Continuous Speech and Time-Frequency Transform Using the Kalman Filter

Mario Barnard
Department of Electrical and Computer Engineering
Oakland University
Rochester, MI, USA
Email: mmbarna2@oakland.edu

Mohamed Zohdy
Department of Electrical and Computer Engineering
Oakland University
Rochester, MI, USA
Email: zohdyma@oakland.edu

***Abstract*—In this paper, a Radial Basis Function-based Kalman filter has been utilized to in order to extended to the time-frequency transform, also called a spectrogram or spectrograph, and also been applied to simple continuous speech.**

Keywords*— *Kalman Filter, Radial Basis Function, Speech Recognition, Time-Frequency Transform, Continuous Speech

1. INTRODUCTION

This paper attempts to expand upon the fused multi-sensor data using Kalman filter [1] and speech enhancement and recognition using the Kalman filter modified via the radial basis function (RBF) [2] in order to include simple continuous speech and time-frequency analysis. The purpose of this paper is to take the concept of the Kalman filter modified with the radial basis function that was developed [2] and to expand that to continuous speech and the time-frequency transform. The time-frequency transform is also known as time-frequency analysis. The graph of the time-frequency analysis is called a spectrogram. A spectrogram is a visual representation of an audio signal with respect to the frequency spectrum and how those frequencies vary with time. The x-axis denotes time in seconds (s) and the y-axis denotes frequency in hertz (Hz). As a side note, males often speak in the 65 Hz to 260 Hz range, while females speak in the 100 Hz to 525 Hz range. Thus, the speech frequency range from about 100 Hz to 260 Hz is just as "masculine" as it is "feminine."

2. ORIGINAL DATA

A word bank was setup using audio recordings. [2] Audio signals such as “Hello”, “Estimation”, and “Oakland” were recorded with a single microphone. The time-domain plots of the signals are shown in Figures 1-3. The frequency-domain plots of the signals are shown in Figures 4-6.

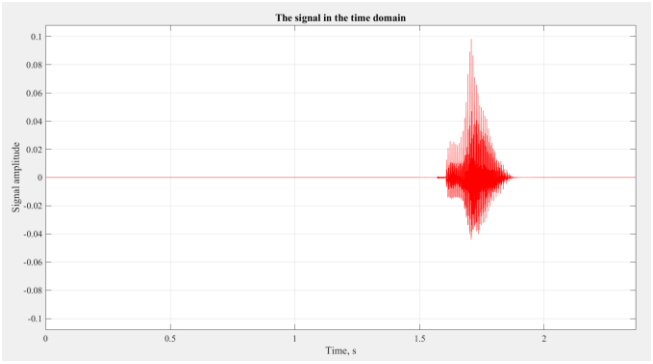

Figure 1: Time-Domain of “Hello”

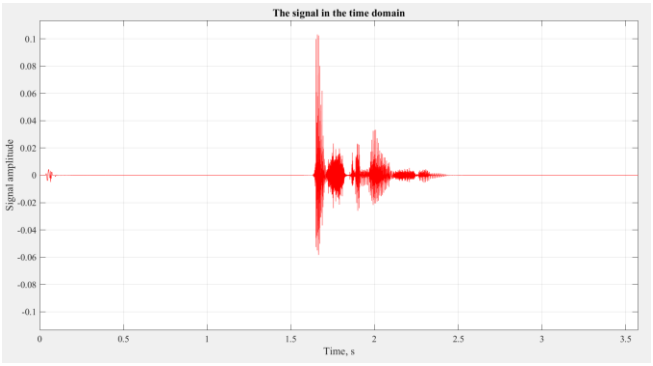

Figure 2: Time-Domain of “Estimation”

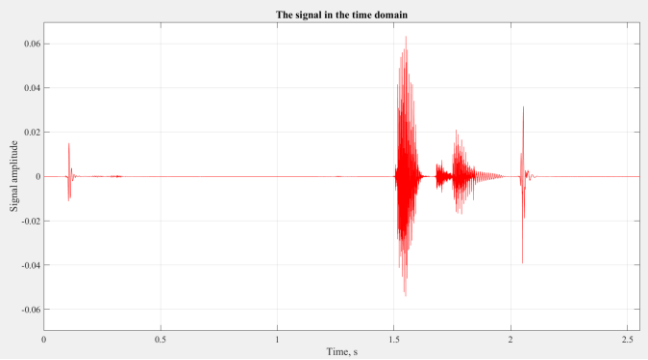

Figure 3: Time-Domain of "Oakland"

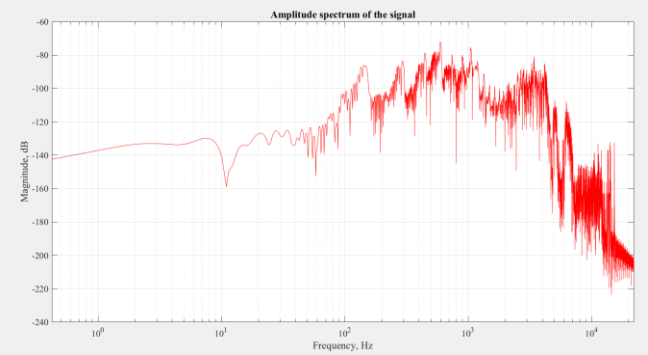

Figure 4: Frequency-Domain of "Hello"

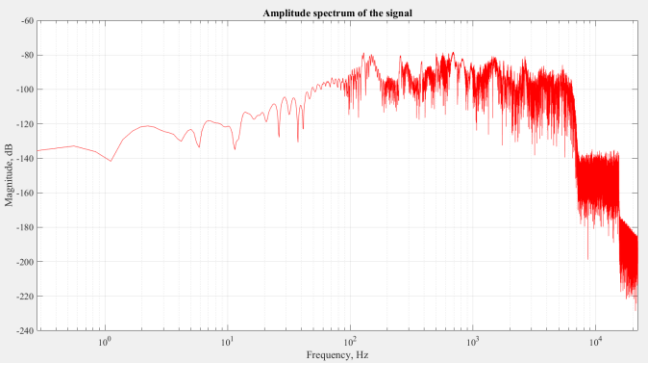

Figure 5: Frequency-Domain of "Estimation"

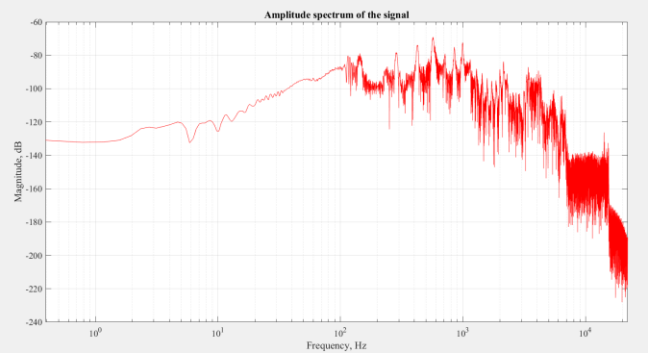

Figure 6: Frequency-Domain of "Oakland"

3. TIME-FREQUENCY TRANSFORM

The time-frequency plot allows the time-domain plot and the frequency-domain plot to be shown in a single spectrogram. The time-frequency plots of the audio signals, "Hello", "Estimation", and "Oakland", are shown in the Figures 7-9.

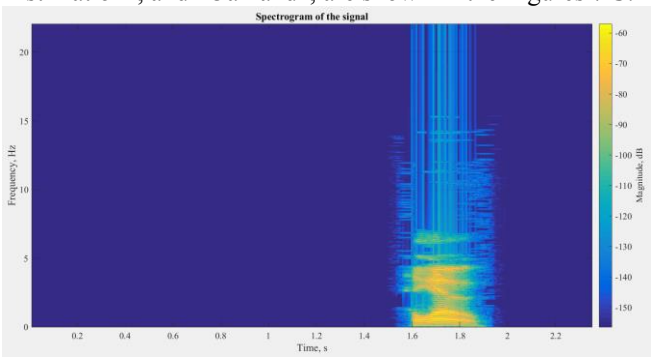

Figure 7: Time-Frequency Domain of "Hello"

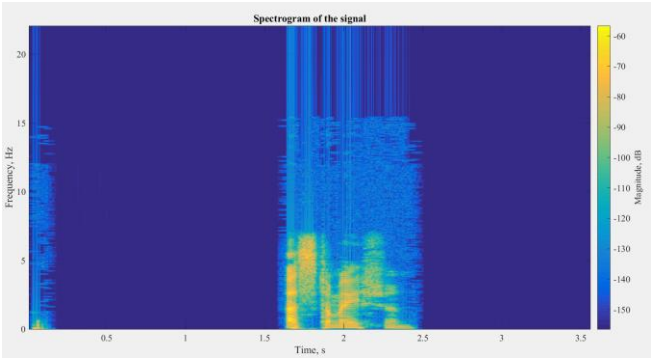

Figure 8: Time-Frequency Domain of "Estimation"

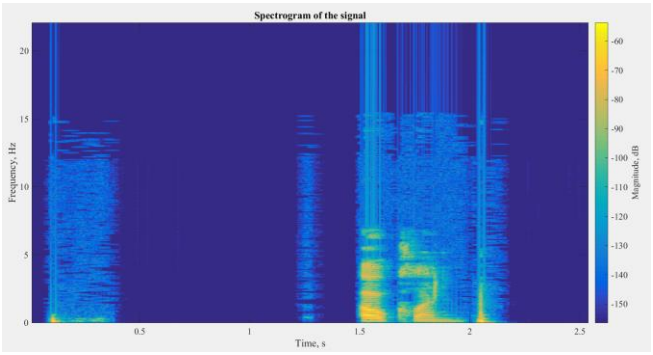

Figure 9: Time-Frequency Domain of "Oakland"

4. CONTINUOUS SPEECH

Four different stages were used for continuous speech analysis. Stage 1 consisted of one isolated word ("Hello"). Stage 2 consisted of two isolated words ("Hello … (short pause)… Estimation"). Stage 3 consisted of two consecutive

words ("Hello…Estimation"). Stage 4 consisted of more than two consecutive words ("Hello…Estimation…Oakland"). The time-domain plots are shown in Figures 10-13. The frequency-domain plots are shown in Figures 14-17. The time-frequency plots are shown in Figures 18-21.

The performance decrease is due to insufficient training data and the noisy nature of the recordings. In this work we deal with the problem of noisy observations through a time-inhomogeneous dynamical system formalism, including observation noise. Under the assumption that we model speech as a Gaussian process at the frame-rate level, a linear state-space dynamical system can be used to parameterize the density of a segment of speech. [3]

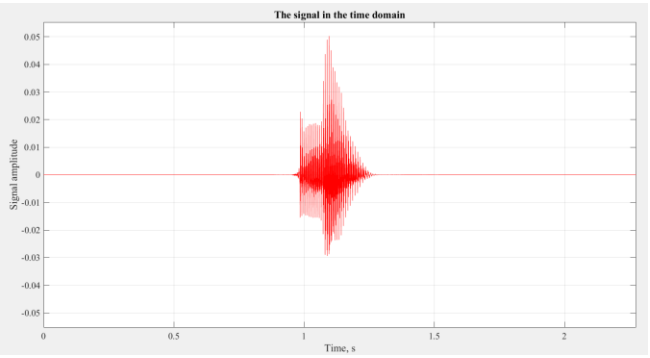

Figure 10: Time-Domain of "Hello"

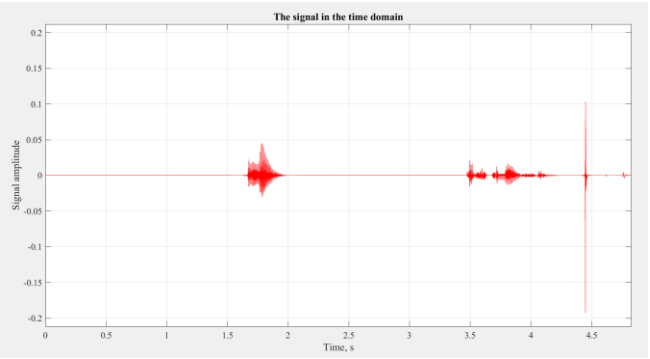

Figure 11: Time-Domain of "Hello … (short pause)… Estimation"

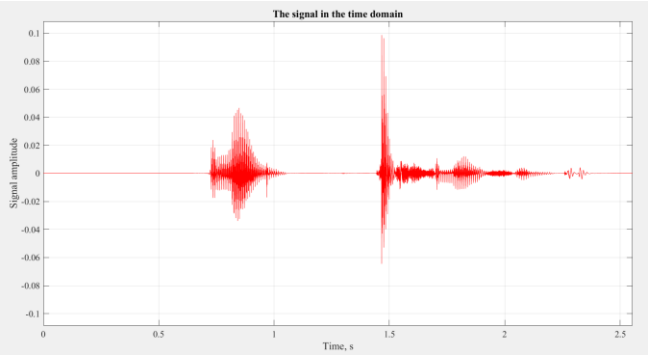

Figure 12: Time-Domain of "Hello…Estimation"

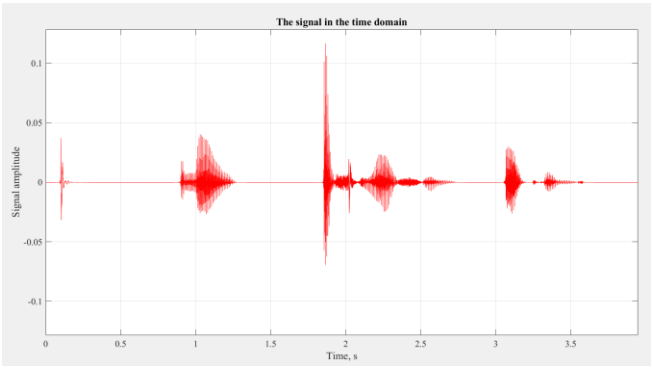

Figure 13: Time-Domain of "Hello…Estimation…Oakland"

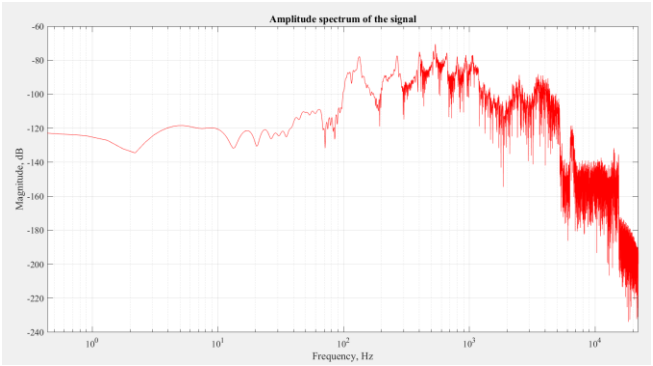

Figure 14: Frequency-Domain of "Hello"

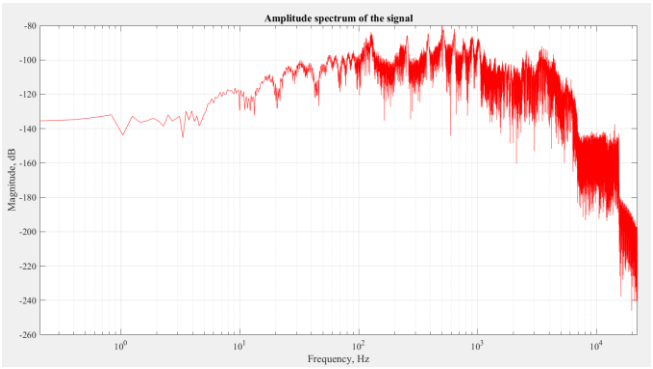

Figure 15: Frequency-Domain of "Hello … (short pause)… Estimation"

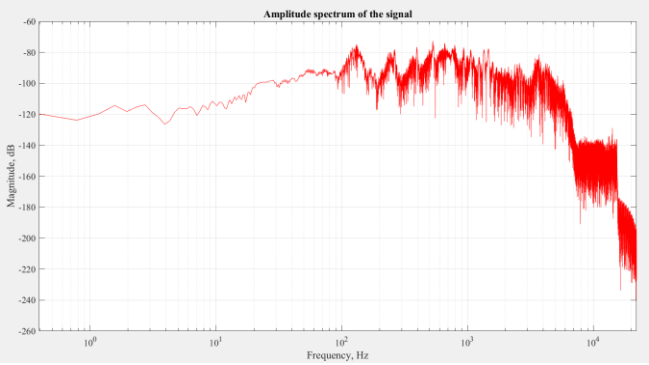

Figure 16: Frequency-Domain of "Hello…Estimation"

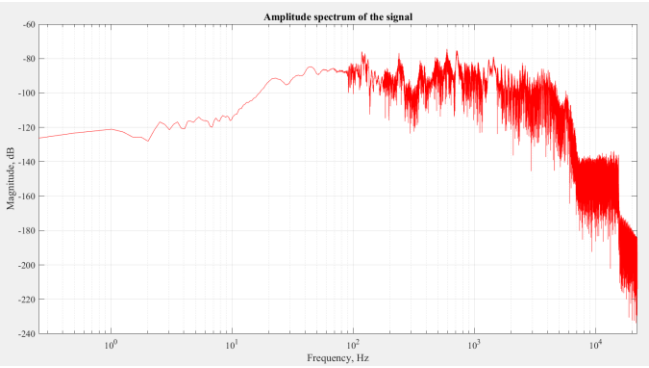

Figure 17: Frequency-Domain of "Hello…Estimation…Oakland"

The spectrograms that are shown in Figures 18-21 were generated using MATLAB.

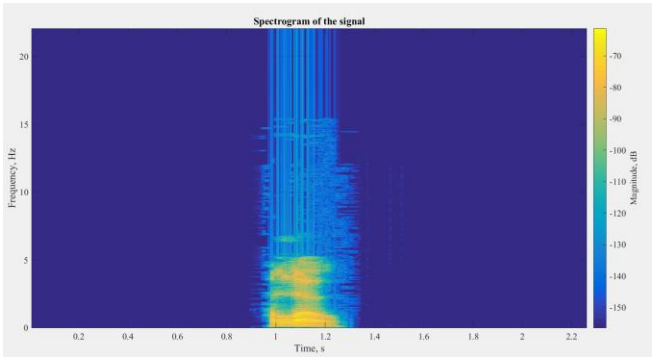

Figure 18: Time-Frequency Domain of "Hello"

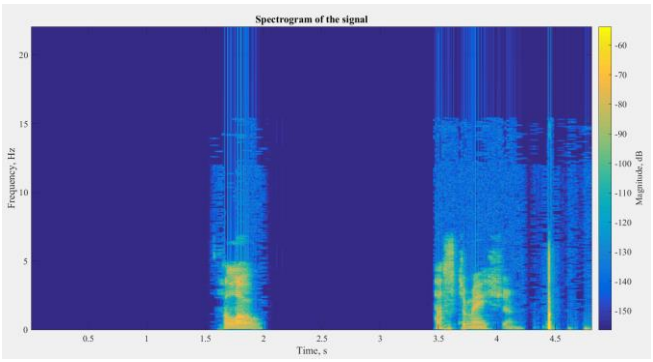

Figure 19: Time-Frequency Domain of "Hello … (short pause)… Estimation"

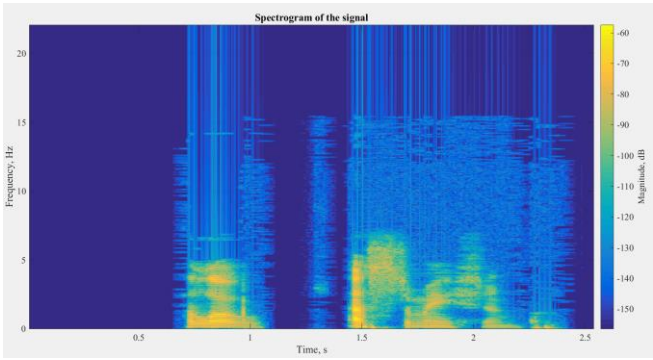

Figure 20: Time-Frequency Domain of "Hello…Estimation"

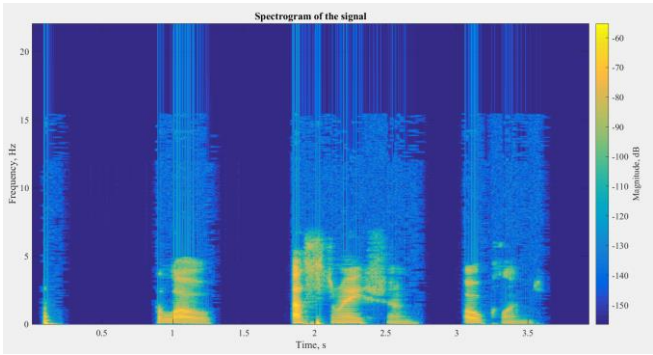

Figure 21: Time-Frequency Domain of "Hello…Estimation"

5. RESULTS

Speech enhancement using a single microphone system has become an active research area for audio signal enhancement. The aim is to minimize the effect of noise and to improve the performance in voice communication systems when input signals are corrupted by background noise. There are various filtering techniques for speech enhancement like spectral subtraction, signal subspace , Wiener filtering, and Kalman filtering. On analysis of SNR values using colea (a MATLAB signal processing tool) we observed that these techniques have some drawbacks and are not efficient compared to adaptive Kalman filtering. A Kalman filter is simply an optimal recursive data processing algorithm. There are many ways of defining optimal, dependent upon the criteria chosen to evaluate performance. One aspect of this optimality is that the Kalman filter incorporates all information that can be provided to it. To overcome the drawback of conventional Kalman filtering for speech enhancement, this algorithm only constantly updates the first value of state vector X(n), which eliminates the matrix operations and reduces the time complexity of the algorithm on it. It is difficult to know what the environmental ambient noise consists of and the effect that the noise on the Kalman filtering algorithm application. In addition to the Kalman filtering algorithm, a real-time adaptive algorithm can be used to estimate the ambient noise to be filtered out for processing. [4]

Not only the ambient noises were considered when the authors' were recording audio data, but also the quality of the microphone must to be considered. Thus, the noise from multiple sources needed to be filtered out.

The continuous speech recordings required editing of the signal to omit some noises that were present in the audio signals. Figures 22-25 show the edited signals.

The following are the Kalman filter equations that were used to analyze the audio recordings from [2].

$$X(k) = \varphi x(k-1) + Gu(k) \quad (1)$$

where the dimension of X(k) matrix is the (p×1) state vector matrix, while the dimension φ is the (p×p) state transition matrix that uses LPCs calculated from noisy speech according to 1.8, G is the (p×1) input matrix and u(k) is the noise corrupted input signal at the kth instant. When speech is corrupted with noise, then the output y(k) is given as:

$$y(k) = x(k) + w(k) \quad (2)$$

where w(k) is the measurement noise, a zero-mean Gaussian noise with variance σ^2_w.

In vector form, this equation may be written as the following:

$$y(k) = Hx(k) + w(k) \quad (3)$$

where, H is the observation matrix with a dimension pf (1×p), which is given by

$$H = (0\ 0 \cdots 0\ 1) \quad (4)$$

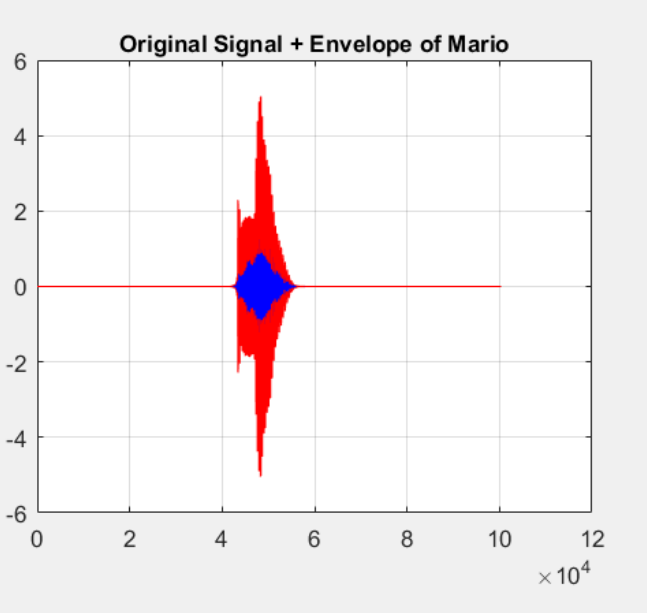

Figure 22: Edited Signal of "Hello"

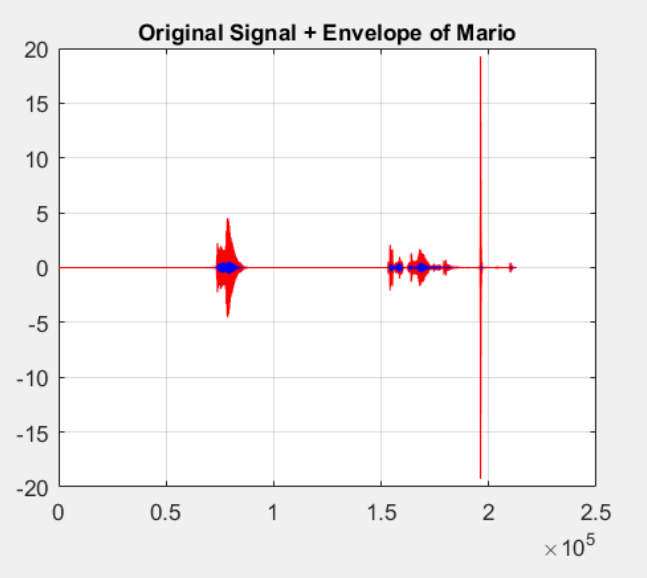

Figure 23: Edited Signal of "Hello … (short pause)… Estimation"

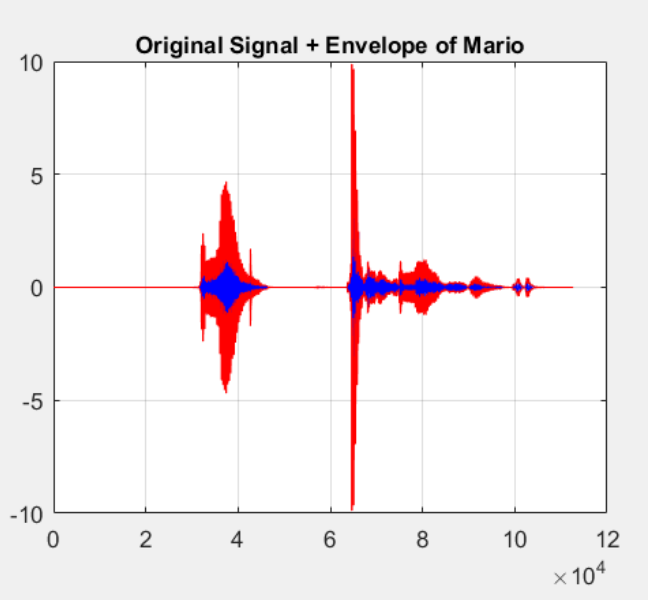

Figure 24: Edited Signal of "Hello…Estimation"

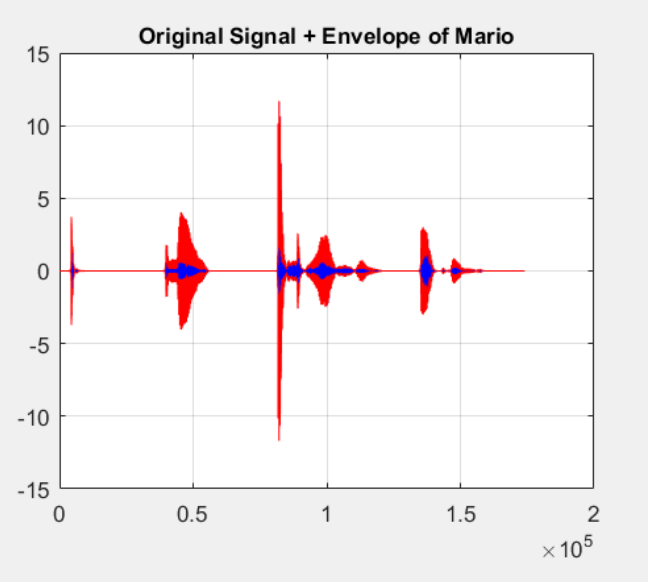

Figure 25: Edited Signal of "Hello…Estimation…Oakland"

The following is the author speaking the various phrases showing the single sided magnitude spectrum in Figures 26-29.

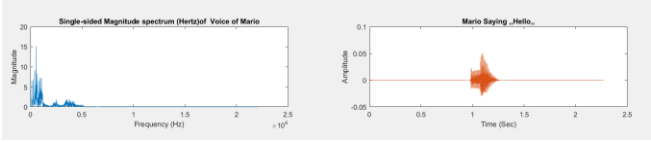

Figure 26: Single Sided Magnitude Spectrum of "Hello"

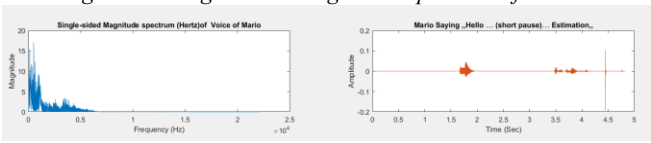

Figure 27: Single Sided Magnitude Spectrum of "Hello … (short pause)… Estimation"

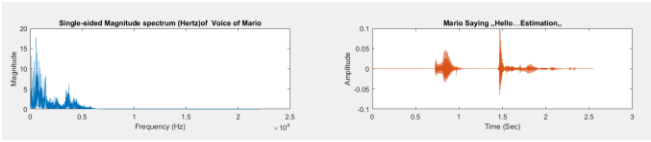

Figure 28: Single Sided Magnitude Spectrum of "Hello…Estimation""

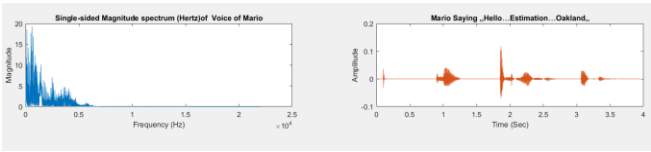

Figure 29: Single Sided Magnitude Spectrum of "Hello…Estimation…Oakland"

6. Conclusion

Starting words such as “Hello”, “Estimation”, and “Oakland” were used as a selection of words for starting the word bank. Continuous speech was best implemented as a single word, and then added consecutive words with and without pauses in speech. It seems that the Kalman filter provides decent filtering of the noise and duration of the silent audio during pauses in speech for continuous speech. For future work on this topic, the authors’ would implement an Extended Kalman filter (EKF) and an Unscented Kalman filter (UKF) in order to gain better results as speech is nonlinear. Other languages would also be investigated as well.

References

[1] M. A. Zohdy, Aftab Ali Khan, Paul Benedict, "Fused multi-sensor data using a Kalman filter modified with interval probability support”, American Control Conference, June 1995.

[2] Mario M. Barnard, Farag M. Lagnf, Amr S. Mahmoud, M. A. Zohdy, “Speech Enhancement and Recognition using Kalman Filter modified via Radial Basis Function”, Oakland University December 2016.

[3] V.Digalaki, J.R. Rohlieek, M. Ostendor, “A Dynamical System Approach to Continuous Speech Recognition”, [Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing, Toronto, Ont., 1991, pp. 289-292 Vol. 1. doi: 10.1109/ICASSP.1991.150334.

[4] Prof. M.V. Ramanaiah, N. Sirisha, P. Ravali, B. Vinay Singh, T. Thirupathi ‘Single Channel Adaptive Kalman Filtering-Based Speech Enhancement Algorithm” International Journal of Advanced Research in Electrical, Electronics, and Instrumentation Engineering, Vol. 4 Issue 4, April 2015.