NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

Electrical and Electronics Department

EE313: Digital Signal Processing Laboratory

Kalman Filter in Speech Enhancement

Project Report

Shraddha Smriti 171EE140
Chaitali Shah 171EE212
B. Alagu Parvathy 171EE251

Contents

Abstract

Speech signal processing is among the important sub-domains of digital signal processing. Speech, being the primary mode of communication among humans, has been the subject of extensive research and analysis in the past few decades. Among the main challenges faced in speech processing, is the extraction of a clear speech signal from a corrupted signal. Every measured signal has the presence of unwanted components called noise which affect the quality of the signal. Therefore, to improve the intelligibility, pleasantness and ability of the signal to be processed further, the noise component of the signal has to be eliminated. Speech is a non-deterministic and non-stationary signal and estimation of the original uncorrupted signal becomes an uphill task. Various algorithms in the time and frequency domain have been developed in the past years to deal with this dilemma.

The objective of the project is enhancement of speech signal using the very famous adaptive filter, the Kalman filter. Using a linear auto regressive model for the speech signal, the state equations and matrices of a given signal can be found. The Kalman filter parameters are found and tuned to estimate the original signal and the noise is eliminated. The effectiveness of the algorithm is checked by listening tests and spectral analysis.

1 Introduction

Speech enhancement includes techniques and algorithms that ameliorate the quality of a given speech signal. This is primarily done by reduction or elimination of noise. Noise can be of various types – periodic, wideband, interference, impulsive et cetera. Speech enhancement is important because every recorded speech has some noise, either due to lack of sophisticated measuring equipments or presence of unintended background sounds. Separation of this random noise signal from the speech signal is among the major theoretical and practical challenges in speech processing. The speech is non- stationary which means that its statistical properties like mean, covariance, et cetera are not constant but rather change with time. It is important to derive an appropriate model of the signal for processing and enhancement. Speech enhancement has many uses such as in mobile phones, telecommunication, hearing aids or in pre-processing of the signal to make it suitable for further processing like speech recognition.

The speech signal is corrupted with white noise. White noise is the presence of unwanted components at all the frequencies in a spectrum. The Signal to Noise ratio(SNR) is known. Kalman filter, described by recursive equations, is used to estimate the signal based on the state equations obtained by assuming an auto regressive(AR) model. Kalman filter was chosen because it is an optimum minimum mean square error estimator and has been shown to perform better than previously used techniques.

2 Literature Survey

Speech enhancement has been done by various techniques in the past. Noise was usually an additive slowly varying signal and hence could be removed by eliminating low frequency components of the signal. Spectral elimination was among the most used techniques in the past [1]. Since this does not work for noise with higher frequency and may even lead to loss of the original signal, various other techniques were explored like stationary adaptive filters like Wiener filter, Kalman filter and transform domain filter (DFT based, signal subspace method) [2]. The Kalman filter was proposed by Rudolf Kalman in the 1960s as a solution to foresee unknown states of dynamic systems [3]. Later it was proposed that speech signals could be represented as time-varying parameters related to the transfer function of the vocal tract and characteristics of the excitation [4]. This made it possible to find the state representation of a speech signal. Utilizing the AR model thus obtained, Kalman filter was used in speech enhancement and the results were compared with other techniques like the Wiener filter [5]. Various algorithms were explored and modified to tune the filter to give better performance. The measurement covariance matrix R of the filter was calculated by Yule Walker Equations [6] [7] , by spectral energy of frames in the signal [8] and also by Power Density Spectrum [9]. The process covariance matrix Q was calculated by defining two performance metrics - sensitivity and robustness metric [10] [11]. In this project, R is found by Power Density Spectrum approach [11] and Q is found using the above mentioned performance metrics [8].

3 Methodology

3.1 Autoregressive model for speech: Mathematical formulation

Autoregressive model specifies that the output variable depends linearly on its own previous values. The speech is modelled as output of linear time-varying filter of order p. This is done by Linear Prediction of speech where in, the speech is defined as the output of an all-pole filer driven by white noise sequence. The value of the signal at an instant k depends on linear combination of previous p values along with the random noise as shown below:

$$\begin{aligned} x(k) &= -a_1x(k-1) - a_2x(k-2) - .... - a_px(p) + u(k) \\ &= -\sum_{i=1}^{p} a_ix(k-1) + u(k) \end{aligned} \tag{1}$$

where the a_ks are the linear prediction coefficients and u(k) is the driving white noise sequence. On inspection the state-space model can be found.

$$\begin{bmatrix} x(k-p+1) \\ x(k-p+2) \\ . \\ . \\ x(k) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & . & . & 0 \\ 0 & 0 & 1 & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ -a_p & -a_{p-1} & -a_{p-2} & . & . & a_1 \end{bmatrix} \begin{bmatrix} x(k-p) \\ x(k-p+1) \\ . \\ . \\ x(k-1) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ 1 \end{bmatrix} u(k) \tag{2}$$

or

$$\mathbf{X}(k) = \mathbf{A}\mathbf{X}(k-1) + \mathbf{G}u(k) \tag{3}$$

where $\mathbf{X}$(k) represents the state matrix of order $(p*1)$, $\mathbf{A}$ is the state transition matrix (also known as the system matrix) calculated using the LPC and is of order $(p*p)$, $\mathbf{G}$ is the $(p*1)$ order input matrix (also known as control matrix) and u(k) is the input signal at the instance k.

From the state space equations (2),

$$\mathbf{X}^T = \begin{bmatrix} x(k-p+2) & x(k-p+2) & . & . & x(k) \end{bmatrix} \tag{4}$$

$$\mathbf{A} = \begin{bmatrix} 0 & | & I & \\ -- & -- & -- & -- \\ -a_p & . & . & -a_1 \end{bmatrix} \tag{5}$$

where I is an identity matrix of order $(p-1*p-1)$.

The corrupted signal is described as

$$y(k) = x(k) + n(k) \tag{6}$$

where $n(k)$ is the additive noise component in the measurement.

The state equations in matrix form is as follows.

$$y(k) = \mathbf{CX}(k) + n(k) \tag{7}$$

where **X**(k) is the state vector from equation (4) and **C** is the observation matrix defined as follows.

$$C = \begin{bmatrix} 0 & 0 & 0 & . & . & 1 \end{bmatrix} \tag{8}$$

3.2 Calculation of a_ks: Linear Prediction Coefficients

The signal is divided into frames of 40 ms each with 10 ms overlap. The Linear Prediction coefficients are found for each segment using the MATLAB function **lpc**. **[Alpha,Q1] = lpc(X,p)** finds the coefficients of a pth-order linear predictor. It returns the present value of the time series X based on past samples. It also finds Q0, the variance of the prediction error.

The Alpha found by the function is the following matrix.

$$\alpha = \begin{bmatrix} 1 & a_1 & a_2 & . & . & a_p \end{bmatrix} \tag{9}$$

The values of the coefficients are used in the last row of the matrix **A** defined in equation (5).

3.3 Measurement Noise Covariance R

3.3.1 Auto Correlation Function

An auto correlation function (ACF) is a measure of likeness in two points separated at a lag. The MATLAB function **xcorr** is used to find the ACF. **[c,l] = xcorr(x)** gives the autocorrelation sequence of x along with the lag, l. Theoretically, the ACF R_xx at lag l, is given by,

$$R_{xx}(l) = E[x(k)x(k-l)] \tag{10}$$

where E[x] is the mathematical expectance or expectation, which is essentially the mean. For the signal given, the ACF is related to the Linear Prediction coefficients as follows:

$$\begin{bmatrix} R_{xx}(0) & R_{xx}(-1) & . & . & R_{xx}(1-p) \\ R_{xx}(1) & R_{xx}(0) & . & . & R_{xx}(2-p) \\ . & . & . & . & . \\ R_{xx}(p-1) & R_{xx}(p-2) & . & . & R_{xx}(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ a_p \end{bmatrix} = - \begin{bmatrix} R_{xx}(1) \\ R_{xx}(2) \\ . \\ R_{xx}(p) \end{bmatrix} \tag{11}$$

3.3.2 Power Spectral Density

The power spectral density (PSD) is the Fourier transform of the auto correlation function and is given by

$$S(f) = \int_{-\infty}^{+\infty} R_x x(\tau) e^{-2\pi j f \tau} d\tau \tag{12}$$

White noise, contains components of all possible frequencies. Hence the power spectrum density plot will contain all the frequency. On the other hand, if we had a pure signal of single frequency, the power spectrum, would contain a sharp spike at that particular frequency, just like its frequency spectrum. On the other hand white noise will contain a fairly flat, that is, constant power spectrum that indicates the presence of all frequencies.

Hence when we plot the power spectral density of the noisy signal which contains some silent parts and some voiced parts, the silent parts will show a flat plot where as the voiced parts will show peaks at the fundamental frequency and its harmonics. To clearly distinguish between voiced and silent frames, the ratio of the geometric mean to the arithmetic mean of the power spectrum is taken .

$$SpectralFlatness(SF) = \frac{\sqrt{\prod_{n=0}^{N-1} x(n)}}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)} \tag{13}$$

where x(n) = magnitude of the nth bin in the power spectrum.

White noise will have a SF of 1 and the pure audio will have a SF of 0. These are the two extreme cases. The unfiltered speech that we use will have a SF between 0 and 1.

3.3.3 Calculating R

For each of the frames of the signal, Power spectral Density (PSD) is calculated. The PSD is trimmed to have values in the range [0.1 kHz, 2 kHz] frequency as that is the range of human speech. A cut-off value (co=$\frac{1}{\sqrt{2}}$) is chosen. The frames are divided into voiced and silent depending on if their SF values are above or below co.The maximum variance of all silent frames is R.

3.4 Process Noise Covariance Q

Process Noise Covariance Q is estimated from the noise that is caused due to the processing model. To calculated the same, we define two matrices J_1 and J_2 which represents the sensitivity and robustness of the system model respectively. Two values of Q's are obtained. First, Q_c for the voiced frames and Q_2 (less than Q_c) for silent frames.
To calculate J_1 and J_2, we define two scalar quantities A_k and B namely the present state estimation error covariance and the process noise covariance in the measured output respectively, at the instance k. Also the scalar value R is the same constant value for all frames. B is a constant for one particular frame.

$$A_k = \mathbf{C}(\mathbf{A}\mathbf{P}(k-1|k-1)\mathbf{A}^T)\mathbf{C}^T \tag{14}$$

$$B = \mathbf{C}(\mathbf{G}Q\mathbf{G}^T)\mathbf{C}^T = \sigma_u^2 = Q_f \tag{15}$$

We can calculate J_{1k} and J_{2k} for any instant K in a particular frame as follows:

$$J_{1k} = [(A_k + B + R)^{-1}R] = \frac{\sigma_w^2}{A_k + \sigma_u^2 + \sigma_w^2} \tag{16}$$

$$J_{2k} = [(A_k + B)^{-1}B] = \frac{\sigma_u^2}{A_k + \sigma_u^2} \tag{17}$$

where σ_w^2 is the variance of the noise.

The overall filter performance parameters J_1 and J_2 are estimated as the mean value of J_{1k} and J_{2k}.

$$J_1 = \frac{1}{N}\sum_{k=1}^{N} J_{1k} \tag{18}$$

$$J_2 = \frac{1}{N}\sum_{k=1}^{N} J_{2k} \tag{19}$$

Since J_1 defines inconsistency between actual R of the measurement and assumed R, it is called the sensitivity metric. Similarly, the metric J_2 defines the inconsistency between assumed process noise covariance σ_u^2 and the actual process noise covariance which is due to error in modelling,it is called robustness metric.

The process noise $Q_f = (\sigma_u^2)$ is estimated for different combination of J_1 and J_2. The Q_f which offers the best robustness and sensitivity (this happens when the values of J_1 and J_2 intersect) to the filter is chosen as Q_c or Q_2 depending on the frame that is being considered.

3.5 Kalman Filter Equations

The Kalman filter is designed by applying the Kalman filter equations to AR model. The equation (3) is known as the State Estimation equation and equation (7) is called the Measurement Equation. Kalman filter can be applied to find a minimum mean squared error (MMSE) estimate of the state vector. This estimate can be denoted as $\hat{X}(k|k)$, also known as the a posteriori state estimate. The corresponding MSE covariance matrix of the estimate is **P**(k), the a posteriori error covariance matrix. The one step estimate **X**(k) is (k—k), the a priori state estimate. Its error covariance matrix is **P**(k—k-1), the a priori error covariance matrix.

$$\hat{X}(k|k) = \hat{X}(k|k-1) + \mathbf{K}(k)(y(k) - (C)\hat{X}(k|k-1)) \tag{20}$$

$$\hat{X}(k|k-1) = \mathbf{A}\hat{X}(k-1|k-1) \tag{21}$$

$$\mathbf{P}(k|k) = (\mathbf{I} - \mathbf{K}(k)\mathbf{C})\mathbf{P}(k|k-1) \tag{22}$$

where,

$$\mathbf{K}(k) = \mathbf{P}(k|k-1)\mathbf{C}^T(\mathbf{C}\ \mathbf{P}(k|k-1)\mathbf{C}^T + R)^{-1} \tag{23}$$

and

$$\mathbf{P}(k|k-1) = \mathbf{A}\mathbf{P}(k-1|k-1)\mathbf{A}^T + \mathbf{G}Q\mathbf{G}^T \tag{24}$$

The equation (20) is the updated equation for the state estimate. Equation (21) projects the state estimate into the next time interval. (21) is the update equation for error covariance. The mean square errors are contained in the diagonal elements of the covariance matrix. The Kalman Gain **K**(k) is calculated by the equation (22). Equation (23) is used to project the error covariance matrix to the next time interval.

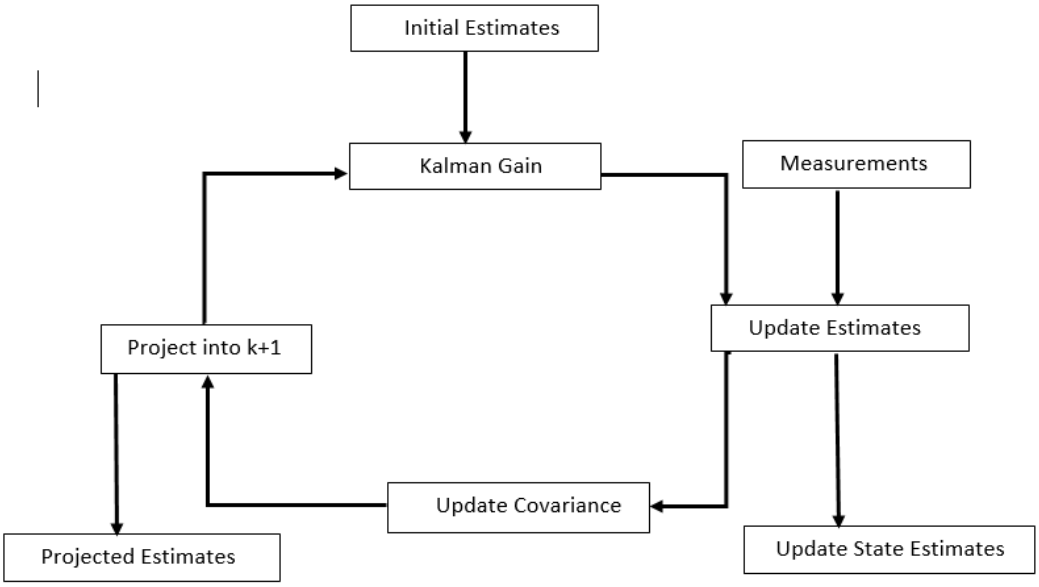

Figure 1: Algorithmic loop of Kalman filter

Kalman Gain has a direct correlation with Q. A high value of Kalman gain shows the dependency of previous estimate more on noisy input, on the other hand low value of gain indicates previous estimate more on the present estimate. Since we need more information to be taken from the voiced frame as it contains the original signal, the Kalman Gain for such frames is kept high. Likewise, the gain for silent frames is kept low so as to minimize the information taken from these frames.

3.6 Error Calculations

To measure the quality of the enhanced audio, and to compare it to the original clean speech, we are using the segmental SNR method. The difference between the segmental SNR of noisy and enhanced speech is the basis to evaluate the performance of the filter. The segmental SNR is given by:

$$SegSNR = \frac{1}{N}\sum_{i=1}^{N} 10\log_{1}{}_{0}[\frac{\sum_{n\in frame_k}|s(n)|^2}{\sum_{n\in frame_k}|\hat{s}(n)-s(n)|^2}] \tag{25}$$

Segmental SNR is expressed in decibels (dB) and a higher value of segmental SNR indicates that more signal component is present than the noise component.

4 Simulation and Experimentation

The code is implemented in MATLAB R2019a Version and the following results were obtained. Tabulated results for different SNR and order are given below. The optimum results for value of SNR have been highlighted.

SNR(dB)	Order	Seg SNR Noisy(dB)	Seg SNR Processed(dB)
SeaGreen 0	8	-14.1474	-1.4257
0	10	-14.1474	-1.8244
0	13	-14.1474	-2.2529
0	15	-14.1474	-2.6412
0	15	-14.1474	-3.2718
0	18	-14.1474	-3.5592
0	20	-14.1474	-3.9632
5	8	-11.6474	-
5	10	-11.6474	-
JungleGreen 5	13	-11.6474	-0.7324
5	15	-11.6474	-0.9583
5	15	-11.6474	-1.6381
5	18	-11.6474	-1.8997
5	20	-11.6474	-2.2257
10	8	-9.1474	-
10	10	-9.1474	-
ForestGreen 10	13	-9.1474	1.0530
10	15	-9.1474	0.7237
10	15	-9.1474	0.2285
10	18	-9.1474	-0.0092
10	20	-9.1474	-0.3117

Table1: Segment SNR for different orders

- The optimum order for SNR = 0 is observed to be 8.
- The optimum order for SNR = 5 is observed to be 13.
- The optimum order for SNR = 10 is observed to be 13.

5 Results

5.1 SNR = 0 dB

For SNR = 0, the optimum performance was observed at order = 8. The plots obtained, when the speech signal was processed for order = 8, are given below.

5.1.1 Time domain plot

The graphs of the signals as a function of time are given below.

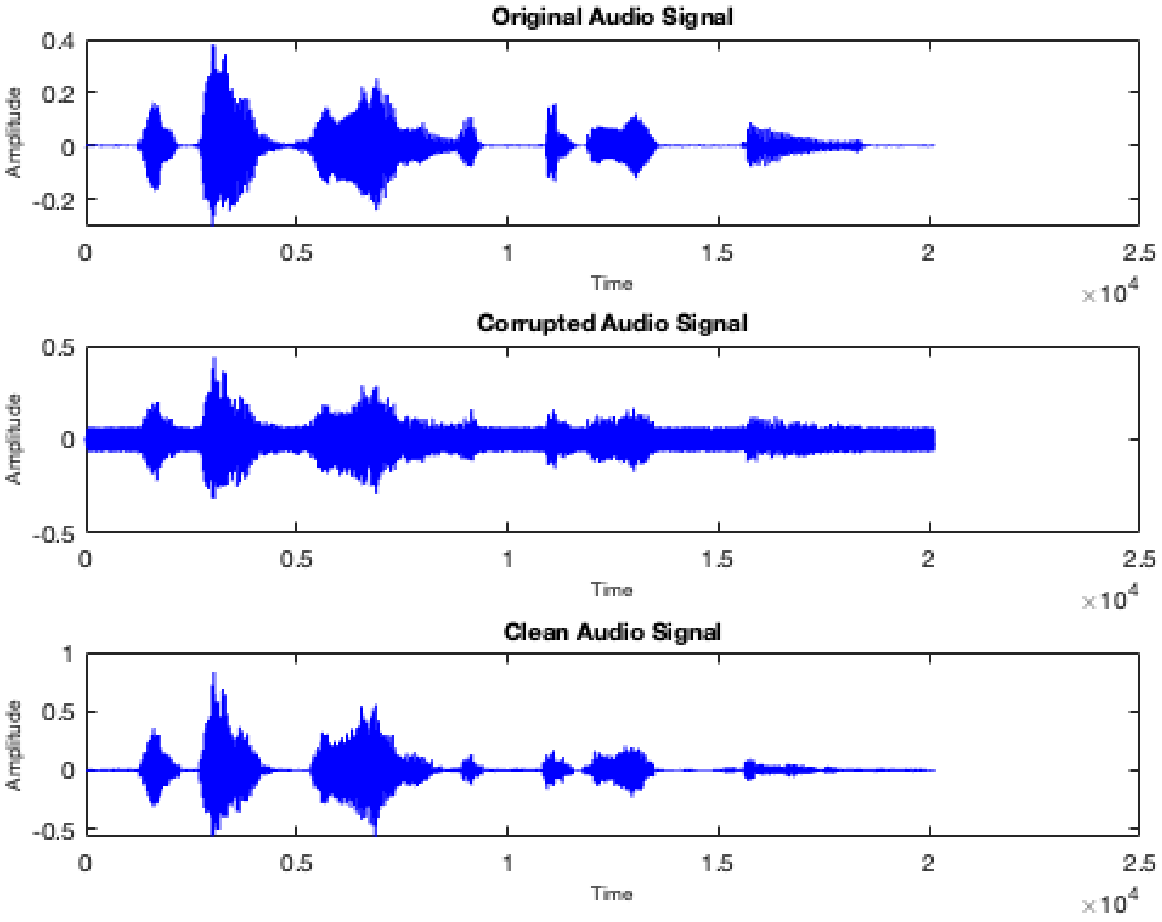

Figure 2: Signal plots in time domain

5.1.2 Frequency domain plot

The frequency spectrums are plotted below.

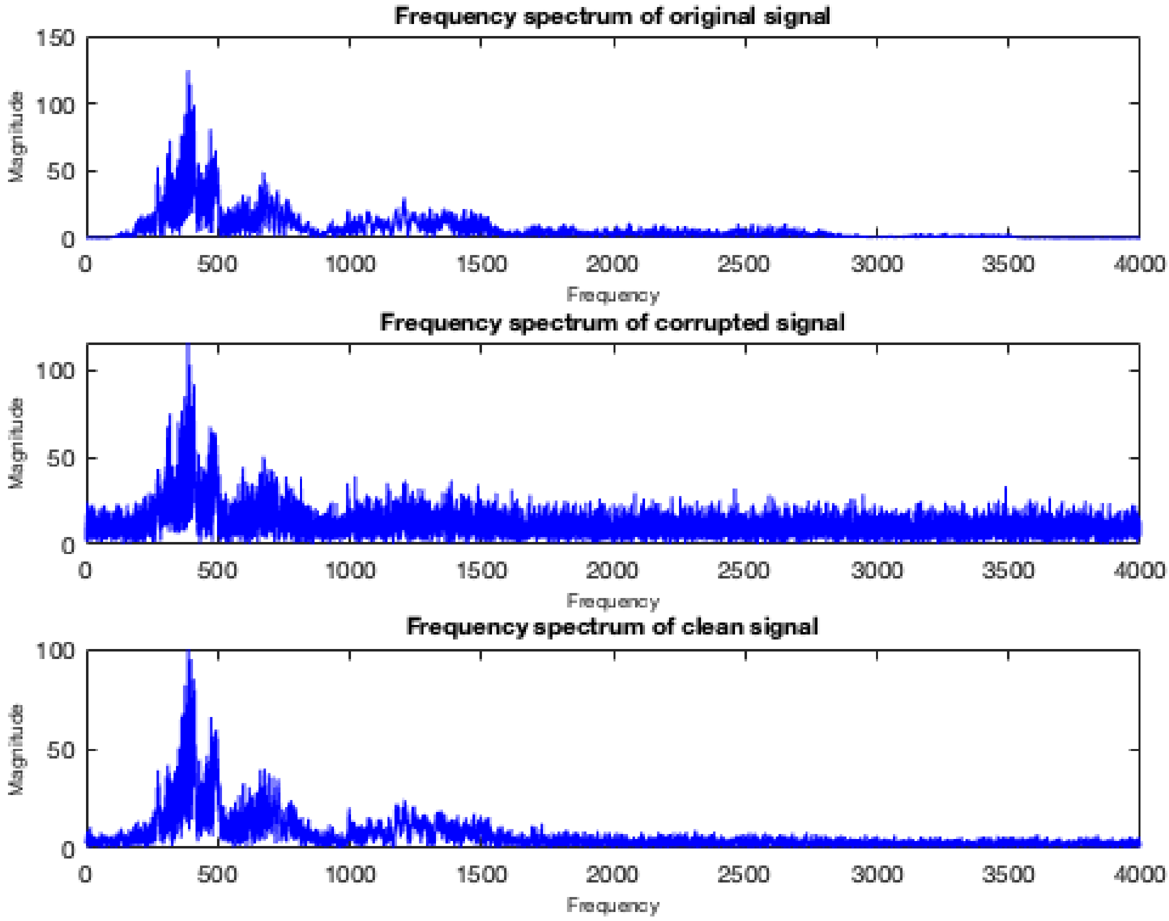

Figure 3: Signal plots in Frequency domain

5.2 SNR = 5 dB

For SNR = 5, the optimum performance was observed at order = 13. The plots obtained, when the speech signal was processed for order = 13, are given below.

5.2.1 Time domain plot

The graphs of the signals as a function of time are given below.

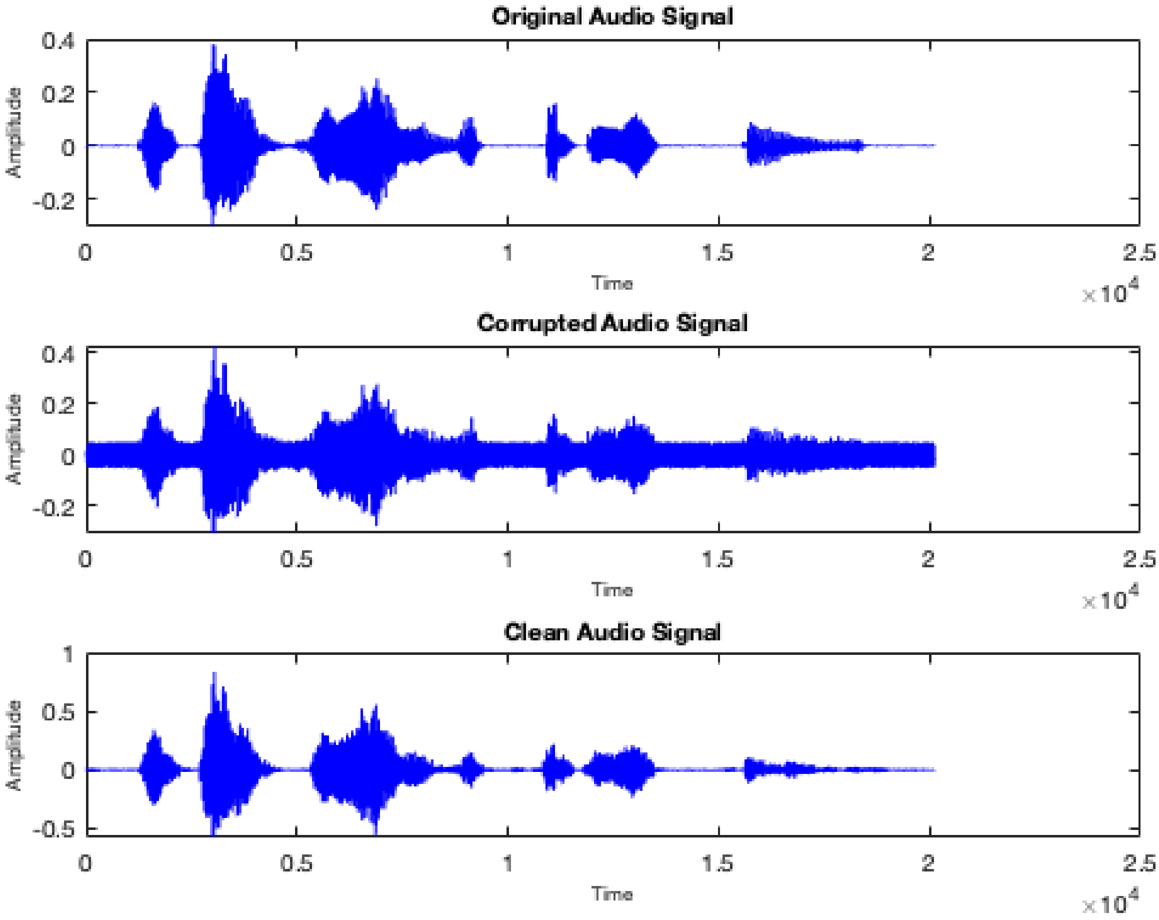

Figure 4: Signal plots in time domain

5.2.2 Frequency domain plot

The frequency spectrums are plotted below.

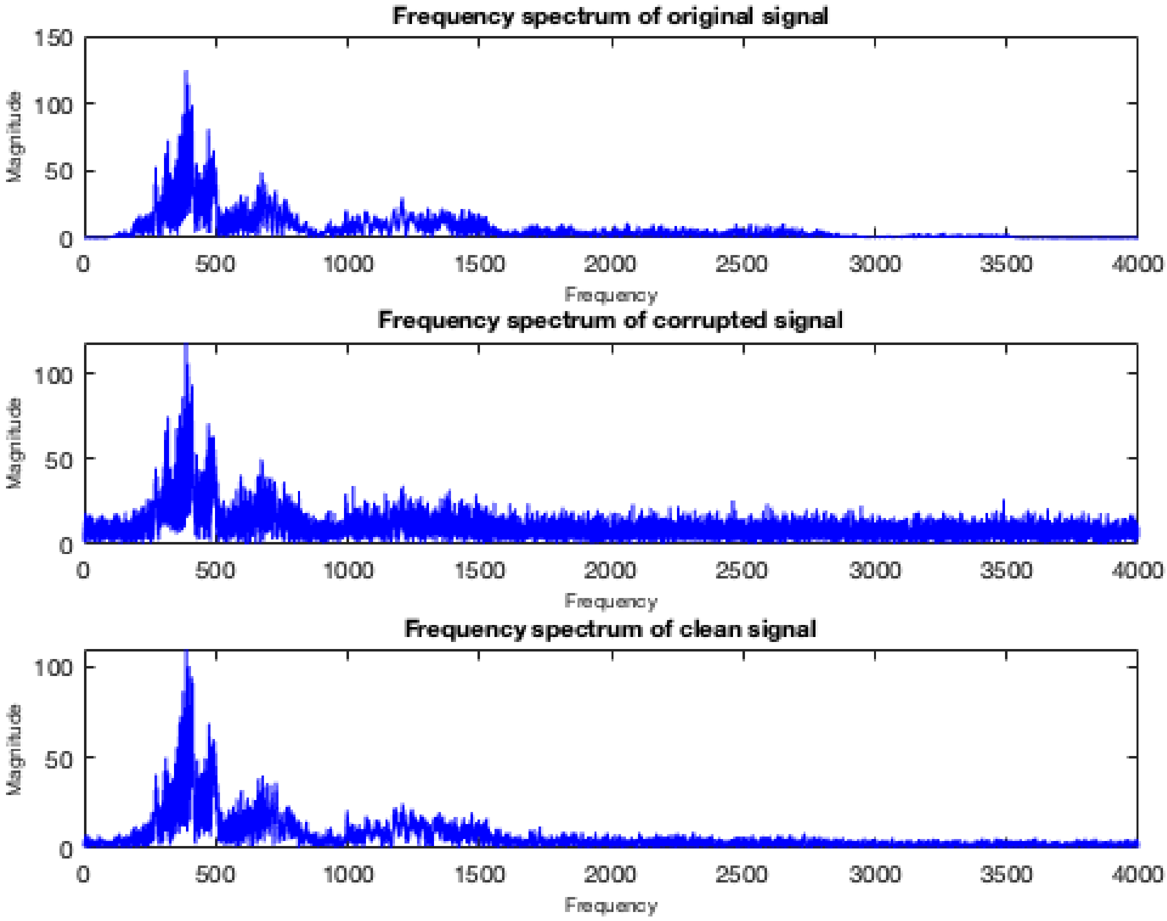

Figure 5: Signal plots in Frequency domain

5.3 SNR = 10 dB

For SNR = 10, the optimum performance was observed at order = 13. The plots obtained, when the speech signal was processed for order = 13, are given below.

5.3.1 Time domain plot

The graphs of the signals as a function of time are given below.

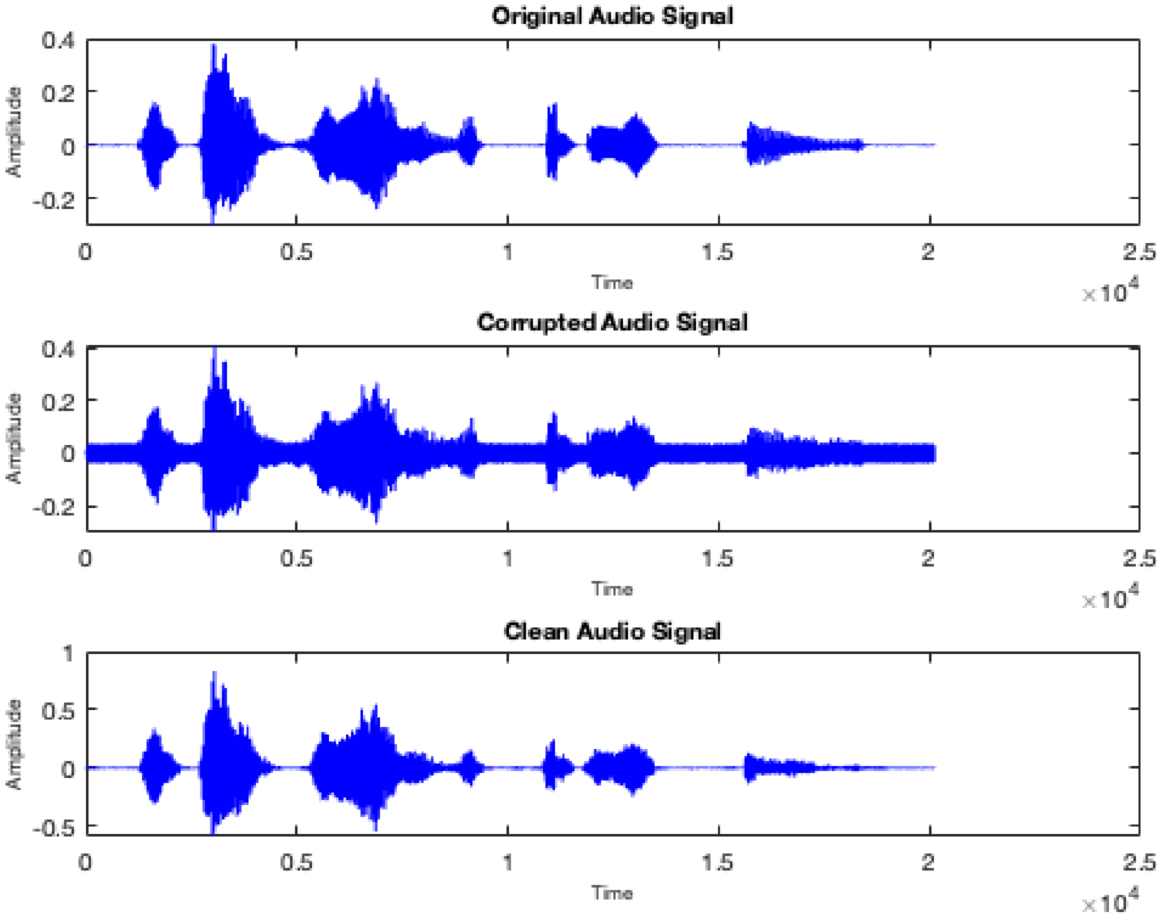

Figure 6: Signal plots in time domain

5.3.2 Frequency domain plot

The frequency spectrums are plotted below.

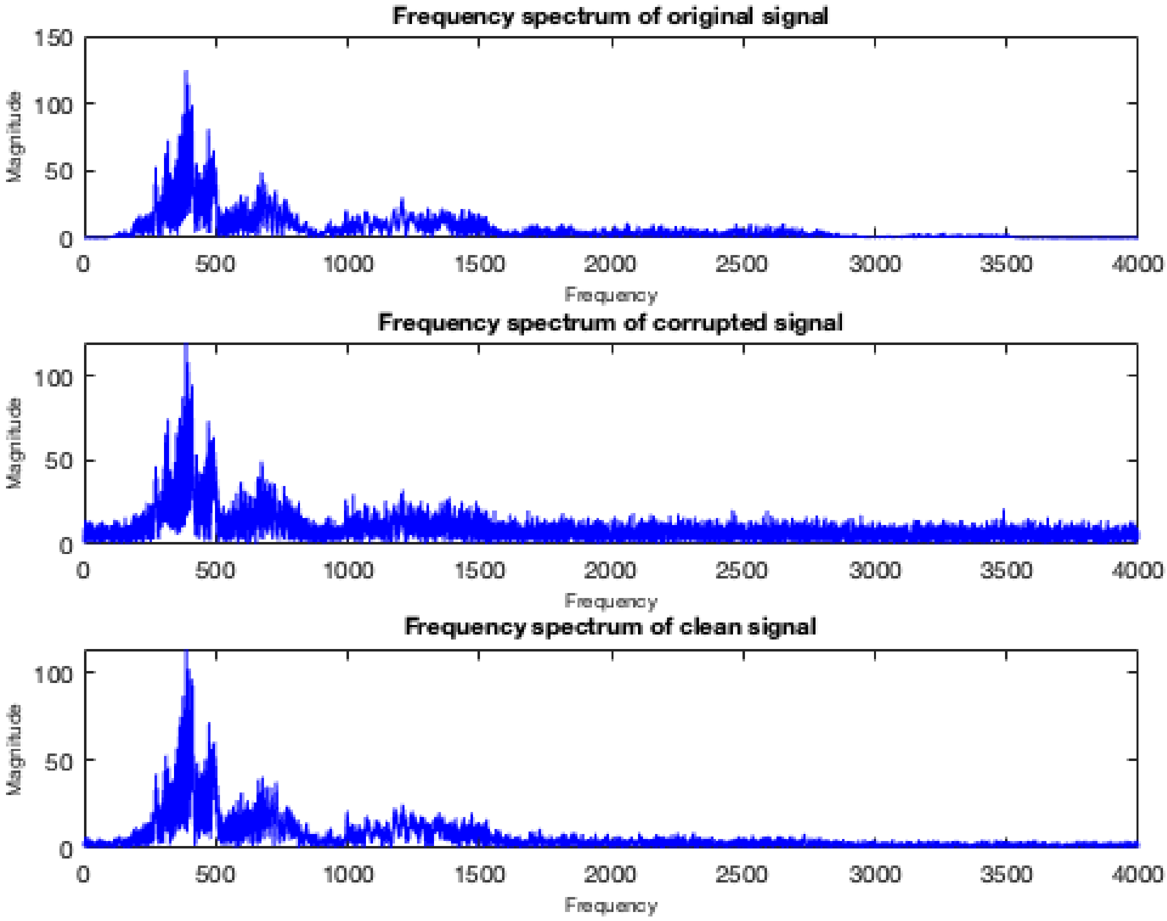

Figure 7: Signal plots in Frequency domain

6 Discussion

Based on the results obtained, it can be said that the Kalman filter has managed to de-noise the signal efficiently. It is evident in the time domain plots that the filtered speech signal resembles the original clean signal. The removal of noise can be seen better in the silent sections of the signal, that is, the parts where the amplitude of the speech is low. There is diminutive change in the signal but listening tests need to be employed to see the effect of this distortion on the coherence of the voice signal. The frequency spectra also point to similar conclusions. It is seen that the corrupted signal has components at all values of frequency. These are generated by the presence of the white noise. These components have been eliminated from the signal on passing through the Kalman filter. Also, it is seen that the magnitude of cleaned signal is lesser than that of the original signal.

According to the method chosen to measure error, a higher value of segmental-SNR represents a better signal. We observe that for different SNR values, optimum order is different for the same audio. It is seen that segmental SNR is greater for lower order systems than for higher order systems.The amplitude of the filtered signal has also reduced from the original signal. These were the quantitative results that we observed.

Furthermore, for qualitative results, the filtered signals were written to wav files and listening test was carried out. However, listening test results vary from person to person and may be deemed ambiguous. Yet, the improvement in the speech is extremely clear in this case. The intelligibility of the original speech has not been compromised with. Listening to the original, noisy and filtered audio, it was observed that in different cases of SNR the quality of the audio was improving but it's extent depends on the order chosen. It can also be observed that even after filtering, the audio has a small amount of noise left which is similar to humming.

7 Conclusion

The random process of speech was modelled.The measurement noise covariance R was found by applying probability theory and used as a parameter for a Kalman filter. The Kalman filter was implemented on the speech signal to eliminate white noise. The analysis of the performance of the Kalman filter was carried out by inspection of time and frequency domain plots and employing listening tests. The performance of the Kalman filter was satisfactory proving the success of the new algorithm proposed [8] to find the measurement noise covariance R. Nevertheless, there is a scope for improvement and more research is required to improve the filter design by tuning the filter parameters.

References

[1] A. Chaudhari and S. Dhonde, “A review on speech enhancement techniques,” in *2015 International Conference on Pervasive Computing (ICPC)*, pp. 1–3, IEEE, 2015.

[2] D. S. Kulkarni, R. R. Deshmukh, and P. P. Shrishrimal, “Article: A review of speech signal enhancement techniques,” *International Journal of Computer Applications*, vol. 139, pp. 23–26, April 2016. Published by Foundation of Computer Science (FCS), NY, USA.

[3] R. E. Kalman, “A new approach to linear filtering and prediction problems,” 1960.

[4] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” *The journal of the acoustical society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[5] K. Paliwal and A. Basu, “A speech enhancement method based on kalman filtering,” in *ICASSP’87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, pp. 177–180, IEEE, 1987.

[6] K. Paliwal, “Estimation of noise variance from the noisy ar signal and its application in speech enhancement,” in *ICASSP’87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, pp. 297–300, IEEE, 1987.

[7] G. Eshel, “The yule walker equations for the ar coefficients,” *Internet resource*, vol. 2, pp. 68–73, 2003.

[8] O. Das, B. Goswami, and R. Ghosh, “Application of the tuned kalman filter in speech enhancement,” in *2016 IEEE first international conference on control, measurement and instrumentation (CMI)*, pp. 62–66, IEEE, 2016.

[9] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, 2001.

[10] M. Saha, R. Ghosh, and B. Goswami, “Robustness and sensitivity metrics for tuning the extended kalman filter,” *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 4, pp. 964–971, 2013.

[11] O. Das, “Kalman filter in speech enhancement,” 2016.