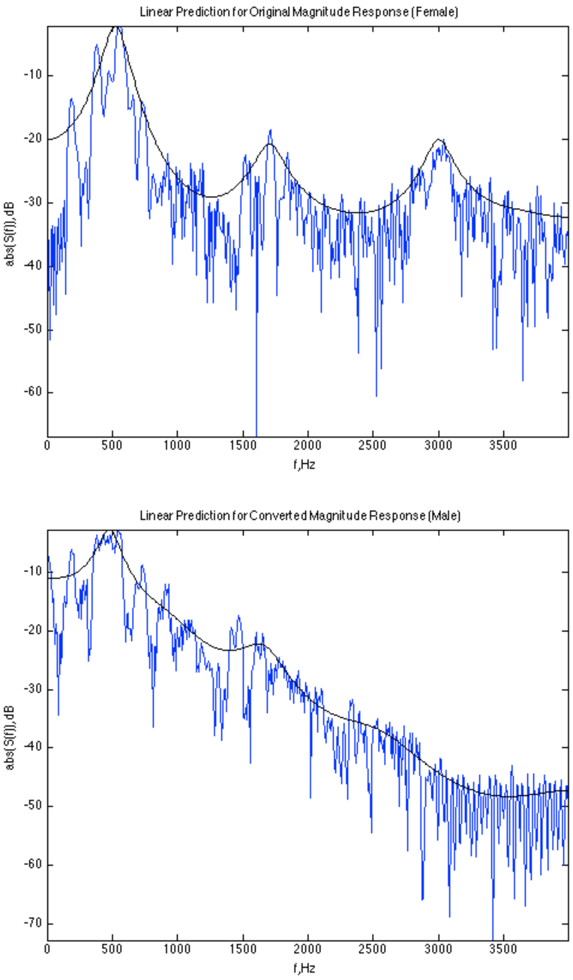


Fig. 8. Linear prediction and magnitude plots for original female speech sample (top) and converted male speech sample (bottom).

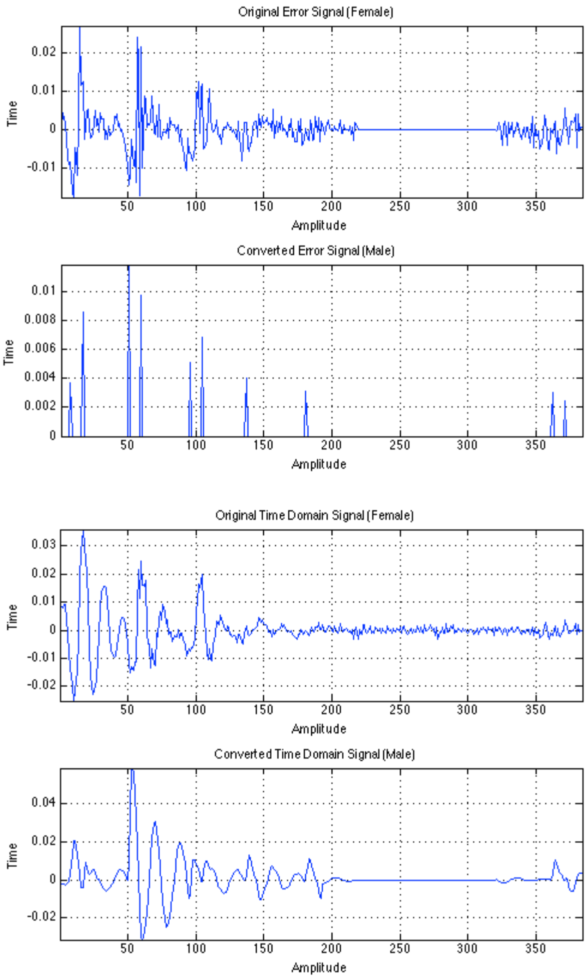


Fig. 9. Plots for unprocessed and processed error signal and reconstructed signals.

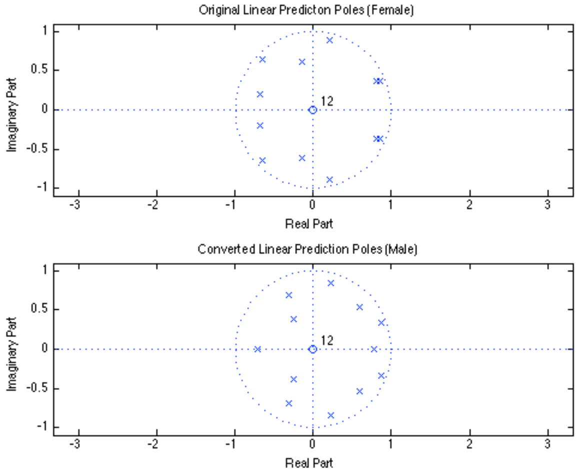


Fig. 10. Pole plot for original female speech sample (top) and male speech sample (bottom).

Both the male to female conversion and female to male conversion were successful in that the words can be somewhat understood, and the voice matches the target gender closer than the previous gender. However, the biggest issue involes lack of clarity and "robotic" nature of the processed speech samples. This is most likely due to periodic high frequency artifacts created in the process, which are different than those that would naturally occur in the voice. The biggest source of error is in the changing of the fundamental period. An issue with changing the fundamental period without speeding up or slowing down the signal is that either audio will have to be removed or silence will have to be added. In this case, a limited number of peak samples are used in the edited excitation and as seen in earlier tests, this results in loss of clarity in the reconstructed signal. In order to prevent speeding up and slowing down and maintain a "human" sounding voice, much more advanced audio signal processing would have to be performed on the error signal.

The formant shifting is also a source of error, as seen from the spectrogram, linear prediction, and pole plots. This is possibly because while shifting a few formants works in theory, linear prediction calculates a larger number of formants and shifting all of these can result in undesired behavior. Both conversions show a loss of spectral clarity in the processed versions, where clear formants are much more difficult to observe. However, the shifting was successful in changing the overall distribution of energy, with conversion to male leaving more spectral energy in the low range, and conversion to female leaving more spectral energy in the higher range. To avoid loss of spectral clarity, only select poles could be shifted, or the number of poles calculated in the linear prediction could be reduced. Also, a final source of error may have been in the windowing, which if there are any errors would create harmonic artifacts.

IV. Conclusion

This project displayed the usefulness of linear prediction in speech processing, in order to analyze and edit the vocal excitation and vocal tract transfer function independently of one another. The biggest lesson learned from this project is the sensitivity of the human ear to "unnatural" artifacts in human speech. It is very easy to hear when a glottal pulse is unnatural, even when filtered by the proper vocal tract transfer function. Also, the shifting of formants will create a drastic perceptual change in the characteristics of the audio. While creating a basic model for speech processing using linear prediction is relatively simple, advanced processing that maintains the "natural" sound of human speech would require much more advanced techniques.