3-30-2004

An Online Strategy for Wavelet Based Analysis of Multiscale Sensor Data

Alok K. Buch
University of South Florida

AN ONLINE STRATEGY FOR WAVELET BASED ANALYSIS OF MULTISCALE SENSOR DATA

by

ALOK K. BUCH

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Industrial Engineering
Department of Industrial and Management Systems Engineering
College of Engineering
University of South Florida

Major Professor: Tapas K. Das, Ph.D.
Jose L. Zayas-Castro, Ph.D.
Ashok Kumar, Ph.D.

Date of Approval:
March 30, 2004

Keywords: wavelets based multiresolution, real-time multiscale analysis, sprt

DEDICATION

Dedicated to my beloved parents and fiance

ACKNOWLEDGEMENTS

It has been a very good experience at the Industrial and Management Systems Engineering Department in University of South Florida while pursuing my Masters degree in Industrial Engineering. It gives me immense pleasure and happiness to be on the verge of earning my Masters of Science in Industrial Engineering degree. I take this opportunity to thank many distinguished personalities, friends and relatives who have assisted me during this journey so far.

To begin with, I would like to thank the Dept. of Industrial and Management Systems Engineering for accepting me as their Graduate student and provide me with the skills and knowledge that will help me in my professional life.

I am sincerely grateful to Dr. Tapas Das for being my mentor and a great motivator. I thank him for keeping faith in me and give me an opportunity to do quality research and bring out the best in me. Inspite of being involved in so many different activities, he always made himself available to assist me with my work or any other problems It has been a privilege to get acquainted with someone like him, whose perseverance and positive thinking make things look easier. I also take this opportunity to thank his family for allowing me to take their share of time with Dr. Das.

I sincerely thank Dr. Jose Zayas Castro and Dr. Ashok Kumar for being my committee members and assisting me in my research.

I thank Rajesh Ganesan for helping me with my research on various stages. His contribution helped me immensely in completing my thesis successfully. I thank my room mates Arun, Sanjit, and Subodh and friends like Atul, Devashish, Sanket, Santosh, and Vasanta for helping me in various ways.

Lastly, I would like to thank my parents and fiance for their constant support and love for me and also my Uncle Pradeep Buch for giving me the opportunity to study in the United States.

TABLE OF CONTENTS

LIST OF FIGURES

AN ONLINE STRATEGY FOR WAVELET BASED ANALYSIS OF MULTISCALE SENSOR DATA

Alok K. Buch

ABSTRACT

Complex industrial processes are represented by data that are well known to be multiscaled due to the variety of events that occur in a process at different time and frequency localizations. Wavelet based multiscale analysis approaches provide an excellent means to examine these events. However, the scope of the existing wavelet based methods in the fields of statistical applications, such as process monitoring and defect identification are still limited. Recent literature contains several wavelet decomposition based multiscale process monitoring approaches including many real life process monitoring applications, such as tool-life monitoring, bearing defect monitoring, and monitoring of ultra-precision processes such as chemical mechanical planarization (CMP) in wafer fabrication. However, all of the above mentioned wavelet based methodologies are offline and depend on the visual observations of the wavelet coefficients and details. The offline analysis paradigm was imposed by the high computation needs of the multiscale analysis, whereas the visual observation based approach was necessitated by the lack of statistical means to identify undesirable events. One of the most recent multiscale application, that deals with detecting delamination in CMP, addressed the need for online analysis by developing a moving window based approach to reduce computation time. This research presents 1) development of a fully online multiscale analysis approach where

the speed of wavelet based analysis of the data matches the rate of data generation, 2) development of a statistical tool based on Sequential Probability Ratio Test (SPRT) to detect events of interest, and 3) development of an approach to display the analysis results through real time graphs for ease of process supervisory decision making. The developed methodologies are programmed using MATLAB 6.5 and implemented on several data sets obtained from metal and oxide CMP of wafer fabrication. The results and analysis are presented.

CHAPTER 1

INTRODUCTION

Process monitoring refers to the task of determining quality characteristics at various stages of a manufacturing process with the objective of detecting abnormal process operations resulting from the shift in the mean and / or the variance of one or more process variables. The most popular technique used for process monitoring is by using control charts. Some of the most popular univariate parametric control charts used, as tools of process monitoring are CUSUM chart, MA chart and EWMA chart. These tools have a fixed scale. Figure 1 shows the unique scales at which some of the single scale charts represent the measurements. From the figure, it is observed that the Shewhart chart represents data at the sampling interval, which is of finest scale, and the CUSUM chart represents data at the scale of all measurements, which is of coarsest scale. In general, each of these control charts filter the data at a single scale and are suited to detect specific magnitudes of shift. For example, Shewhart filter is suited for detecting large shifts in mean, whereas MA, CUSUM, and EWMA filters are suited for smaller shifts in mean. The single scale filters or linear filters perform well only when the measurements are non-correlated. For complex processes, measured data are inherently correlated and multiscale in nature due to the deterministic/stochastic events occurring with different localizations in time and frequency. Therefore, a multiscale analysis of data becomes imperative for better understanding of a process that results in efficient monitoring.

Recent literature has presented different wavelet decomposition techniques used for multiscale data analysis, which are conceptually superior to those of the

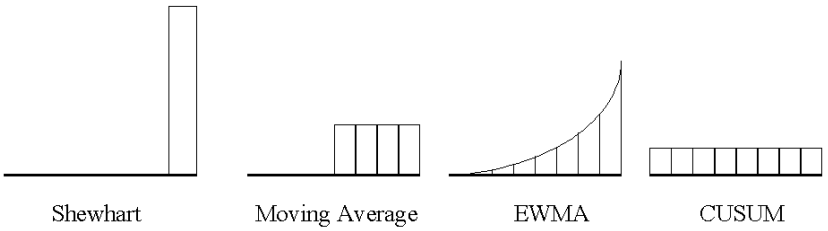

Figure 1.1. Single Scale Filters.

traditional single scale methods. In most of these methods, the multiscaled time-domain data is decomposed separately for each variable into several scales (frequency bands) by using wavelets. The data at each of these scales are then monitored to detect any process inconsistencies. In addition to wavelet decomposition, tools like thresholding methods, wavelet reconstruction and other charting techniques are also involved. Even though such wavelet based strategies have been applied successfully in many online manufacturing applications such as tool wear monitoring, detection of broken tools and monitoring of abnormal forces on the machine spindle, the online application of these strategies in ultra-precision manufacturing processes such as chemical mechanical planarization (CMP) is in the initial stages of evolution. The major reasons prohibiting the use of the existing strategies online in ultra-precision processes are high computational time taken for multiscale analysis and the absence of any statistical methodology to detect abnormal events.

Chemical Mechanical Planarization (CMP) is a nanomanufacturing process that is critical in fulfilling the requirements of semiconductor device manufacturing, such as continual feature size reduction, introduction of new materials for higher processing speeds and improved reliability, multilevel metallization (MLM) or interconnections, and increased productivity through larger wafer sizes. Traditional approaches to planarization in wafer fabrication, such as spin-on-glass (SOG), Borophosphosilicate glass (BPSG) reflow, bias sputtering, dry etching, and deposition-and-etch back process, have fallen short of fulfilling the above industry needs. However, the increased sophistication of the CMP process has brought dif-

ficult manufacturing challenges, that include defects such as delamination, dishing, under/over polishing, process monitoring and process control.

Recent literature presented offline analysis of the acoustic emission (AE) signal using Wavelet based multiresolution to detect abnormal events in CMP such as delamination by taking the longest possible dyadic length (2^{16}) of the data. Wavelet decomposition and energy analysis had to be done for one level at a time due to the high computation needed. Moreover, the dyadic discretization (wavelet decomposition with downsampling by 2) introduces a time delay in the computation of the coefficients at non-dyadic location, and this problem is severe at coarser scales (Bakshi [35]). To overcome this time delay, and implement continuing defect identification during the process, a moving window methodology for online defect identification using wavelet analysis of AE signal was developed. However, this online moving window strategy strategy is still computationally slow to analyze the data at different levels and lacks statistical means to identify undesirable events on a real time basis. Morover, this method does not display the analysis results through real time graphs for ease of process supervisory decision making.

In order for real time analysis of sensor signal, it is necessary to develop a strategy where the speed of analysis matches the rate of data generation and there is a statistical tool to detect events of interest. The strategy should also provide the means to display the results through real time plots. This research focuses on development of such a strategy and its implementation on sensor data collected for a CMP process.

The next chapter reviews the different elements of wavelet based multiscale analysis methods, their performance, and applications followed by a review of the most recent research in CMP process monitoring.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, a systematic review of the research done in the area related to multiscale process monitoring is presented. First, the basic structure of multiscale algorithms is presented. This is followed by 1) a review of different elements of the Multiscale Methods such as denoising, Monitoring, Width of the testing window, Selection of Wavelet type, Border distortion or end effects, On-line and Off-line Monitoring and Contribution Plots, 2) Performance of the Multiscale Process Monitoring Methods, and 3) Applications of Multiscale Methods. Finally, a description of the current trends in the CMP process monitoring and control is provided.

2.2 Basic Structure of Multiscale Process Monitoring Algorithms

The general structure of the existing multiscale algorithms for process monitoring is shown in Figure 2.1. The input data for multiscale algorithms is either univariate or multivariate, and has one or more of the following characteristics: stationary or nonstationary, Gaussian or non-Gaussian corrupted with random or gross errors, independent or auto/cross correlated, linear or nonlinear, and deterministic or stochastic. As a first step, many of the existing algorithms consider prescreening of the time domain data to look for missing data or outliers. Prescreening also reveals the extent of noise and the presence of autocorrelation in the data. The variables are then decomposed into different scales on a set of basis functions, which are orthonormal or nonorthonormal. The wavelet coefficients at all the scales are then

denoised using different thresholding techniques, which are reviewed in the next section. Denoising is followed by the formation of details and approximation, which are monitored using different charting techniques. The denoising step is a crucial one, as insufficient denoising will distort the readings during the monitoring by introducing errors, and excessive denoising will over-smooth the sharp features of the signal by considering them as outliers. Subsequent to monitoring of both details and approximation, the time domain signal is reconstructed using the thresholded wavelet coefficients from the scales that are found to be significant in the monitoring process. The reconstruction is accomplished by the inverse wavelet transform of the thresholded coefficients. Monitoring of the reconstructed time domain signal is then carried out by various charting techniques.

2.3 Elements of the Multiscale Methods

There are several elements in multiscale methods that are important and must be addressed while designing an efficient monitoring approach. These include denoising, monitoring, depth of decomposition, width of the testing window, selection of the wavelet type, border distortion or end effects, issues involved in moving from off-line to on-line monitoring, and contribution charts. The above elements are briefly described in the following sub sections.

2.3.1 Denoising

Industrial process data is often multivariate and is generally corrupted with different forms of noise. Therefore, it is essential that the true signal be extracted from the noisy data before monitoring methods are applied. Literature has presented multiscale denoising and process monitoring for processes with models and without models. A detailed review of the different denoising and monitoring methods

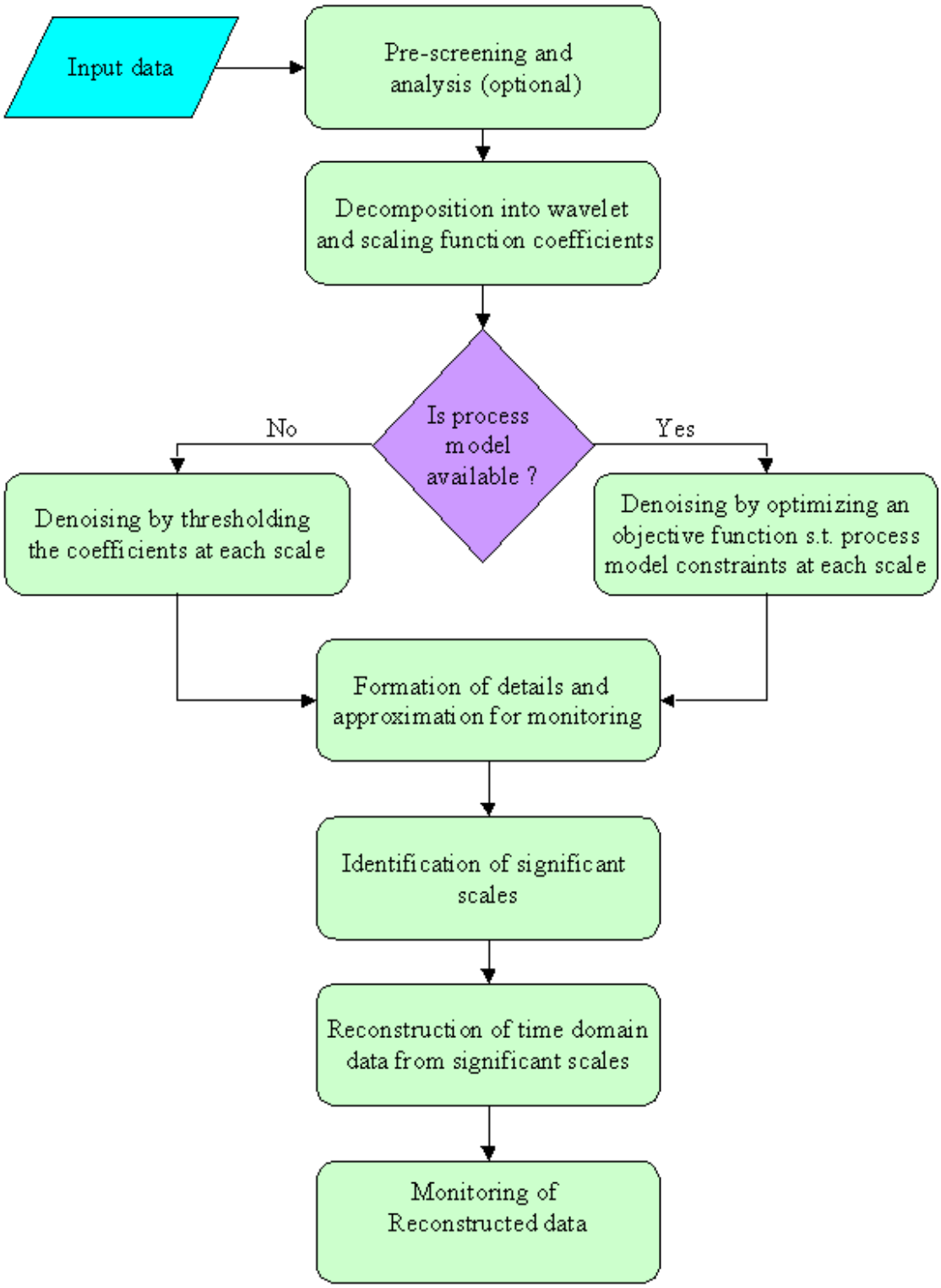

Figure 2.1. Structure of Multiscale Process Monitoring Algorithms.

for processes without models has been done here. A detailed review of denoising and process monitoring methods for processes with models can be found in Ganesan [1].

The process of removing noise with maximum retention of the underlying process information is an important preliminary step and is referred to as data cleansing, filtering or denoising. Errors in the data due to noise could be stationary or nonstationary. Based on the underlying probabilistic distribution, noise can also be classified as Gaussian (random) and non-Gaussian (gross). Random noise can be further classified as white or colored based on the autocorrelation function (ACF) or power spectrum.

2.3.1.1 Single Scale Denoising

Some of the traditional single scale denoising methods without process models are the linear low pass filtering techniques, such as mean filtering and exponential filtering that work by taking a weighted sum of previous measurements in a window of finite length (Finite Impulse Response Filter (FIR)) or infinite length (Infinite Impulse Response Filter (IIR)). Linear filters are represented as

$$\hat{x}_t = \sum_{i=0}^{I-1} w_i x_{t-i}, \tag{2.1}$$

where, I is the filter length and w_i is a finite (FIR) or infinite (IIR) sequence of weighting coefficients which satisfy the condition

$$\sum_i w_i = 1. \tag{2.2}$$

The weighting coefficients (w_i) are the impulse response of the filter. Some examples of linear filters are Shewhart, MA, EWMA, CUSUM, Butterworth. The disadvantage of linear filtering is that, they are single scale and have a fixed time-frequency localization, i.e., they analyze data only at a single frequency. These

filters do not perform well when the measurements in the data are correlated and non-Gaussian distributed. FMH is an offline filtering method and is more effective for filtering piecewise constant signals.

2.3.1.2 Multiscale Denoising

In recent years, wavelet based multiscale denoising methods have gained tremendous popularity. Potential of these nonlinear filtering techniques are still being explored. Wavelets have the ability to represent the deterministic features (spikes, shifts in mean/variance, trends) of the data with only a small number of coefficients. The remaining coefficients represent the stochastic components. This property is used in both denoising and in detecting features of the signals. Selecting a proper threshold value is essential for effective denoising. A universal threshold suggested by Donoho *et al.* [2] (Visushrink) is given by

$$t_j = \sigma_j \sqrt{2log(n)}, \tag{2.3}$$

where n is the signal length and σ_j is the standard deviation of the noise at scale j. The value of σ_j is estimated from the median of absolute deviation (MAD) of the wavelet coefficients at scale j as

$$\sigma_j = \frac{1}{0.6745} median(|d_{j,k}|), \tag{2.4}$$

where $d_{j,k}$ are the wavelet coefficients for all k translations at scale j. The significant wavelet coefficients are then extracted by applying soft or hard thresholding as shown in Figure 2.2. Hard thresholding maintains the same value for the coefficients that exceed the threshold limits, whereas soft thresholding shrinks their values towards zero by the value of the threshold limit. It has been shown that hard thresholding can result in larger variance in the signal after reconstruction, and produce occasional

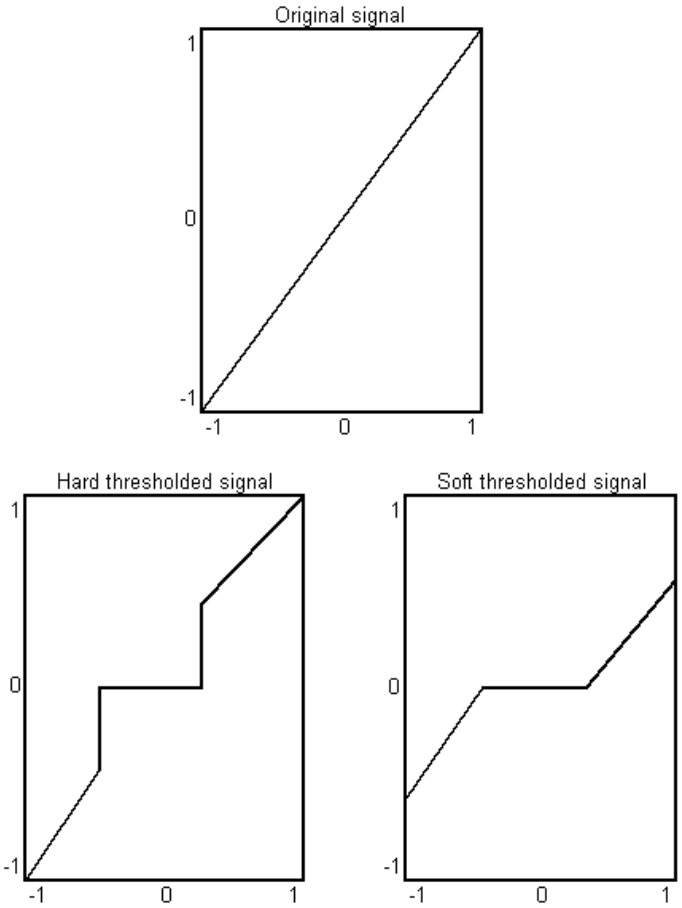

Figure 2.2. Hard and Soft Thresholding of the Signal.

artifacts that can roughen the appearance of the reconstructed signal. However, they can better represent peaks and discontinuities. Soft thresholding on the other hand has larger bias but gives better visual quality of filtering. At scale j, the thresholded coefficients are determined as follows.

Hard thresholding:

$$\tilde{d}_{j,k} = \begin{cases} d_{j,k} & |d_{j,k}| \geq t \\ 0 & |d_{j,k}| < t. \end{cases} \tag{2.5}$$

Soft thresholding:

$$\tilde{d}_{j,k} = \begin{cases} sign(d_{j,k})(|d_{j,k}| - t) & |d_{j,k}| \geq t \\ 0 & |d_{j,k}| < t, \end{cases} \tag{2.6}$$

where $sign(d_{j,k})$ is the positive or negative sign of the wavelet coefficient $d_{j,k}$. To compromise the trade-offs between variance and bias, a firm thresholding was suggested (Bruce and Gao [3]), which requires two threshold values and thus more computation. In practice, the universal threshold tends to oversmooth and depends only on the data through the σ_j estimate. In fact, for large samples the universal threshold will not only remove almost all the noise but also a portion of the underlying deterministic signal.

Online treatment of Gaussian and non-Gaussian errors, both separately and in a combined manner have been proposed by online multiscale filtering (OLMS) (Bakshi and Nounou [4], [5]), Bakshi [6]). In this method the measured data is decomposed within a moving window of dyadic length using a boundary corrected wavelet filter. The wavelet coefficients are then thresholded and the signal is reconstructed. Only the last data point of the reconstructed signal is retained for online use. When new data points are available the window is moved to include the latest data point while maintaining the maximum dyadic window length. Similar to the offline BCTI

technique, the OLMS procedure can be extended to include non-Gaussian errors by combining it with multiscale linear filtering.

2.3.2 Monitoring

After denoising the data, several monitoring schemes can be applied to detect features of the data. The monitoring schemes can be broadly classified into single scale and multiscale methods.

2.3.2.1 Single Scale Monitoring

The research literature contains several single scale approaches to analyze multivariate data, e.g., Hotelling's T^2 [7] which is a Shewhart type control chart, multivariate EWMA (Lowry [8], Prabhu and Runger [9]). These multivariate monitoring techniques are reasonably effective as long as the number of process variables is not very large. However, for industrial processes with many variables, the average run length (ARL) performance value of these charts increase with the increase in the number of variables. Often the variability in the process is not equally distributed and maximum variability is found in a relatively small subset of the original variables. To deal with such situations, process monitoring is often preceded by methods such as CVA, PLS and PCA (Ganesan [1]). All the mentioned PCA approaches have been used for single scale multivariate data analysis and some of them have been extended to multiscale analysis. Several other papers on both parametric and nonparametric approaches to single scale multivariate process control procedures exist in literature.

Single scale monitoring approaches have many limitations and some of these are reviewed here. As stated earlier, single scale monitoring approaches are suited to detect specific types of process disturbances (e.g., Shewhart chart is used in the detection of large shifts, whereas MA, EWMA and CUSUM are more suited for detecting small shifts). Another limitation of the single scale monitoring approach

is its ineffectiveness in controlling autocorrelated data. However, as the number of process variables increase, these approaches are not practical and they lack multivariate extensions. Similarly, for single scale multivariate process monitoring, where PCA is used, one of the shortcomings is the limited ability of the PCA method to reduce the error by eliminating some principal components. Due to this limitation, some errors usually leak past into the model. Single scale multivariate methods also suffer from the fact that the data along each principal component is monitored using single scale charts. Most of the above limitations may be overcome by the use of multiscale approach to process monitoring.

2.3.2.2 Multiscale Monitoring

In recent years there have been considerable developments in the use of wavelet based multiscale analysis methods. Signal processing, image processing, and data compression are three areas where wavelet methods have been already proved to be of significant value. A good review of the literature on wavelet applications can be found in Meyer [10], Mallat [11]. Though there are not many applications of multiresolution analysis in the area of process monitoring, the research is picking up in this area and its potential to contribute in the development of process monitoring tools is still largely unexplored. The first definition of wavelets can be attributed to Morlet [12]. Meyer [10] developed wavelets with infinite support and exponential decay. The work done by Meyer is regarded as the theoretical breakthrough in wavelets. Mallat [11] and Meyer [10] developed orthonormal wavelet bases functions in a general framework called multiresolution analysis. Daubechies [13] derived wavelet bases that have compact support. The Daubechies wavelets are the most widely applied wavelet families. These developments have made wavelet methods more suitable in analyzing various signals or processes.

Recent research has shown the ability of wavelets to analyze process data at multiple scales. It was shown that wavelet decomposition can approximately decorrelate the measurements when they are autocorrelated. The idea of detecting changes in signals using wavelet transformation was first developed by Grossman *et al.* [14]. Mallat and Zhong [15] had applied wavelet transformation to detect change points (inflexion points, discontinuities). Jump and sharp cusp detection and edge detection by wavelets were studied by Wang [16] and Song and Jutamulia [17] respectively. Cheung and Stephanopolous [18] developed a methodology for representing process trends. This method describes the qualitative and quantitative features of a signal. Bakshi and Stephanopolous [19] developed a methodology to extract features at multiple scales using wavelet approach. They made use of wavelet interval trees to extract stable trends at multiple scales, based on which they developed a methodology for fault diagnosis and pattern based supervisory control (Bakshi and Stephanopoulos [20]). A few other wavelet based process monitoring applications are Alexander [21], Ibrahim *et al.* ([22], [23]).

Multiscale statistical process monitoring has been applied to both univariate and multivariate process data. In the case of a univariate process, the time domain data is wavelet decomposed into coefficients at each scale, which are then denoised using the thresholding rules discussed earlier. The details and approximation are derived from the thresholded coefficients and conventional SPC techniques are used to monitor them. The time domain data is then reconstructed from the thresholded coefficients of the significant scales. The reconstructed data is monitored using conventional SPC techniques (Bakshi [6], [24]).

Multiscale principal components analysis (MSPCA) (Bakshi [25]) is among the recent multivariate multiscale statistical process control (MSSPC) approaches without process models. MSPCA uses wavelet decomposition to approximately decorrelate the autocorrelated data, and also uses linear PCA to remove the cross

correlation among the multivariate data. The description of the method can be found in (Ganesan [1]) . The MSPCA approach not only extracts the significant signal features and monitors them but also adapts to the nature of the signal features. This approach has been further extented into a nonlinear MSPCA (NLMSPCA) for process monitoring and fault detection (Shao *et al.* [26], Fourie and Vaal [27]). These authors have used an input trained neural network to extract the latent nonlinear structure from the PCA transformed data set. Bivariate plots with both nonparametric and parametric control limits were used to monitor the output of the neural network. Differential and residual contribution plots were also used to relate a defect to the responsible variable.

2.3.3 Depth of Wavelet Decomposition

When a discrete wavelet transform (DWT) is used in multiscale denoising or process monitoring, it is necessary to fix the depth of decomposition of the time domain data. Excessive decomposition and thresholding at coarse scales would result in higher compression of the data and thus involve the risk of eliminating important features and over smoothing of the signal. Also, excessive decomposition means more control charts, difficulty of type-I error allocation, and increase in the computational time. On the other hand, inadequate decomposition will result in a considerable portion of the noise to be retained in the reconstructed signal. Thus the depth of decomposition needs to be optimized in order to have the best quality of filtered signal. For a signal of length n in an offline mode, or a window of length n in an online mode, empirical evidence suggests that the depth of decomposition should be half the maximum possible depth. For example, for a dyadic data length $n = 2^j$, the recommended depth of decomposition is $j/2$. In dealing with long patches of outliers with multiscale median filtering, the depth of decomposition is chosen such that the

effective median filter length at the coarsest scale is longer than the longest patch of outliers (Bakshi [5]).

2.3.4 Width of the Testing Window

For offline analysis, the entire (dyadic) signal length is considered as a single window. The edges are corrected by the methods explained in Subsection 2.3.6. However, for online monitoring the length of the test window must be suitably selected. Clearly, longer window length means higher possible depth of decomposition, which might result in excessive smoothing and also increased sensitivity to smaller shifts at lower levels of decomposition. The online monitoring procedure is to move the window in time and include the most recent measurement while maintaining the maximum dyadic window length. The window length is then held constant after reaching a sufficient upper limit. This upper limit is decided on a case-by-case basis depending on the nature of the signal and the objective of the study.

2.3.5 Selection of Wavelet Type

Several wavelet basis function types are available in the literature. Some of these are the Haar, Daubechies, coiflets, symlets, bi-orthogonal wavelets, etc. The Haar basis was known even before the wavelets were developed. Though Haar has a compact support, it does not have good time-frequency localization. Moreover, it is unsuitable for representing classes of smoother functions due to its discontinuities. Some of the desirable properties of the basis functions are good time-frequency localizations, various degrees of smoothness (number of continuous derivatives), and large number of vanishing moments (ensures maximum number of zeros of the polynomial at the highest discrete frequency). For a detailed description on the mathematical properties of the wavelet basis function, refer to Strang and Nguyen [28]. The most widely used wavelet is the Daubechies basis function. The Haar filter is best suited

to represent step signals or piecewise constant signals, whereas the Daubechies filter is better for smoother signals.

2.3.6 Border Distortion or End Effects

When the signal is of finite length, then filters other than Haar would require additional data at either ends of the signal for decomposition. In general, for orthonormal wavelets the dyadic (2^j) point gets supported on $N/2$ coefficients (N = filter length). However, with boundary problems, such a support does not exist at every scale. Since Haar wavelet has only one vanishing moment, the contribution of the last point in the moving window is directly transferred to the last detail and approximation coefficient for every scale. However, with online implementation using higher order wavelets, the last scaled signal, which is used for monitoring purpose, would be the least accurate one due to inaccuracies induced by border distortion. Not only does the finite length of the signal but also the noncausal nature of most wavelets lead to this border distortion. Due to this problem, the original signal is not accurately represented at the edges. Hence, to deal with finite length signals, one would need to extrapolate the data, or use other methods to handle boundary problems. Some of the ways to deal with this issue is by using periodic wavelets, folded wavelets, or boundary wavelets. Eventhough dealing with boundary corrected filters for convolution of the edge translations of the discrete wavelet transform (DWT) seems effective theoretically, it is computationally impractical. There exists an alternate option of dealing with the edge of the signal for minimizing boundary distortion effect. This involves the computation of a few extra coefficients at each stage of the decomposition. Three popular approaches have been suggested in the literature. These are zero padding, symmetrization, and smooth padding. Theory behind these approaches can be found in Strang and Nguyen [28]. Zero padding assumes zero values for the extended section of the signal. This is a default procedure similar to not

taking any action to deal with the discontinuities created at the edge of the signal. Symmetrization assumes a symmetric value replication outside the original support of the signal. The disadvantage of this approach is that discontinuities are artificially created in the first derivative of the function (signal). Finally, smooth padding applies a simple first order derivative extrapolation. This approach has been shown to work well in general for smooth signals.

2.3.7 Online and Offline Monitoring

Most of the wavelet based methods for denoising or process monitoring have a significant disadvantage when it comes to online implementation of these methods. The noncausality of the wavelet introduces a time delay in the computation of the coefficients at nondyadic locations and this problem is severe at the coarser scales (Bakshi [5]). Also the dyadic discretization (wavelet decomposition with downsampling by 2) requires the signal to be of dyadic length for the decomposition to take place. Hence, this restricts the application of the wavelet-based methods to offline use. However, dyadic discretization has a significant advantage over integer discretization, since it allows decomposition using orthonormal wavelets, which also approximately decorrelates the autocorrelation in the data. The resulting coefficients at each scale are uncorrelated and Gaussian distributed with equal variance.

In online monitoring applications decisions are made based on the last sampled data point. Since decomposition using DWT requires data of dyadic length, it forces a delay in deriving process information. For example, at scale j=1, for data sampled at unit time interval apart, the wavelet coefficients (that depicts the process condition) are obtained at times 2, 4, 6, 8,....., which means that at times 1, 3, 5, 7,..... no process information is available until the next data point is obtained. This delay increases geometrically with scales, e.g., at j=3, process information is

obtained only at times 8, 16, 24, 32,..... Such delays can be eliminated by using a moving window of dyadic length and considering only the last coefficients at all scales. This approach is called uniform or integer discretization. However, in this procedure the wavelet coefficients are no longer orthonormal to each other and they lose the property of approximately decorrelating the autocorrelated data. The variance at each scale is still proportional to the power spectrum of the measured data, but level dependent thresholding must be used due to the presence of autocorrelation. A detailed description of the selection of threshold limits by maintaining equal false alarm rate and the adjustments needed to account for autocorrelation are discussed by Aradhye *et al.* [29].

2.3.8 Contribution Plots

Once a fault is detected by the monitoring step (T^2, Q or nonparametric charts) of a multivariate process monitoring procedure with transformed data (using PCA, PLS etc.), it is imperative to identify the original variable(s) responsible for the defect. This can be done by implementing a process variable contribution plot as suggested by Miller *et al.* [30]. An alternate procedure is based upon the assumption that the partial derivative of a function with respect to a specific variable can indicate the relative influence of the variable on that function. This is called a differential contribution plot (Shao *et al.* [26]).

2.4 Performance of the Multiscale Process Monitoring Methods

A review of the existing literature, presented in this paper, reveals that almost all the research done so far has been focused on developing new methodologies for multiscale data denoising and process monitoring. The benefits of multiscale methods have been projected in the literature in terms of their superior ability to detect process features, such as defect, fault, change point, edge, process operating

region, and system states in situations where other traditional methods are likely to be inefficient. The other claimed benefits for multiscale methods are 1) denoising capabilities, 2) ability to study signals localized in time and frequency, 3) capability to simultaneously handle small and large shifts in the mean and/or variance within the same monitoring framework, 4) feature extraction capabilities, and 5) the ability to handle other data features such as autocorrelation, Gaussian or non-Gaussian distributions for data and errors, and linearity or nonlinearity of the process. One of the key measures in judging the efficiency of a monitoring scheme is the average-run-length (ARL) performance under varying process conditions. Performances of a few of the available multiscale methods have been studied for univariate process data by comparing their ARL with those of the single scale approaches (Aradhye *et al.* [29] and Bakshi [6]). A comparison of a multiscale method with MA and Shewhart charts (for Gaussian iid measurements under both dyadic and integer discretization) for detecting mean shifts was conducted by Aradhye *et al.* [29]. The results indicate that at smaller mean shifts the MA chart performance is the best, and at large shifts Shewhart performance is superior. The performance of the multiscale method was shown to be in-between MA and Shewhart. Aradhye *et al.* [29] also examined the performance of MSSPC for univariate data in comparison to the methods of weighted batch means, moving center line exponentially weighted moving average (MCEWMA), and the residuals for a stationary autocorrelated process. This study was repeated for a nonstationary process using MCEWMA. Multiscale performance has been shown to be in-between weighted batch means and the residuals in the stationary case, and better than MCEWMA at large shifts for the nonstationary scenario. The above simulation based studies have assumed the data to be corrupted with white noise and has used Haar wavelet for the multiscale analysis.

Very little has been done so far in evaluating the ARL performances of different multivariate MSSPC techniques. The only study in the multivariate MSSPC

domain is the work by Aradhye et. al [29], which compares PCA and DPCA with MSPCA and MS-DPCA using a linear uncorrelated model and a correlated time series model. For uncorrelated data, the ARL performance of PCA and MSPCA match each other. In the autocorrelated situation DPCA performs better with dyadic discretization and MS-DPCA performs better with integer discretization. Clearly, a comprehensive comparison of the ARL performances of the multivariate MSSPC techniques with the other single scale statistical tools under various process conditions is not available in literature.

2.5 Applications of Multiscale Methods

Wavelet based multiscale analysis has found many interesting applications in areas including engineering. The multiscale methods are analogous to Fourier analysis and can represent various objects, such as signal, function, and images. Wavelet theory is growing very rapidly and its applications cover physical, medical,engineering, and social sciences.

Advanced use of wavelets is found in the fields of 2D and 3D image processing, data compression, image and video compression, and other digital signal processing applications. A few examples of such applications are: FBI finger print analysis, medical image analysis (analysis of one-dimensional physiological signals obtained by phonocardiography, and electrocardiography (ECG)), biomedical image processing algorithms (e.g., noise reduction, and image enhancement), image reconstruction and acquisition schemes (tomography and magnetic resonance imaging (MRI)), and multiresolution methods for the registration and statistical analysis of functional images of the brain (positron emission tomography (PET)). Other application areas include speech recognition, music synthesis, geophysical sensing and seismic studies.

Wavelets have been extensively used in processing data from chromatography, electrochemical studies, voltammetry, spectroscopic studies, physical chemistry, and in chemical process monitoring. Several applications have been found in manufacturing systems, such as flexible manufacturing systems (FMS) and other automated manufacturing set-ups. Some of the key issues in manufacturing is earlier detection of components that wear, such as cutting tools and bearings, quick detection of broken tools, monitoring abnormal forces on the machine spindle, and vibrations. These issues are studied by monitoring different signals that are captured through various transducers mounted on the machine components. One such signal is the acoustic emission (AE) signal, which is a high frequency signal with frequencies much higher than the normal machine vibrations and environmental noise. Wavelet transform analysis of this AE signal has resulted in revealing a wide range of information that are not usually seen in other methods, such as the Fast Fourier transforms (FFT). Some references for such applications include Qi [31], Li *et al.* [32], Wang and chu [33], Paul [34], and Chen *et al.* [35].

In the last decade, research in wavelets has been extended to numerous statistical applications. Linear and nonlinear nonparametric regression was one of the first type of applications where the potential of wavelets to recover an underlying function from a noisy data was shown (Antoniadis [36], [37]. Density estimation is another area where wavelets have been applied (Safavi *et al.* [38], Donoho *et al.* [39], and Vannucci [40]). Denoising or data rectification using wavelets as discussed earlier in this paper, have resulted in a wide range of feature extraction applications due to the capability of wavelets in separating the high frequency components of the signal from the low frequency ones by successive decomposition of the signal. A good summary of statistical applications with numerous references can be found in Abramovich *et al.* [41]. Many research papers exist that deal with detection of change point, edge, and discontinuity. Among the recent research in statistical area,

multiscale process monitoring has drawn considerable attention for both univariate and multivariate process data as reviewed in this paper (Ikonomopoulos and Endou [42] and Rosen and Lennox [43]).

2.6 CMP Process Development, Monitoring and Control

The most important CMP process outputs that are of interest include, material removal rate (MRR), within wafer non-uniformity (WIWNU), within die non-uniformity (WIDNU), wafer-wafer non-uniformity (WWNU) and surface quality such as roughness, presence of impregnated particles, and corrosion resistance. A significant number of input parameters influence these output parameters.

Most of the process development studies in the literature are focused on understanding the effect of polishing conditions and consumable sets to improve WIWNU. This is achieved by developing a CMP process on unpatterned wafers in order to optimize WIWNU. However, to determine the effectiveness of the CMP process on patterned wafers, planarity is checked through the measurement of step height (thickness) of certain features. However, recent studies have suggested that step height measurement is not a very good measure of planarity as it provides us with information about local planarity, rather than the global planarity within die or within the whole wafer. Other studies have suggested that global planarity is more closely associated with the within die non-uniformity (WIDNU) (Stine *et al.* [44], and Fang *et al.* [45]). WIDNU can be improved by changing polishing parameters (i.e., speed and pressure) and the consumable set. However, WIDNU is not a universal parameter, since it is strongly dependent on the choice of measurement locations and on the device layout. In addition to the above, measuring the thickness of the wafer as described in the previous section also monitors material removal rate. However, recent studies have explored the use of AE signal in material removal monitoring (Hwang [46], Chang [47], Chang *et al.* [48], Dornfeld [49]). A report on the research

done by Tanzawa *et al.*[50] describes the development of a data logging network system for the CMP process. This system help building up a proper process control model that represents the CMP process.

Thickness control using *in-situ* sensors and online metrology has resulted in improved CMP yield. The *in-situ* sensors are based on indirect measurement and are designed to provide real time process control, and can also correct incoming thickness variation (variation due to film deposition). Since information from *in-situ* sensors have to be correlated to thickness on a wafer, the development of a control algorithm for CMP is not straightforward. Additionally, these sensors can be sensitive to other CMP process variations leading to poor repeatability, especially at end point detection. However, online metrology has been proven to be accurate and reliable in such scenarios and can be used for run-to-run process control of CMP. Such an attempt has been made by Boning *et al.* [51] using an EWMA controller to adjust polishing time. A wavelet based multiscale method for CMP has been proposed by Ganesan *et al.* [52] to detect the delamination defect of low-k dielectric layers by analyzing AE signal. An offline strategy and a moving window based strategy for online implementation are developed. The results indicate that the the wavelet based approach using the AE signal offers an efficient means for real time detection of delamination defects in CMP processes, thus offering a viable tool for CMP process control.

This research addresses the use of wavelet decomposition based multiresolution analysis approach for real time analysis of sensor data at different scales and develop a statistical tool to detect events of interest. In the next chapter, a brief overview of the theory on wavelet based multiresolution methodology is provided.

CHAPTER 3

MULTIRESOLUTION AND SPRT

3.1 Introduction

In this chapter, the basic concepts of the wavelet based multiresolution approach are introduced. Any process is characterized by its parameters. An in-control process operating under chance causes of variation will typically consist of a single feature, i.e., data collected from the process will have a stable probability distribution. However, most manufacturing processes do not behave in an ideal way. For example, the measurements representing a process may be contaminated with noise or subjected to various parameter changes. This implies that the state of a process in general would not consist of a single feature but rather would have multiple features (each describing a change in the parameters). A typical example of a process with multiple features is shown in Figure 3.1.

A process behavior with multiple features may be caused by various events associated with the process, namely tool failure, faults, sensor failure and machine parts degradations. Identification of a change in a parameter usually leads to identifying the various sources responsible for such a change. To get a description of these changes, it is necessary to extract the relevant features. To obtain a better description of these features, it would be better to understand how these features get represented in the frequency and time domains. In what follows, the word *signal* means data set representing a process. The frequency description of a signal can be given by the Fourier transform of the signal, which represents approximation of the signal on a set of sine and cosine basis functions. The frequency content of the signal

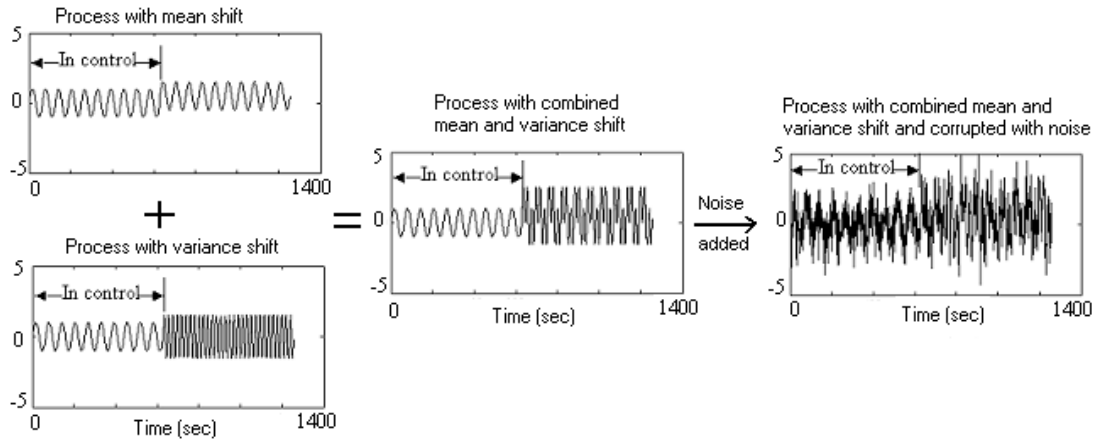

Figure 3.1. A Process With Multiple Features.

describes the nature of the signal and also provides a basis for using various filtering methods to analyze a signal on a particular frequency range. For example, a step change in a signal (i.e., a change in the process mean) indicates that signal is more localized in time but not in frequency. Then, several sine and cosine basis functions would be needed to approximate the signal. On the other hand, consider the signal noise with variance change. This signal can be approximated by a finite set of sine and cosine basis functions, and thus it is localized in frequency. Thus, a change in variance is more localized in frequency than in time domain. Hence, it is clear that different process features are better represented at various domains and thus should be examined accordingly. A common approach for an analysis of a process data set should be a time-frequency approach, which would describe the time localization as well as the frequency localization of the data.

The traditional method for analyzing data in the time-frequency space is done through filtering. Filters, are weight functions, which are used to remove or extract data over a range of frequencies. That is, the filters are used to extract various features of the data based on their frequency range. The best example would be the geometric moving average (MA) or the EWMA filter. Typically, an EWMA filter with a small value of filter parameter (λ) behaves like a low pass filter, i.e.,

it removes a wide range of high frequencies in the measurements. Hence, filtering using low pass filter would result in a smooth version of the signal. To extract time localized features the value of λ can be kept high. For example, a high value of $\lambda = 1$) would result in a Shewhart filter (which is also localized in time domain) and the features extracted will be more localized in time. For a step change (mean change), the frequency components are present all over the frequency domain. Hence, a step change would be retained as a step change. In order to extract features from other frequency range, it is necessary to vary the value of the filter parameter. There are problems associated with varying the value of the filter parameter. For example, if we use a high value filter parameter (like EWMA) we obtain measurements which usually contain noise and various changes in the signals. Hence, applying threshold or decision rules would result in improper detection of the change (false alarm). Also, applying a smaller filter parameter results in over smoothing and the change might go undetected even after applying appropriate decision rule. These problems arise due to the lack of adaptability of the filter parameter to the various changes in the process. Thus a method is needed that can adapt to the time and frequency domains and provide a complete time-frequency description of the signal. Recent developments in the wavelet decomposition methods have provided an effective tool for multiresolution analysis (MRA) of signals. The following sections, provides a brief description of the frequency domain methods, convolution theory, scale definition, scale space filtering, and an example of multiscale analysis. The wavelet expansion, dilation equation, and its multiresolution capabilities are discussed next.

3.2 Fourier Analysis

The basic concept of frequency-domain analysis is that a signal can be considered as the sum of sinusoidal basis functions. A continuous sine function $\sin(\omega t)$ is a single frequency wave of frequency ω radians/second, and the frequency

domain description consist of a single value at a particular frequency. The Fourier analysis is done for periodic and non-periodic signals. For periodic signals, the analysis is done by Fourier series and for nonperiodic signals, the analysis is done by Fourier transform.

3.2.1 Fourier Series

A periodic signal $f(t)$ can be decomposed on a set of sine and cosine basis function. The complex exponential form of a Fourier series is obtained by substituting the exponential equivalents of the sine and cosine terms into the original form of the series, which is given as follows.

$$f(t) = \sum_{n=-\infty}^{\infty} a_n e^{jn\omega_1 t}, \tag{3.1}$$

where a_n is the Fourier coefficient given as

$$a_n = \frac{1}{T} \int_{t_1}^{t_1+T} f(t) e^{-jn\omega_1 t} dt. \tag{3.2}$$

3.2.2 Fourier Transform

The Fourier transform is applied for nonperiodic signal. For a signal $f(t)$, the Fourier transform is given as follows

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt, \tag{3.3}$$

where

$$f(t) = \int_{-\infty}^{\infty} F(\omega) e^{j\omega t} d\omega. \tag{3.4}$$

The Fourier transform of a signal indicates the frequency content of the signal. A filter can be designed according to its Fourier transform. For a lowpass filter (which

filters out high frequency component), the Fourier transform of the filter should be bounded within a low frequency range and vice versa for a highpass filter.

3.3 Convolution Theory

For a good understanding of the filtering techniques, it is necessary to understand the convolution theory. This theory is also useful in finding densities of random variables. The following example illustrates a convolution with respect to density functions. Consider the random variable $Z = X + Y$. It is possible to estimate the probability density function of Z knowing the joint density $f_{XY}(x, y)$. The probability distribution function $F_Z(z)$ is given by

$$F_Z(z) = P(Z \leq z). \tag{3.5}$$

Since $Z = X + Y$, it can be written as

$$F_Z(z) = P(X + Y \leq z), \tag{3.6}$$

$$F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_{XY}(x, y) dx dy. \tag{3.7}$$

Then the density function is given as

$$f_Z(z) = \frac{d[F_z(z)]}{dz} = \int_{-\infty}^{\infty} f_{XY}(z - y, y) dy. \tag{3.8}$$

If the random variables X and Y are independent, then Equation 3.8 leads to the following simplification

$$F_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(Y) dy. \tag{3.9}$$

The above integral is the convolution of the probability density function $f_X(x)$ and $f_Y(y)$. In other words, if a random variable is expressed as the sum of independent random variables then the probability distribution of the former is given by the convolution of the density functions of the two independent random variables. In general, the convolution of two functions f and g is another function h and is given by

$$h(x) = \int_{-\infty}^{\infty} f(u)g(x-u)du, \tag{3.10}$$

which means that, for a particular value of X, say x_1, $h(x_1)$ is a linear sum of values of $f(u)$, weighted by the function $g(x-u)$.

Convolution theorem is used in filtering. For example, when a filter is applied to a set of measurements the output obtained is the convolution of the measurements with the filter. The following equation describes the operation for an EWMA filter

$$Z_i = \sum_{i=1}^{\infty} X_i h(n-i). \tag{3.11}$$

The numbers $h(n), \ldots, h(0)$ are filter coefficients that multiplies X. Convolution in time domain has a very important relationship in the frequency domain. The relationship is given by the following Fourier transform pair

$$f(t) * g(t) \Leftrightarrow F(f) \cdot G(f). \tag{3.12}$$

Thus, the convolution of two functions in time domain is equivalent to product of their Fourier transform. This implies that, if we convolve a function with a low or high pass filter, their Fourier transform gets multiplied. Hence, their output obtained will be based on the frequency range of the filters.

3.4 Properties Required for Multiresolution Analysis Methods

Multiscale analysis methods represent signals in the entire time-frequency space. However, for such a representation of the signals, it is necessary for the filters that are employed to have certain properties. These properties, as discussed in Bakshi and Stephanopoulous [19] are: 1) The filters should span the entire scale space (range of time and frequency). They should be able to extract various temporal features in a process: features that are localized in time and in frequency. 2) The relevant information should be more explicit in the multiscale representation than in the original data. For example, a step change in the process should be brought out clearly after applying the filter. The change is then detected by applying necessary threshold rule or decision rule. 3) The description at various scales should have minimum redundancy. This means that, as the value of the filter parameter is changed, the filter removes measurements having various frequency ranges. However, an important aspect when varying the filter parameter is that the information extracted should not be redundant in the various outputs. 4) The representation should allow data compression. Data compression is achieved by eliminating unnecessary information. In any process, the measurements are contaminated with noise. Therefore, the relevant features in the process such as changes in the mean and variance, and the trends in the mean are not explicitly described. The filtering method should filter the noise and any other irrelevant information. 5) The multiscale representation should allow stable and complete reconstruction. Stability allows the multiscale representation of the signal to be robust to minor changes or disturbances in the measurements, and reconstruction implies that no information is lost, and 6) extraction of various features in the signals should be possible and a good description of temporal features should be evident. In other words, the filter should be able to adapt to the signal feature, thus extracting the relevant features.

3.5 A Multiresolution Example

Consider a multiresolution analysis example shown in Figure 3.2. The Figure 3.2a represents the original signal and the set of the Figures 3.2b through 3.2g represent a multiresolution representation of the signal. Figure 3.2c shows the highest frequency component of the signal, which when extracted converts the original signal to what is shown in Figure 3.2b. Similarly, Figures 3.2e and 3.2g show the extracted high frequency components from data in Figures 3.2b and 3.2d respectively. Thus the data shown in Figures 3.2(c, e, and g) are the high frequency components of the data (also called *details*), and the remaining low frequency part of the data is shown in Figure 3.2f (also called the *approximations*). It may be noted that information contained in Figures 3.2(c, e, g, and f) can be combined to reconstruct the original signal. An extensive literature review of wavelet based multiresolution methodologies and their applications in statistical field, such as process monitoring can be obtained from Ganesan *et al.* [53].

3.6 Dilation and Translation

The dilation (scaling) of a function can be described by the Fourier representation of the signal. Consider the Fourier approximation of a function $f(t)$ as

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{j}(a_j cos(jt) + b_j sin(jt)). \tag{3.13}$$

The function is approximated by a set of sine and cosine basis functions for various j. It can be seen from Figure 3.3 that the sine and cosine functions at various j or levels represent the dilated versions of the sine and cosine basis functions. It is also observed that the function frequency increases with increase in the dilation level. Translation (shifting) of basis function is explained by Figure 3.4, which shows the translations of a Haar basis function.

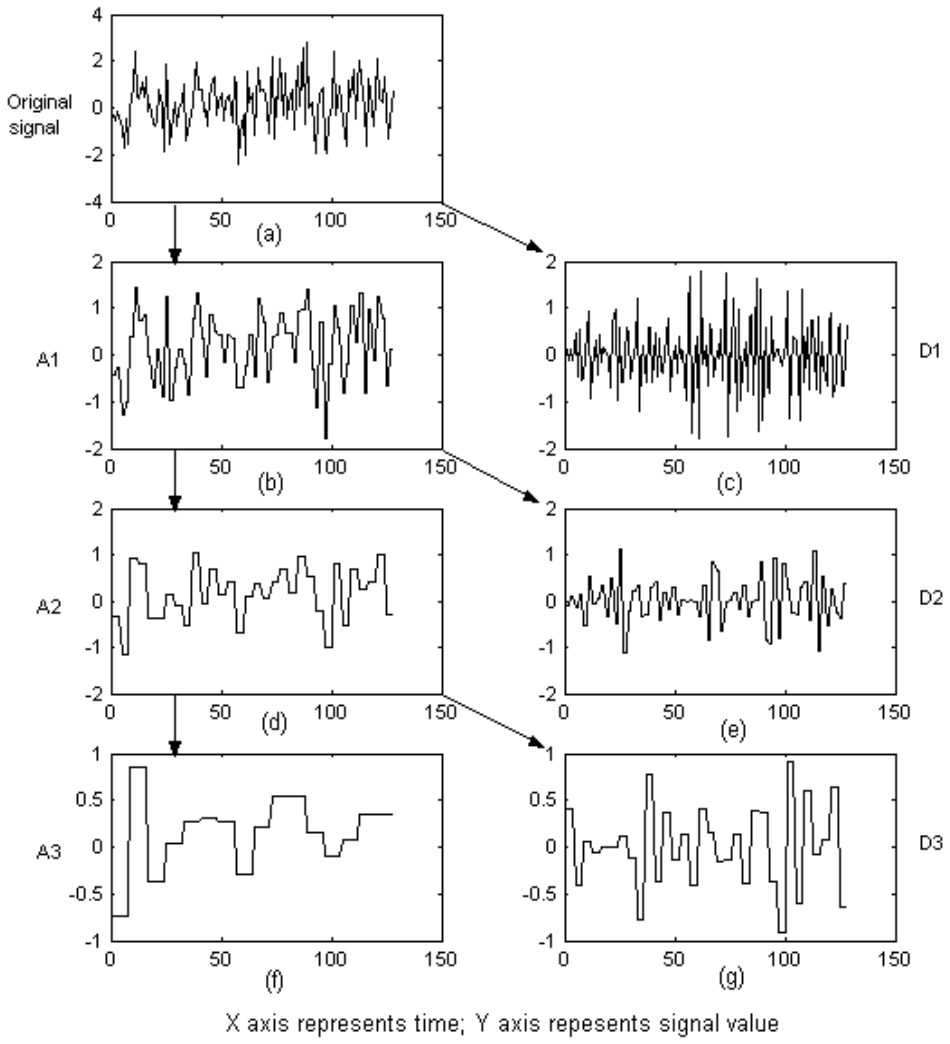

Figure 3.2. A Multiresolution Analysis Example.

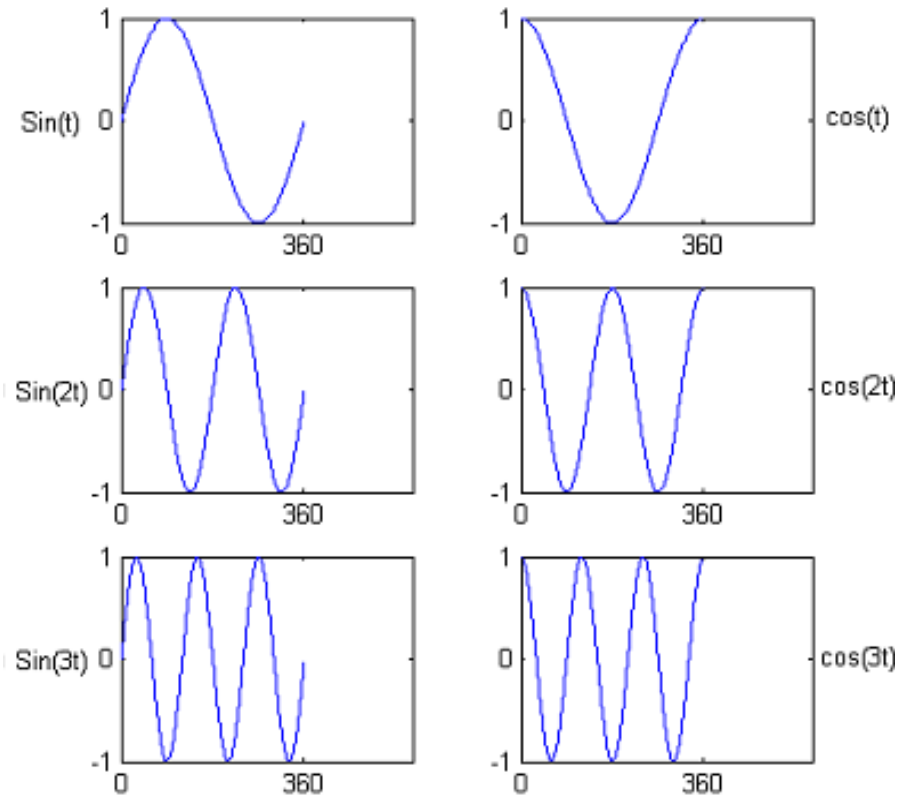

Figure 3.3. Dilations of $Sin(x)$ and $Cos(x)$.

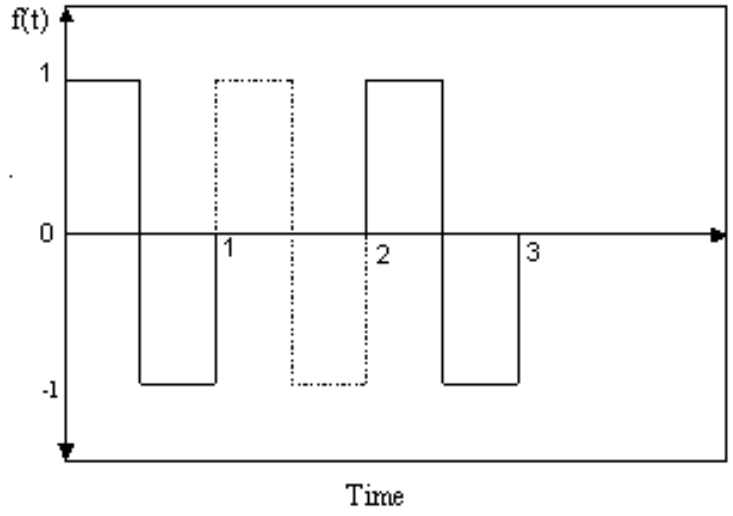

Figure 3.4. Translations of a Haar Basis Function.

3.7 Wavelet as a Multiresolution Analysis (MRA) Tool

The basic idea behind signal processing with wavelets is that the signal can be decomposed into its constituent elements through the use of basis functions. These basis functions can be obtained from the scaled (dilated) and shifted (translated) versions of the mother wavelet (ψ). The wavelet analysis uses linear combinations of basis functions (wavelets), localized both in time and frequency, to represent any function in the $L^2(R)$ space. For example,

$$f(t) = \sum_{j,k} b_{j,k}\psi_{j,k}(t), \tag{3.14}$$

where j and k are dilation (or scale) and translation indices respectively, $\psi_{j,k}$ denotes a collection of basis functions, and $b_{j,k}$ are the coefficients of these functions. The next section briefly describes the characteristics of a wavelet system.

3.8 Characteristics of a Wavelet System

There are three important characteristics for a wavelet system (Burrus *et al.* [54]): 1) The wavelet system is generated from dilation and translation of scaling functions (ϕ) (refer Subsection 3.8.1) or wavelets (ψ). The two dimensional parameterization is achieved from the function $\psi(t)$ by

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \qquad j, k \in Z, \tag{3.15}$$

where Z is the set of integers. The factor $2^{j/2}$ maintains the norm for the wavelet function. 2) All wavelet systems satisfy the multiresolution condition. In other words, if a signal can be expressed as a weighted sum of a larger set of basis functions, then the same signal can be expressed as a weighted sum of a smaller set of basis functions. The coefficients representing the larger set of basis functions are known as the high

resolution coefficients and those of the smaller set are the lower resolution coefficients, and 3) these lower resolution coefficients can be calculated from higher resolution coefficients by a tree-structured algorithm (pyramid tree) called *filter banks*.

3.8.1 Scaling Function

The scaling function forms the basic step for the derivation of wavelet bases. A set of scaling functions, in terms of its translates, is given by

$$\phi(t) = \phi(t-k), \qquad k \in Z, \qquad \phi \in L^2(R). \tag{3.16}$$

The subspace of $L^2(R)$ spanned by the scaling functions is defined as

$$V_0 = span\{\phi_k(t),\ k \in Z\}, \tag{3.17}$$

which implies that

$$f(t) = \sum_l a_l \phi_l(t), \qquad \text{for} \qquad f(t) \in V_0. \tag{3.18}$$

A two dimensional family of scaling functions are generated from the basic scaling function through their dilations and translations as given by

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k), \tag{3.19}$$

where

$$V_j = span\{\phi_{j,k}(t),\ k \in Z\}, \tag{3.20}$$

which implies that

$$f(t) = \sum_l a_l \phi_l(2^j t + k), \qquad \text{for} \qquad f(t) \in V_j. \tag{3.21}$$

For larger values of j, the scaling function has a larger span hence it can represent the signal in more number of steps.

3.8.2 Multiresolution Analysis

Multiresolution provides a formal approach for constructing orthonormal basis functions. It provides a particular framework for the understanding of wavelet bases. The idea behind multiresolution analysis is to express a function or signal as a limit of successive approximations and each of the approximation gives a smoother version of the function or signal. These successive approximations correspond to different resolution, hence it has the name multiresolution. The multi-resolution analysis of $L^2(R)$ is defined as a sequence of closed subspaces V_j of $L^2(R)$, $j \in Z$, with the following properties.

Completeness in $L^2(R)$.

$$\ldots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \ldots , \tag{3.22}$$

$$\cap_j V_j = \{0\}, \ \overline{\cup_j V_j} = L^2(R). \tag{3.23}$$

This implies that the space that contains the higher resolution signals would also contain the lower resolution signals.

Scale Invariance.

$$f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}, \tag{3.24}$$

This implies that elements in a space are scaled versions of elements in the next space.

Shift invariance.

$$f(x) \in V_j \Leftrightarrow f(x+k) \in V_j, \qquad k \in Z. \tag{3.25}$$

V_0 *has an orthonormal basis* $\phi(t-k)$, and there exists a scaling function $\phi(t) \in V_0$ such that

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k) \qquad j,k \in Z, \tag{3.26}$$

which satisfy the property of *Riesz basis* and are linearly independent.

¿From this it is known that $\phi(t) \in V_0 \in V_1$, and $\phi(2t)$ is a basis for the subspace V_1 or in other words, V_1 is the space spanned by $\phi(2t)$. This implies that $\phi(t)$ can be expressed as a weighted sum of shifted $\phi(2t)$ as

$$\phi(t) = \sum_k h(k)\sqrt{2}\phi(2^j t - k) \qquad n \in Z, \tag{3.27}$$

where the coefficients $h(k)$ are a sequence of numbers which are called the scaling function coefficients (scaling filter or scaling vector) and $\sqrt{2}$ maintains the norm. This recursive equation is of primary importance in derivation of the scaling function, which is the complete basis set. The recursive equation is called the *Dilation equation* or *refinement equation* or *two-scale equation.* The scaling function has the following property

$$\int_{-\infty}^{\infty} \phi(t)dt = 1 \qquad \text{and} \qquad \sum_k h(k) = 1. \tag{3.28}$$

3.8.3 Wavelet Function

The resolution of a signal can be described by scaling functions and they describe smoother versions of the signal. In order to identify differences between the scaling spaces (V_n) a different set of functions have to be defined. These functions are called wavelet functions (ψ) and they span the differences between the spaces spanned by the scaling function. The space representations by wavelet functions are orthogonal complements of the scaling functions.

The orthogonal complement of V_n in V_{n+1} is given by W_n. If the wavelet spanned space is defined such that

$$V_1 = V_0 \oplus W_0, \tag{3.29}$$

then

$$V_2 = V_0 \oplus W_0 \oplus W_1. \tag{3.30}$$

Thus, it can be written that

$$L^2(R) = \bigoplus_{j \in Z} W_j = V_{j_0} \bigoplus_{j \geq j_0} W_j. \tag{3.31}$$

The wavelet space $W_0 \subset V_1$, and can be represented as a weighted sum of translated $\phi(2t)$ as follows

$$\psi(t) = \sum_k h_l(k)\sqrt{2}\phi(2t-k), \qquad n \in Z, \tag{3.32}$$

where $h_l(k) = (-1)^k h(1-k)$. This wavelet function generates other equations for the set of expansions of the form

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k), \qquad j,k \in Z. \tag{3.33}$$

3.8.4 Discrete Wavelet Transform

According to the multiresolution analysis given by Equation 3.31, any function $f(t) \in L^2(R)$ can be written using Equation 3.19 (scaling expansion) and Equation 3.33 (wavelet expansion) as

$$f(t) = \sum_k c_{j_0}\phi_{j_0,k}(t) + \sum_k \sum_{j=j_0}^{\infty} d_{j,k}\psi_{j,k}(k). \tag{3.34}$$

The coefficients in the wavelet expansion are called the discrete wavelet Transform (DWT) of the function $f(t)$. If the wavelet system is orthogonal then the coefficients

can be calculated by

$$c_{j_0} = \langle f(t), \phi_{j_0,k}(t) \rangle = \int f(t)\phi_{j_0,k}(t)dt, \tag{3.35}$$

$$d_{j,k} = \langle f(t), \psi_{j,k}(t) \rangle = \int f(t)\psi_{j,k}(t)dt. \tag{3.36}$$

3.8.5 Mallat's Decomposition and Reconstruction Algorithms

In most of the applications, one cannot derive the scaling and wavelet functions easily. It is desirable that an expression for the scaling ($c_{j,k}$) and wavelet coefficients ($d_{j,k}$) be attainable without the need to construct $\phi(t)$ or $\psi(t)$. To do this, Mallat [55] developed a very efficient algorithm (*Pyramidal Algorithm*) for multiresolution analysis using wavelets. This is also referred to as the fast wavelet transform (FWT). Essentially, the scaling coefficients at the higher resolution are used to calculate the scaling coefficients at the next resolution. A similar procedure is used to obtain the wavelet coefficients. However, it should be noted that the above procedure is possible, provided the scaling function coefficients at some resolution level j is known. One such practice is to discretely sample the scaling function coeffient at level j from the function $f(t)$. However, Strang and Nguyen [28] refer to this pro cedure as being *a wavelet crime* but also mention that it is a convenient practice to employ. The Mallat algorithm works perfectly (i.e., obtains perfect reconstruction) if orthogonal wavelets are used. Thus, MRA provides a hierarchical and fast scheme for the computation of the DWT. In signal processing language this type of algorithm is studied as *filter banks.* Filter Banks are structures, which decompose a signal into various subsignals. This algorithm depends on two sequences called the lowpass and highpass filters. The computation of coefficients is called the analysis phase and the reconstruction is called the synthesis phase. These phases are the steps in filter

banks (subband filtering). A detailed description of the above algorithm steps and the filter bank theory can be found in Strang and Nguyen [28].

3.9 Sequential Probability Ratio Test (SPRT)

The role of statistical quality control is to provide decision tools that support production and maintenance activities, and this is achieved through a quality monitoring system (QMS). It is well known that the details from wavelet reconstruction are usually very small in magnitude and changes in these details due to an assignable cause are even smaller. Thus, it is essential to have a very sensitive, efficient, and computationally less intensive QMS that can be implemented on a real time basis. These requirements can be achieved through control charts that use sequential probability ratio test. Another important property of the SPRT is its optimality in reference to the average sampling number (ASN). That is, SPRT provided the minimum sample sizes under any value of θ within the hypothesis. The SPRT is designed to work with wavelet details, which are normally distributed.

3.9.1 Notion of a Sequential Test

In the current theory of testing hypotheses, the size of the sample on which the test is based is treated as a constant for any particular problem. An essential feature of the sequential test, as distinguished from the current test procedure, is that the number of observations required by the sequential test depends on the outcome of the observations, and is, therefore, not predetermined, but a random variable. The sequential method of testing a hypothesis H is described as follows. A rule is given for making one of the following three decisions at any stage of the experiment (at the m^{th} trial for each integral value of m): 1) to accept the hypothesis H, 2) to reject the hypothesis H, 3) to continue the experiment by making an additional observation. Thus, such a test procedure is carried out sequentially. On

the basis of the first observation one of the aforementioned three decisions is made. If the first or the second decision is made, the process is terminated. If the second decision is made, a second trial is performed. Again, on the basis of the first two observations one of the three decisions is made. If the third decision is made, a third trial is performed and so on. The number n of observations required by such a test procedure is a random variable, since the value of n depends on the outcome of the observations. For each positive integral value m, denote the totality of all possible samples $(x_1, ..., x_m)$ by M_m. M_m is also referred to as the m-dimensional sample space. A rule for making one of three decisions at any stage of the experiment is described as follows. For each integral value m, the m-dimensional sample space is split into three mutually exclusive zones, R_m^0,R_m^1, and R_m. After the first observation x_1 is drawn, the hypothesis H that is being tested is accepted if x_1 lies in R_0^0, rejected if x_1 lies in R_1^1, or a second observation is made if x_1 lies in R_1. If the third decision is made and a second observation x_2 is drawn, H is rejected, accepted, or a third observation is drawn, according as the observed sample (x_1, x_2) lies in R_2^0, R_2^1, or R_2. This process is stopped only when the first or the second decision is made. Thus a sequential test is completely defined by defining the sets R_m^0, R_m^1, and R_m for all positive integral values of m. The sets R_m^0, R_m^1, and R_m ($m = 1, 2, \ldots$) defining a sequential test need to be selected properly. The principles for a proper choice of the sets necessitate study of the consequences of any particular choice [56].

3.9.2 SPRT for Testing Two Simple Hypotheses

Wald [56] designed SPRT as a statistical tool for deciding between two simple hypothesis. If a random variable X is distributed $f(x, \theta)$, it is possible to construct the simple hypothesis $H_0 : \theta = \theta_0$ with $H_1 : \theta = \theta_1$ using SPRT. This test is based on the Neyman-Pearson (N-P) Lemma [57], which states that, for a fixed sample size of n, the optimal design (most powerful test) for simple hypothesis can

be obtained from the likelihood ratio (λ_n) as follows:

$$Accept\ H_o \qquad if \quad \lambda_n < k \tag{3.37}$$

$$Accept\ H_1 \qquad if \quad \lambda_n \geq k, \tag{3.38}$$

where

$$\lambda_n = \prod_{i=1}^{n} \frac{f(x_i, \theta_1)}{f(x_i, \theta_0)}, \tag{3.39}$$

k is the decision limit associated with significance α, and i denoted the observation index. The SPRT based on the N-P Lemma is given as follows. When the n^{th} sample is being tested, SPRT makes one of the following three decisions:

$$Accept\ H_0 \qquad if\ \lambda_n < A \tag{3.40}$$

$$Accept\ H_1 \qquad if\ \lambda_n \geq B \tag{3.41}$$

$$Continue\ Sampling \qquad otherwise \tag{3.42}$$

The third decision implies that another observation must be added to the sample and that the test must be performed again. This sequential procedure stops whenever H_0 is accepted or rejected. SPRT can be designed for different density function f and parameter θ. It may be possible to obtain the magnitude of the α and β errors for a given value of A and B. However, Wald showed that an approximation of the errors could be obtained using just the decision limits A and B as follows [58].

$$\alpha = (1 - A)/(B - A) \tag{3.43}$$

and

$$\beta = A(B - 1)/(B - A) \tag{3.44}$$

It has been shown that the above equations overestimate α and β [56].

3.9.3 SPRT for Variance

The SPRT based on N-P Lemma uses two decision limits, upper and lower instead of one. Consequently, there are three decision zones. The hypothesis for the test of variance when the mean is known is set as follows:

$$H_0 : \sigma^2 = \sigma_0^2 \tag{3.45}$$

$$H_1 : \sigma^2 = \sigma_1^2 \quad (\sigma_0^2 < \sigma_1^2), \tag{3.46}$$

For a population $X = N(\mu, \sigma^2)$, where σ_0 and σ_1 are design parameters for in-control and out-of-control standard deviation values. Suppose that, after $n-1$ observations, the test has indicated that there is no evidence for accepting or rejecting H_0. Define λ_n^σ, the n^{th} likelihood ratio for testing the variance as

$$\lambda_n^\sigma = \frac{L(x_1, x_2, ..., x_n; \mu, \sigma_1^2)}{L(x_1, x_2, ..., x_n; \mu, \sigma_0^2)} \tag{3.47}$$

The three decision criteria are as follows.

$$Accept\ H_0\ (Reject\ H_1) \qquad if \quad \lambda_n^\sigma < a_\sigma \tag{3.48}$$

$$Reject\ H_0\ (Accept\ H_1) \qquad if \quad \lambda_n^\sigma > b_\sigma \tag{3.49}$$

$$Stay\ undecided\ and\ keep\ on\ sampling \qquad otherwise \tag{3.50}$$

where a_σ and b_σ are design variables. The region between a_σ and b_σ limits is also referred as the zone of indifference. Wald also showed that approximate magnitude of α and β errors associated with a test can be obtained using just the detection limits a_σ and b_σ as

$$\alpha \approx \frac{1 - a_\sigma}{b_\sigma - a_\sigma} \tag{3.51}$$

$$\beta \approx \frac{a_\sigma(b_\sigma - 1)}{b_\sigma - a_\sigma}. \tag{3.52}$$

Taking logarithms of Equation 3.47, the log-likelihood ratio is computed. Step by step details of this procedure can be found in [59]. Thus, the SPRT for the variance with known mean is as follows.

$$Accept\ H_0 \qquad if\ \ \hat{\sigma}_n^2 < \frac{R_1}{R_2} + \frac{h_\sigma}{nR_2} \tag{3.53}$$

$$Reject\ H_0 \qquad if\ \ \hat{\sigma}_n^2 > \frac{R_1}{R_2} + \frac{k_\sigma}{nR_2} \tag{3.54}$$

$$Stay\ undecided\ and\ keep\ on\ sampling \qquad otherwise \tag{3.55}$$

where

$$h_\sigma = \ln(a_\sigma) \tag{3.56}$$

$$k_\sigma = \ln(b_\sigma) \tag{3.57}$$

$$R_1 = \ln(\frac{\sigma_1}{\sigma_0}) \tag{3.58}$$

$$R_2 = \frac{1}{2}(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}) \tag{3.59}$$

$$\hat{\sigma}_n^2 = \frac{\sum_{k=1}^{n}(x_k - \mu)^2}{n}. \tag{3.60}$$

Chapter (5) explains in detail the real-time strategy using wavelet based multiresolution to analyze sensor data and SPRT for variance as a statistical tool to detect the occurrence of undesirable events.

CHAPTER 4
RESEARCH OBJECTIVES

4.1 Introduction

This chapter gives an introduction to the type of data that will be analyzed and the reasons for developing a Wavelet based multiscale analysis strategy for real time monitoring of the process.

4.2 Problem Description

During the oxide and metal CMP process, the Acoustic Emission (AE) signals generated are sensed by an acoustic sensor and discretized using a data acquisition system. The data is generated at a very high rate (approximately 190 data points per second) and are multiscale in nature. This data carries a lot of information about the process in a very short period of time and at different levels of frequency. Analysis of such data has been done offline using Various multiscale analysis methodologies. An online strategy has also been proposed but is computationally incapable of analyzing the data real time. To the best of my knowledge, none of the existing methods are capable of real time analysis of the data that is generated at the rate mentioned above because of the high computation time taken by them. It may be feasible to use the existing strategies by skipping data, but in doing so vital information may be lost considering the data generation rate. For example, skipping 10 seconds of the processing time will result into skipping of 1900 data points. Since real time analysis of the data is not possible with the existing techniques, it is not only difficult to understand and monitor the process but also makes it impossible to

detect abnormal events during processing, as a result affecting the quality and the performance of the product. In order to monitor the process real time and improve the quality and performance of the product, it is necessary to develop a strategy that is computationally fast to analyze each and every data point at different levels of frequency and also be able to detect events of interest. With the objective of overcoming the shortcomings of the exisiting strategies for real time analysis of multiscale data generated at a high rate, the objective of this research is to: 1) develop a fully online wavelet based multiscale analysis strategy where the speed of wavelet based analysis approach matches the rate of data generation 2) to develop a statistical tool based on Sequential Probability Ratio Test (SPRT) to detect events of interest, and 3) develop an approach to display the analysis results through real time graphs for the ease of supervisory decision making. The research proposes to develop a strategy programmed using MATLAB 6.5 and Wavelet toolbox consisting of 1) decomposing the input signal, 2) reconstructing the details at each level of frequency, and 3) implementing a moving block strategy to analyze the signal real time by reducing the computation time and, 4) applying Sequential Probability Ratio Test (SPRT) for variance of the details at each level of frequency to detect abnormal events, and 5) develop an approach to display the analysis results through real time graphs .

CHAPTER 5

APPLICATION OF WAVELETS AND SPRT FOR REAL-TIME ANALYSIS OF SENSOR DATA

5.1 Introduction

In order to accomplish the research objectives stated in the previous chapter, a wavelet based strategy has been developed for real-time analysis of the sensor data and detection of events of interest by applying SPRT. The strategy was coded using Matlab 6.5 and Wavelet toolbox functions. The sequence of the steps involved in implementing this strategy is as follows:

1. Interface the sensor system with the Data Acquisition (DAQ) system.
2. Interface the DAQ with Matlab to transform the sensor data in a format recognized by Matlab.
3. Group the transformed data in sets termed as block having the desired dyadic length.
4. Execute the Matlab code:
   (a) Read the block of transformed sensor data from its stored location into the Matlab code to analyze it.
   (b) Decompose the data block using wavelets at different levels.
   (c) Reconstruct the wavelet coefficients to get the details at each level.
   (d) Apply a Sequential Probability Ratio Test (SPRT) on the details.

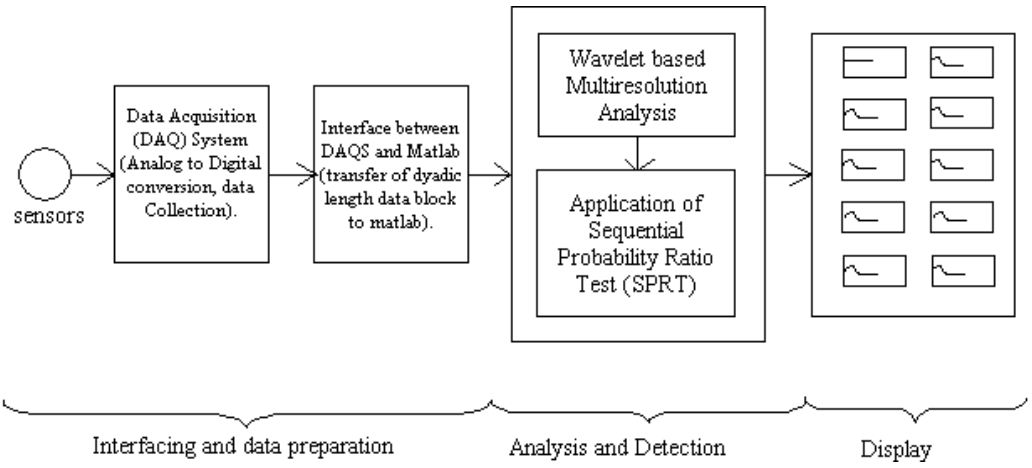

Figure 5.1. Schematic Diagram of Various Stages of the Real-Time Methodology.

(e) Display real-time multi-level plots of the details and SPRT charts that indicate state of the process and raise a flag at the occurrence of an undesirable event.

(f) Stop the plotting at the instance an undesirable event occurs or continue with the analysis of the new data block.

5. Repeat steps (1) to (4) to continue analyzing the data in blocks until an undesirable event occurs or the process under study ends.

In what follows, a detailed description of the steps above is provided.

5.2 Description of Real-Time Implementation of the Methodology

Figure 5.1 shows a schematic diagram of the various stages of the proposed real time methodology, which includes 1) interfacing and data preparation, 2) analysis and details, and 3) display of results, which are described below.

5.2.1 Interfacing and Data Preparation

5.2.1.1 Sensing

Sensing of the various environmental properties such as sound, force, light etc. during any physical phenomena is done by sensors. The role of sensors is to convert the physical phenomena into electric signals and transmit the signal to a Data Acquisition (DAQ) system. The most commonly used sensors for process monitoring and control applications in manufacturing and nano-manufacturing are acoustic emission (AE) sensors, force sensors, ultrasonic sensors and many more. The AE sensors are the most popular choice amongst them. AE sensors operate on the principle of sensing the acoustic emission signals that are generated by the rapid release of energy from localized sources within materials such as metals due to dislocation movement accompanying plastic deformation, thermal stresses etc. The major advantage of using AE sensors to monitor a manufacturing process is that the frequency of the acoustic emission (AE) signal is much higher than those of machine vibrations and environmental noises, and thus provides a good representation of the manufacturing process. The sensitivity (signal to noise ratio) of AE signals increases as the amount material removal reduces making it more suitable to be used in sensing micro to nano scale precision machining operations.

5.2.1.2 Data Acquisition (DAQ) System

The task of acquiring the sensor data and storing it automatically is done by a DAQ system. The electric signals from the process in analog form are acquired by a PC-based DAQ system and subsequently digitized. This data can be analyzed using different computer analysis software for the purpose of monitoring and/ or controlling the system. The performance of the PC based DAQ system depends on various elements like the PC, signal conditioning, DAQ hardware software. The

processing speed and the hard drive capacity of the PC has a significant impact on the maximum speed at which the data is acquired continuously. A decision on the type of DAQ system to be used depends on the specifications such as sampling frequency, resolution, and accuracy. Sampling frequency is the frequency of analog input signal sampling and conversion to a digital value. Resolution refers to the degree of fineness of digital word representing analog value. Accuracy depends on parameters like signal conditioning, linearity, hysterisis, and temperature consideration.

5.2.1.3 Interface between Data Acquisition (DAQ) System and Matlab

Software is used to interface between DAQ System and Matlab. The role of the software is to act as a link between the DAQ system and Matlab to enable transfer of data from the DAQ system into Matlab. As soon as the length of the data digitized by the DAQS attains the desired dyadic length, the interfacing software reads this data, converts it into a format recognized by Matlab as its input and saves the converted data in a matlab (.m) file. The matlab file acts as the first block of input to the Matlab code. Once the input is available, at that instance the matlab code is executed automatically to analyze the input data. While the first block is being analyzed, the DAQ system continues to acquire new data from the process. Once the new data attains the desired dyadic length, the interfacing software converts the data block and saves it in a new matlab file for further processing. This cycle of sensing, acquisition of data, digitizing, and converting to matlab file is repeated until any undesirable event is detected or the process under study ends.

5.2.2 Analysis and Detection

The methodology to detect undesirable events real-time has been coded using Matlab 6.5 and the Wavelet Toolbox functions. The code is executed whenever a data block of the desired dyadic length is available as input or the analysis of the

previous block is completed. The block of input data of dyadic length is read from the matlab (.m) file. The desired dyadic length of the data block is depends on the extent of decomposition required to obtain relevant process information, which can vary depending on the nature of the process. For instance, if the important process information is known to contain within the first eight levels, then the required length of data is 256 (2^8). The analysis starts with the first data block followed by the next data block. This cycle continues until an undesirable event is detected or the process under study ends. The length of the data block is kept constant through the course of block by block analysis. Since the process data is selected and analyzed in blocks with a constant dyadic length, the strategy is termed as a moving block strategy. The moving block strategy is used as against the moving window strategy proposed by Ganesan *et al.* [52]. as is computationally capable of analyzing the process with a high rate of data generation on a real-time basis. In the moving block strategy, not every data point of the block is located at a dyadic location as is the case in the moving window strategy. This prohibited uniform discretization. However, a comparison of energy level plots at different levels for the moving window strategy and the moving block strategy have shown that the inability to maintain uniform discretization while using the moving block strategy does not result in any significant difference in the energy plots displayed or the information content at each level. In the moving window or the moving block strategy the wavelet coefficients are no longer orthonormal to each other and they lose the property of approximately decorrelating the autocorrelated data. After reading the data block as input, it undergoes wavelet analysis.

5.2.2.1 Multiresolution Analysis (MRA) Using Wavelets

Multiscale wavelet analysis is done on the entire block starting with its decomposition up to the desired levels using wavelets. The depth of decomposition of

the data in the time domain is fixed to obtain wavelet coefficients at each level. The depth of decomposition is critical as excessive decomposition means more plots and increase in the computational time, whereas inadequate decomposition could result in inability to identify some process events at lower frequencies. The signal energy plots indicating the extent of information available at each level has been shown to be useful ([52]) in deciding the maximum depth of decomposition. Several wavelet basis function types such as Haar, Daubechies, coiflets, symlets etc. are available in literature. The application of wavelets depends on nature of the input data to be analyzed and the extent and accuracy of information desired. For instance, the Haar wavelet has a compact support but it does not have good time-frequency localization. It is also unsuitable for representing classes of smoother functions due its discontinuities. Haar wavelets are best suited to represent step signals or piecewise constant signals. Daubechies family of wavelets is the most widely used basis functions because they have good time-frequency localization and a large number of vanishing moments (ensures maximum number of zeros of the polynomial at the highest discrete frequency). Daubechies wavelets are better for smoother signals. After decomposing the block of data at all levels, the wavelet coefficients undergo reconstruction one level at a time to obtain the details at each level. After reconstruction of the wavelet coefficients to get the details, SPRT for variance is conducted.

5.2.3 Implementation of SPRT for Variance

SPRT for variance test to detect an undesirable event is implemented as follows:

$$Reject\ H_0(Accept\ H_1 : Undesirable\ event\ occured)\ if\ \hat{\sigma}_n^2 > \frac{R_1}{R_2} + \frac{k_\sigma}{nR_2} \quad (5.1)$$

$$Stay\ undecided\ and\ keep\ sampling\ otherwise \quad (5.2)$$

where R_1, R_2, $\hat{\sigma}_n^2$, and k_σ are as defined earlier.The design parameters of the SPRT chart are σ_0, σ_1, α, and β. The α error (false alarm) occurs when a point falls outside the upper control limit and the H_0 is rejected when undesirable event is not reached. Similarly, a β error occurs when a point falls below the lower control limit and H_0 is accepted, even though the true process is out of control. However, in the case of SPRT for the variance of wavelet details, an increase in variance (point falling outside the upper control limit) is of concern to us. Hence, a point falling below the lower control limit does not signal any alarm, and consequently β error does not have any significance. The above can also be proved by fixing α and varying β. It is noticed that as β is increased, the lower control limit moves closer to the centerline while the upper control limit remains practically unchanged. A similar result is observed when α is increased maintaining β constant, where the upper control limit moves closer to the centerline and the lower control limit remains the same. A higher value of α naturally means a higher chance of false alarm. Thus, a small value of α is preferred.

The values for σ_0 and σ_1 are obtained from the wavelet details for each block at each level. Since, the variance is plotted, it is quite natural to fix these design parameters through an S-chart. The value of σ_1, the upper limit, is first fixed using 3 sigma criteria.

It is noted from Equation 5.1 that the region below the upper limit is the zone of indifference. Thus, σ_0 value can be fixed anywhere below σ_1. It was found that maintaining σ_0 close to σ_1 showed good detection of the undesirable event.

$$\sigma_1 = \bar{S} + 3\bar{S}(\sqrt{1 - c4^2})/c4 \tag{5.3}$$

$$\sigma_0 = \bar{S} + 2.95\bar{S}(\sqrt{1 - c4^2})/c4 \tag{5.4}$$

$$c4 = \frac{4(n-1)}{4n-3}, \tag{5.5}$$

where $c4$ is a constant which depends on the sample size n, and $\bar{S}$ is the mean standard deviation. Starting with the first data block, the standard deviation of the wavelet details is calculated and is taken as $\bar{s}$ value. This is used to calculate σ_0 and σ_1, which are in turn used to calculate the upper and lower limits. The variance of the wavelet details at any point in the current block is calculated by taking all the data points preceding and including the current point. For example, at level 1, for the first block of dyadic length 256, if n = 5, then the variance is calculated over 5 data points. Similarly, for the same level, if n = 250, the variance is calculated over 250 data points. When SPRT proceeds further to the second block or any further block, large volumes of previous data would have to be stored which causes a decrease in computational speed. Storing detail values of only one previous block and calculating the current variance based on preceding data, which extends only up to the first value of the previous block, can avoid this situation. For example, for the second data block of length 256, at n = 5 the variance of details is calculated over a set of 256 data points. The first 251 points belong to most recent 251 data points from the preceding block and the remaining 5 data points belong to the current block. It is also observed that as SPRT proceeds both σ_0 and σ_1 values stabilize as $\bar{S}$ value stabilizes. Even though SPRT control limits for every window is drawn from the data itself, the sensitivity of the limits are maintained by the averaging effect in $\bar{S}$ calculation. Thus, when an undesirable event occurs, the standard deviation of the details increases in that window but the $\bar{S}$ value is not affected. Consequently, the SPRT control limits retain its sensitivity. This combined effect causes the variance plot of the details to cross the upper control limit of the SPRT chart and raise a flag to indicate the occurrence of an undesirable event.

5.2.4 Testing of SPRT for Variance of Details

After calculating the acceptance limits, rejection limits and the variances values of detail for a data block at all levels, SPRT for variance of the details is applied. Starting with the first level, each value of variance at position n is compared with each value of the rejection limit at the same position n. If at any given position n, the variance value exceeds the rejection limit, the test fails (rejects the null hypothesis), indicating an occurrence of an undesirable event. If the test does not fail at a given level, it is applied to the next level. The testing continues until the criteria for failing is met for the first time at any given level.

5.2.5 Display - Multilevel Plotting and SPRT for Variance

One of the unique feature of real-time methodology is its ability to display real-time SPRT plots enabling *in-situ* process control and displaying of the event of interest at the instance it occurs. In this section, the strategy of plotting the results depending upon the outcome of SPRT is explained.

Plotting of the details and the SPRT chart is based on the outcome of the test described above. Plotting starts from the first level. For any level, if the variance of the details at any position n exceeds or equals the rejection limit at the corresponding position n, then plots of the details and SPRT chart are displayed only till the point where the test failed at that level. A square box or a flag can be seen in the plot for SPRT for variance at that level indicating the time and scale at which the undesirable event has occurred and as a result terminates the program. No further plotting is done for succeeding levels. If the SPRT test does not fail for the entire data block at the level it is being conducted, the details and SPRT chart for the entire block at that level is displayed. After plotting the details and SPRT chart for the level SPRT did not fail, the variance data at the succeeding level undergoes SPRT. The testing continues till the last level is reached.

5.2.6 Repetition of Steps (1) to (4)

For the current data block, if the SPRT for variance does not fail and plots of details and SPRT chart are displayed at all the levels, the analysis moves over to the new data block. This cycle continues until a flag is raised to indicate occurrence of any undesirable event or the process under study ends.

5.3 Salient Features of the Moving Block Strategy

1. The moving block strategy has been found to be computationally capable of dealing with very high data collection rates. The algorithm has been developed by writing a code that is efficient and does not necessitate storing of huge data in the temporary memory of the computer. After every iteration, the volatile memory is cleared ensuring a constant processing speed as the number of data points collected from the process under study increases.

2. A complete understanding and monitoring of the process at different scales can be achieved by observing the real-time multi-level plots

3. It uses Sequential Probability Ratio Test (SPRT) for variance as a statistical tool to detect undesirable events.

4. It is capable of displaying the occurrence of undesirable events on a real-time basis by terminating the program and raising a flag in the form of a square box and thus eliminating the need for constant monitoring of the plots.

5. Excellent time-frequency localization enables in exactly knowing when the event of interest has occurred.

6. The computing requirement of the methodology is nominal as per todays PC standards. The methodology was tested offline on a data set representing a 5 minute long CMP process. The computation time taken was 5.5 minutes on a

computer having a Pentium IV processor, 2.88. GHZ processing speed and a 512 MB RAM.

7. The coding for this methodology has been developed using Matlab 6.5 and Wavelet Toolbox that are extremely popular and are readily available at a reasonable price.

CHAPTER 6

APPLICATION OF THE REAL-TIME METHODOLOGY TO A CHEMICAL MECHANICAL PLANARIZATION (CMP) PROCESS TO DETECT DELAMINATION

6.1 Introduction

In this chapter, an introduction to the CMP process has been provided, followed by the description of various defects arising during the CMP process and the application of the real-time methodology to the CMP process to detect the delamination defect.

6.2 Introduction to Chemical Mechanical Planarization (CMP)

Today's high performance circuits contain many areas with a high density of narrow and tall features, which in turn produces dense and high aspect ratio topography on the surface. When such topography exists, for example on an inter-level dielectric (ILD) film deposited over one layer of metal, it makes it impossible to directly deposit the next metal layer. Hence, multilevel metallization (MLM) is difficult to achieve. A typical MLM structure is shown in Figure 6.1 for which the goal of CMP is to reduce the topography on the wafer, and achieve global planarization. Introduction of new inter-connect materials and dielectric materials are necessary to keep pace with features size reduction and increased performance. In view of this requirement, copper is expected to be a new inter-connect material due to its performance superiority over aluminum. With copper, there is a greater control of electro-migration defect, and due to its higher conductivity than aluminum, it

is possible to make inter-connect lines smaller and yet provide the same current carrying capability. Additionally, using low-k dielectrics between metal lines makes it possible to have tighter packing density within a layer, which allows the use of fewer layers in circuits. The CMP process is the only existing method for making patterning lines of material such as copper, which is difficult to etch and pattern using the conventional dry etching methods. This is achieved by "Copper damascene approach", by depositing copper on silica, and subsequently using CMP to polish the top continuous over layer of copper (conventionally done by depositing dielectric over copper and polishing off the excess dielectric deposits). A schematic diagram for the process step to fabricate copper lines by conventional and copper damascene CMP process is shown in Figure 6.2. Other applications of CMP include, polishing off the excess tungsten (W) deposits that are used to make connections between metal layers and in the shallow trench isolation (STI) process, which is an alternate method of forming isolation regions between active devices. The goal in STI-CMP is to polish the oxide overburden, and to stop as soon as the nitride is exposed. Hence, the CMP process synergistically combines both tribological and chemical effects to planarize metals like copper and tungsten and insulating materials like silica and polymers.

A schematic diagram of CMP process is shown in Figure 6.3. The material removal rate (MRR) in CMP is usually in the range of 100-800 nm/min. in thickness, which is extremely small in comparison to conventional machining processes. The general mechanism of material removal in CMP is chemical etching and mechanical abrasion. The complexities of the CMP process are due to the large number of factors that influence the material removal process, such as the down pressure, back pressure, frictional forces, velocity, consumable parameters, such as the pad topography, pad materials and geometry, slurry abrasive size, slurry abrasive geometry, slurry chemicals and viscosity, wafer parameters, such as materials, geometry, pattern layout,

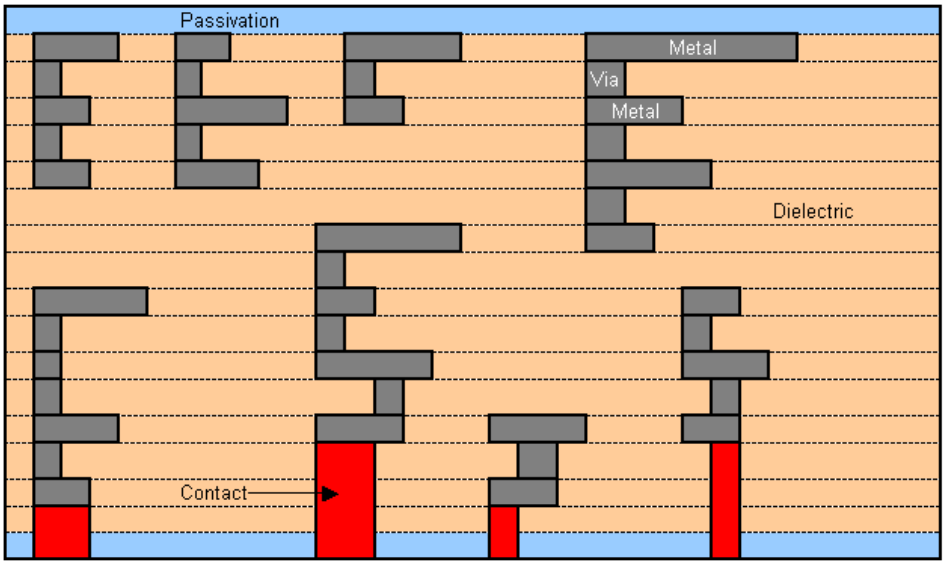

Figure 6.1. Schematic Diagram of a Multilevel Metallization (MLM) Process.

pattern density, wafer size, incoming wafer quality, and other parameters, such as wafer cleaning and temperature distribution over wafer-pad interface. A number of theories exist for the exact mechanism of material removal in a CMP process (Steigerwald *et al.* [60] and Li *et al.* [61]). Explanation of the fundamental mechanism of CMP from different viewpoints of mechanical removal, such as slurry flow (Runnels [62], Runnels and Eyman [63], and Sundarajan *et al.* [64]), slurry abrasive abrasion (Liu *et al.*[65], Shi *et al.* [66], and Chekina *et al.* [67]), and chemical removal (Runnels [62] and Sundarajan *et al.* [64]) are available in literature. Some of the material removal models that examine the input parameter interactions have been developed by Luo and Dornfeld ([68] and [69]. These models are based on solid-solid contact mode in CMP, and use the assumptions of 1) plastic contact over the wafer-abrasive interface and pad-abrasive interface, 2) a normally distributed abrasive size, and 3) a periodic rough surface of polishing pad. Also material removal rate formulations as functions of down pressure and abrasive size have been presented and verified using experimental results (Luo and Dornfeld [69], [70], [71], and [68]). In the solid-solid contact mode the wafer contacts the pad directly when the down pressure is relatively

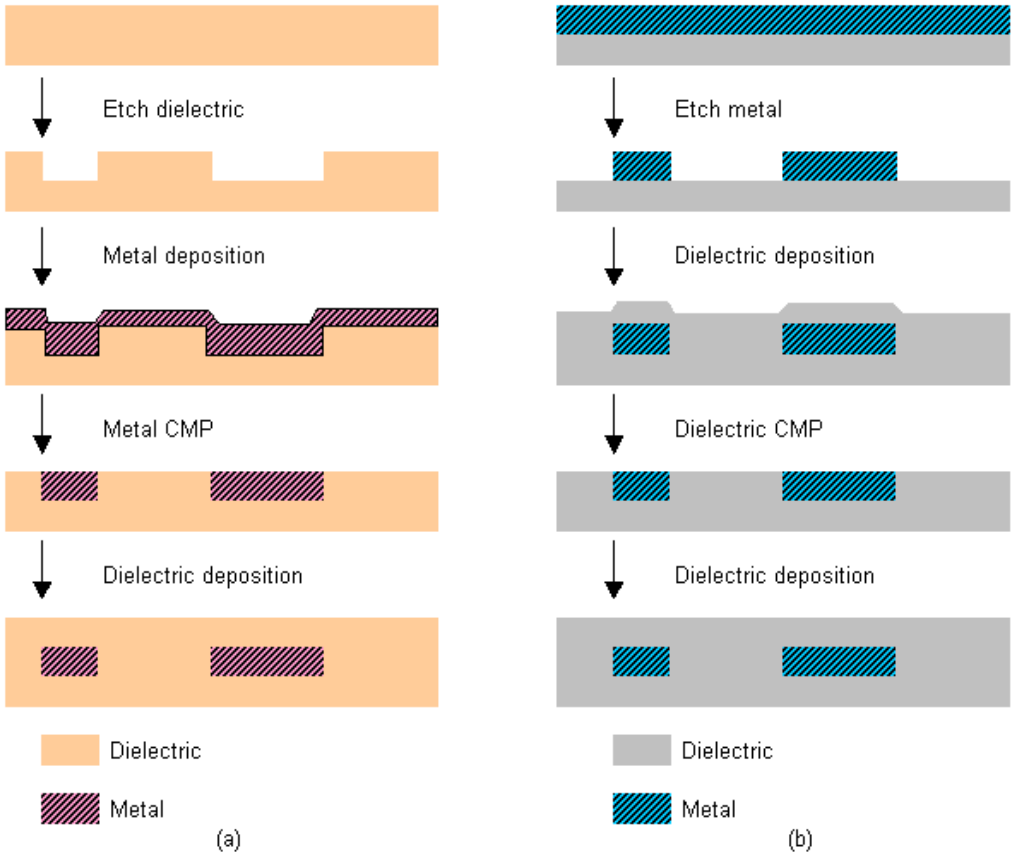

Figure 6.2. A Schematic Diagram for the Process Steps to Fabricate Copper Lines by (a) Copper Damascene Approach, and (b) Conventional Approach.

large and the relative velocity is small. This is the most common mode of CMP in which the two-body abrasion (microscratching) is dominant and fluid flow effects are insignificant. However, when the down pressure becomes small and relative velocity is large, a thin fluid film may be formed between the wafer and the pad. This is called the noncontact mode or three-body abrasion (microindentation). In this case, the mechanical effects of CMP are related to the Heresy number and conveniently referenced to the hydrodynamic bearing theory through the Stribeck curve (Moon [72] and Chang *et al.* [48]).

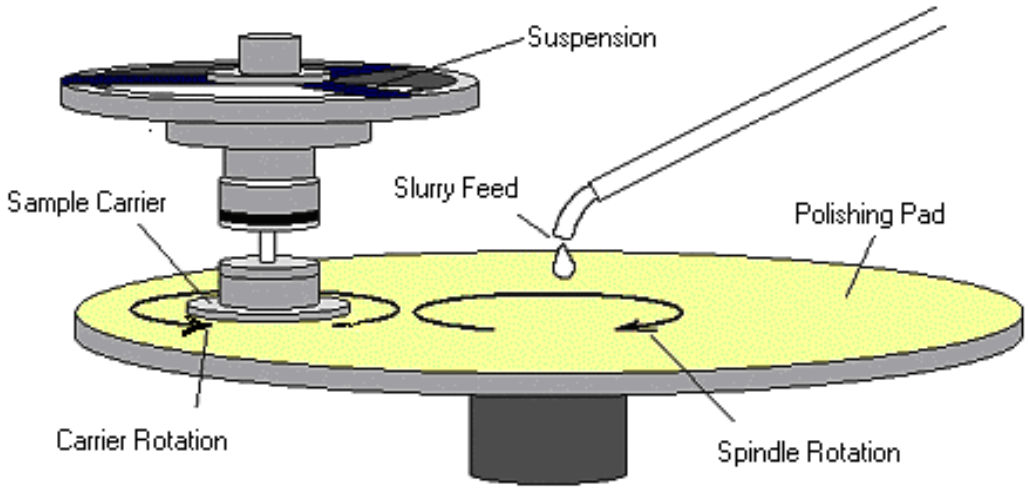

Figure 6.3. Schematic Diagram of the CMP Process.

6.3 Defects in CMP

With the advent of newer technologies in the semiconductor field, defects imparted by the CMP process are becoming more delicate requiring improved monitoring methods. For example, copper, which is fast replacing aluminum due to its higher conductivity and other advantages, is soft and has a low microhardness as compared to other conventional materials. CMP of copper surfaces using silica or alumina based slurries result in a high degree of surface scratches. Pits and craters may also be observed on the soft metal film surface. The large differences between the hardness of the metal and the slurry abrasive can further accentuate a special type of defect known as 'dishing' (Steigerwald *et al.* [60]). A schematic illustration of dishing and erosion effect is shown in Figure 6.4. The formation of a trough shaped dish is generally attributed to the low hardness of copper and the distortion of the pad. These defects are enhanced by higher polishing rates and use of higher hardness slurry particles.

Another chief defect in CMP of metals, oxides or dielectrics is over and under-polishing. Changes in material removal rate due to normal polish pad life

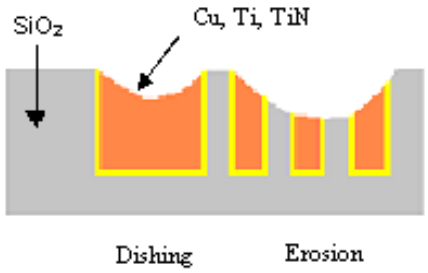

Figure 6.4. Dishing and Erosion Defects in CMP Process.

cycle, variations in slurry and pad, conditioning issues of pads, and a myriad of other potential variables can lead to this defect. Additionally incoming initial oxide or metal layer thickness may also fluctuate. Hence, end point detection (EPD) in CMP is very crucial in order to control this defect. Literature cites the need for EPD in three different processes - copper damascene, shallow trench isolation (STI), and inter-level dielectric (ILD) CMP. In copper damascene, rigorous EPD of the copper layer is required to prevent oxide erosion and copper dishing. Any erosion of oxide or dishing of copper layer would thin down the inter-connect lines, resulting in a significant effect on circuit delays owing to an increase in the RC ($resistance \times capacitance$) constant. For shallow trench isolation CMP, the process of oxide polishing must be stopped as soon as the nitride is exposed. Due to the varied pattern density distribution across the die, nitride is exposed at different times across the chip, thus resulting in different nitride over-polish. For large pattern density variations and depending on nitride thickness, the nitride may be completely removed from some areas, thus exposing the active devices. Such situations are unacceptable, and make EPD very essential in STI-CMP. Another area with a challenging EPD is the inter-level dielectric process. This is because, there is no change of material involved and CMP process must be terminated once the required thickness of the layer being polished is achieved. Additional information on EPD can be found in Hwang ([73] and [46]).

Highly planar post-polish surfaces are expected from a CMP process. However, planarity is often estimated and considered achieved if the step height of certain features is within certain limits. Since step height reduction is correlated to the amount of material removed, pre-CMP deposition thickness is first evaluated by measuring the step heights, and then material removal is monitored against time. Studies have indicated that different conclusions on planarity are achieved depending on the choice of measurement location (Fang *et al.* [45]). This suggests that step height measurement is not a very good indicator of planarity. This only provides information about local planarity, rather than the planarity within a whole die or within a whole wafer (global planarity). Hence, difference in planarity between various points on the wafer surface is another CMP defect and efforts are being made to measure and control global planarity.

Delamination of dielectric layers during CMP is another important defect, which has drawn considerable attention in recent years. Several low-k materials do not meet thermal, mechanical, and electrical requirements for integrating them with metals. Newer wafers are being designed with densely packed circuitry which requires the use of low-k dielectrics between metal interconnections. These low-k dielectrics help to reduce the capacitance, which would otherwise increase as the gaps between the metal lines diminishh due to shrinking chip size and increasing complexity. Low-k dielectrics currently being considered are generally porous in nature, which results in lower hardness, mechanical strength, cohesive force, and modulus of elasticity. They also have poor adhesion to metals in multilevel stacks. This gives rise to a common defect called delamination during the CMP process. Figure 6.5 shows three different scanning electron microscope images of polished Cu-low k wafers with and without delamination. The white portions of these figures are copper metal while the dark background represents dielectrics.

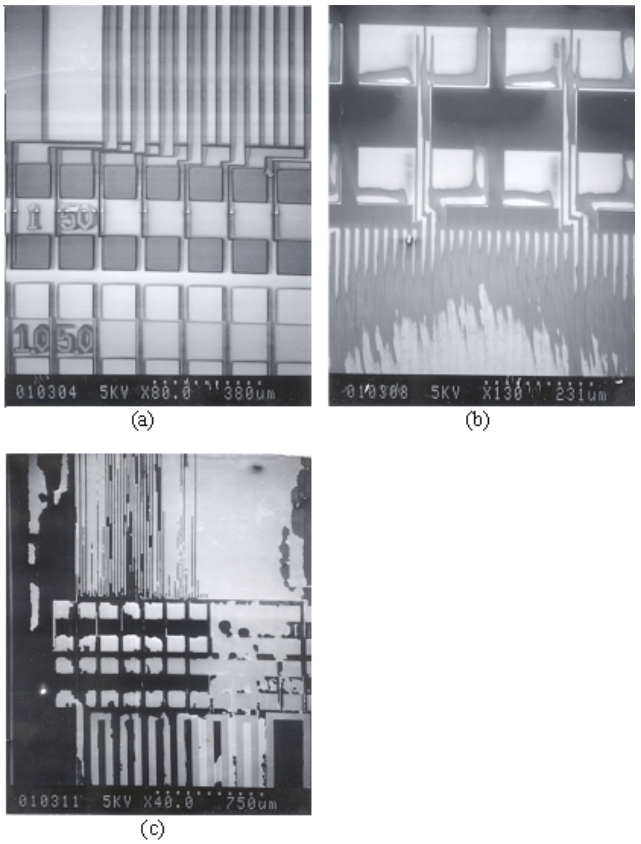

Figure 6.5. Polished Wafers from a Cu-low k CMP Process.

6.4 Application of the Strategy to Detect Delamination Defect

This section explains the application of the real-time methodology described in Chapter 6, to analyze the acoustic emission (AE) sensor data for identification and display of the occurrence of delamination defect of low-k dielectric layers in a copper damascene CMP process offline.

The CMP-Tribometer, manufactured by Center for Tribology Inc. (CETR) and located at the Nanomaterials and Nanomanufacturing Research Center (NNRC), University of South Florida (USF) was used for conducting the experimental trials. The test beds are equipped with Acoustic Emission (AE) sensors and necessary data acquisition systems. Several wafers were planarized under different combinations of rotational speed (rpm) while maintaining same downward pressure (psi), slurry composition and pad materials. The data for the process with delamination and without delamination was collected. The polished wafers were also examined using Scanning Electron Microscope (SEM) to identify between the data representing a process with delamination and without delamination . The methodology was then applied on the data sets with delamination defect to assess the efficacy of the delamination defect detection. Wavelet decomposition was done using the Daubechies 4th order wavelet. This is because the AE signal was a smooth signal. A dyadic window width of 64 was chosen which allows 6 levels of decomposition. It was observed that the initial levels consisted of high frequency noise and the underlying process was well captured at the 3^{rd}, 4^{th}, and the 5^{th} levels of decomposition. After reconstructing the wavelet coefficients to obtain the corresponding details, SPRT for variance was implemented. The design parameters α and β were fixed at 0.001 and 0.010 respectively. The design parameters σ_0 and σ_1 for every block were calculated based on the data set without delamination, which in turn were used to calculate the control limits of the SPRT chart. The data set with delamination was then used to calculate the variances of the wavelet details for each block at six levels. The variances were then

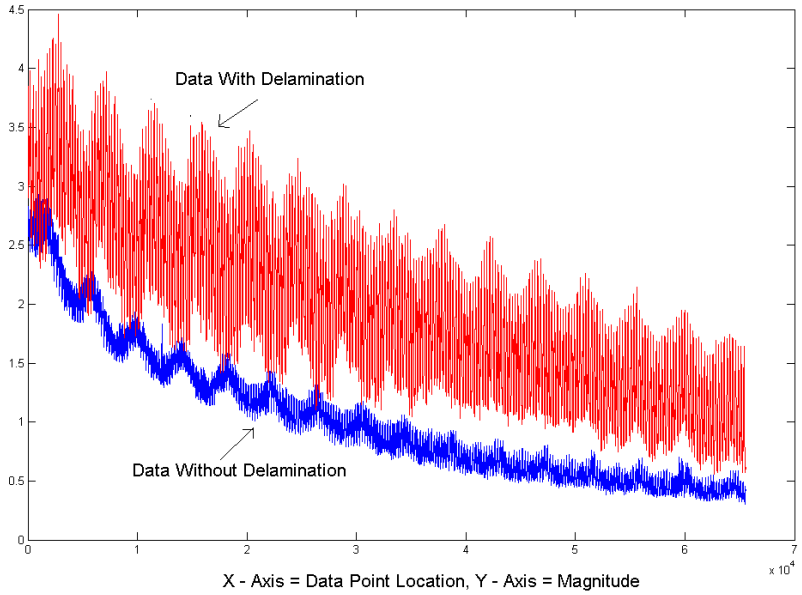

Figure 6.6. Plot of CMP Process Data (Good and Bad).

compared with the upper limit of the SPRT chart and the results were displayed. The methodology was able to exactly display the data point location and scale at which the delamination defect had occurred. The resulting displays have been shown in the next section.

6.5 Results

Figure 6.6 shows the plots of the two datasets representing the CMP process, one with delamination and the other without delamination. It was observed that the data with delamination is more dispersed than the data without delamination. The process data display also indicates its multiscale nature and necessisates the use of the real-time multiscale methodology developed in this research to analyze the process data at different scales.

Various datasets representing a CMP process with delamination were analyzed using the real-time methodology to visualy depict in the SPRT chart for variance, the exact data point location and scale at which delamination occured.

As explained in Chapter (5), real-time plots were displayed after each dyadic data block was analyzed using wavelet based multiresolution followed by the application of the Sequential Probability Ratio Test (SPRT) for variance of wavelet details. Since relevant process information was available at the 3^{rd}, 4^{th},and the 5^{th} level, SPRT for variance was applied only at these three levels.

The left hand side column of the real-time plot displays the wavelet details at each level, whereas the right hand side column displays the SPRT charts with a display of the two control limits and the variance of the wavelet details. The following figures display the multi-level plots for various process data. The occurence of delamination defect is indicated by a square box by the data point position and scale at which occurs. Since, the rate of analysis matches the rate of data acquisition, the exact time at which delamination starts can be found by adding a count of one data block to the count of data blocks already analyzed.

Figures 6.7, 6.9, and 6.11 display the start of delamination at levels 5, 3, and 4 respectively. To ensure that the detection of the start of delamination in different process datasets done by the real-time methodology is accurate, an offline plot of the variance of the wavelet details over the entire length of process data are shown only at the levels (5, 3, and 4) where the real-time implementation detected delamination. The offline plots are shown in figures 6.8, 6.10, and 6.12. It can be concluded that the starting of the delamination defect in both cases is shown at the same data point location and scale, thus ensuring the accurate detection of delamination, when the methodology is implemented real-time.

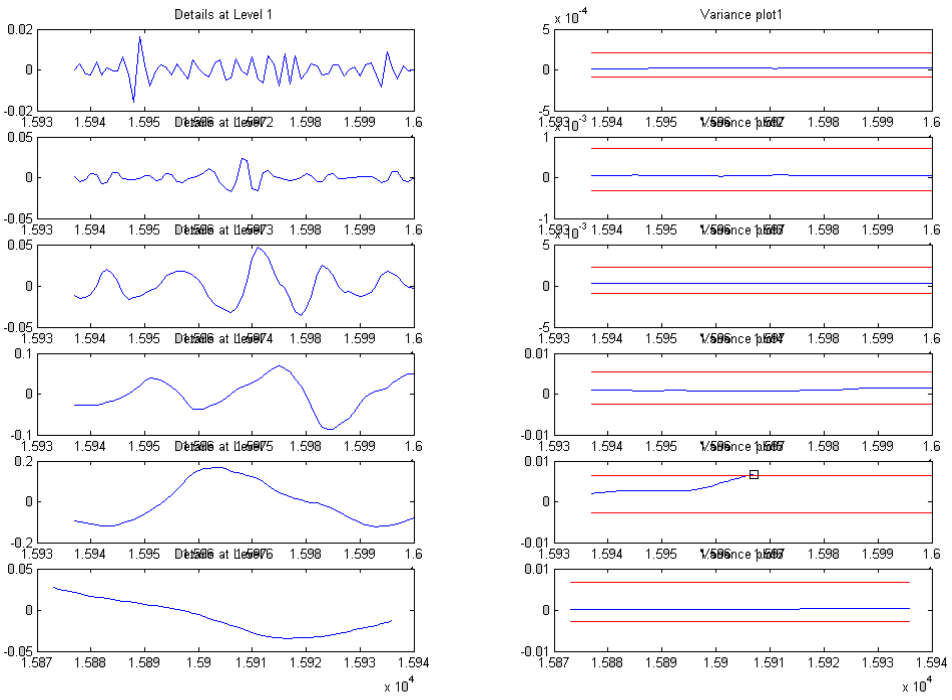

Figure 6.7. Real-Time Plot of CMP Process Dataset - I.

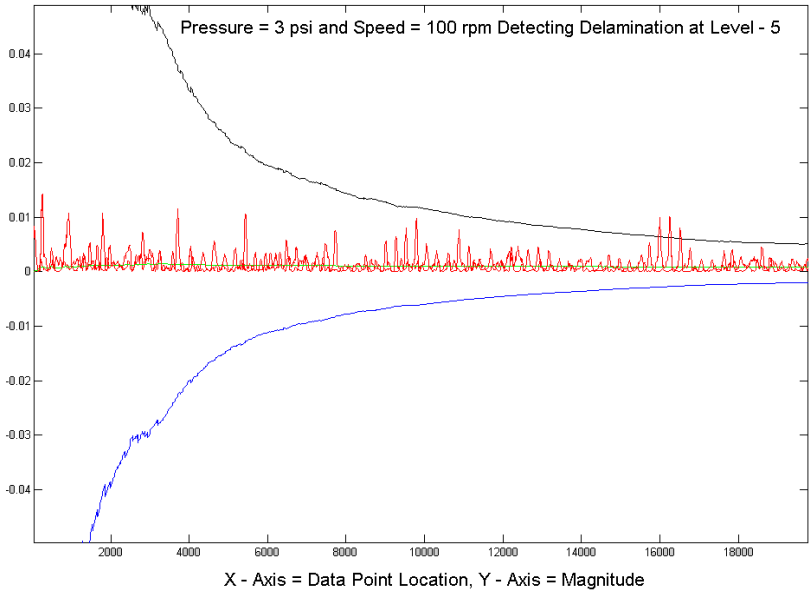

Figure 6.8. Offline Plot of CMP Process Dataset - I.

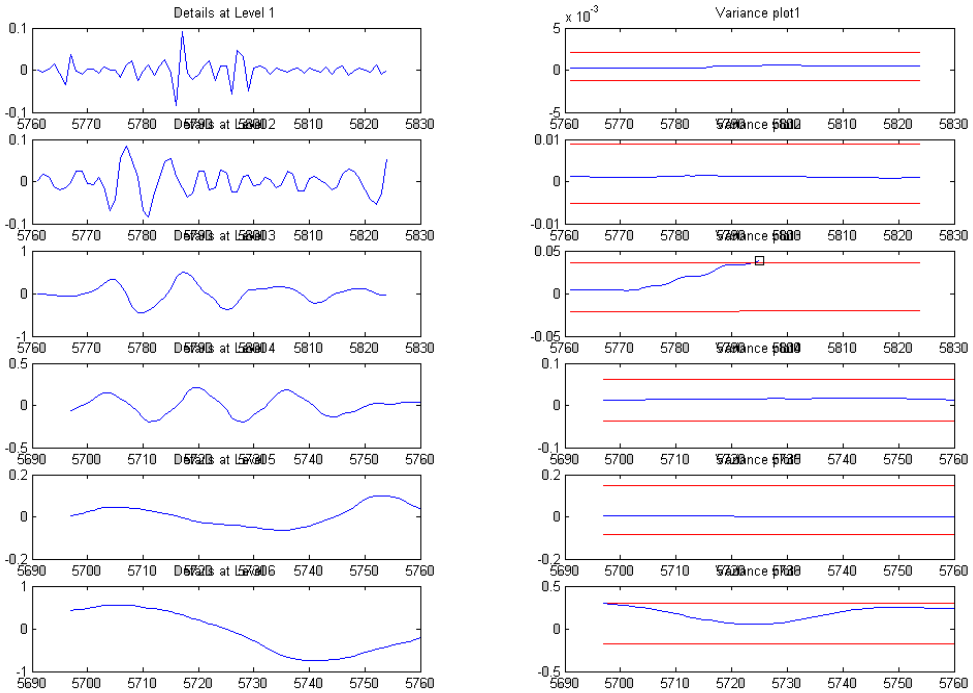

Figure 6.9. Real-Time Plot of CMP Process Dataset - II.

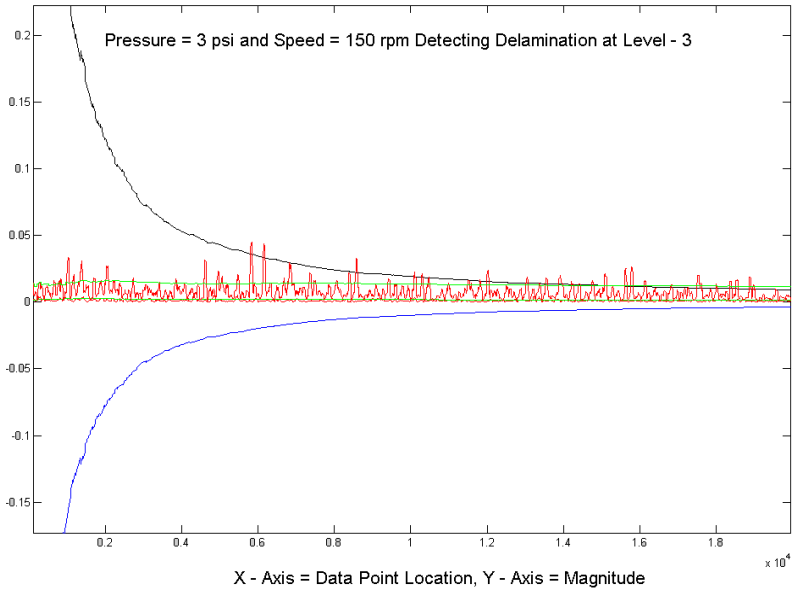

Figure 6.10. Offline Plot of CMP Process Dataset - II.

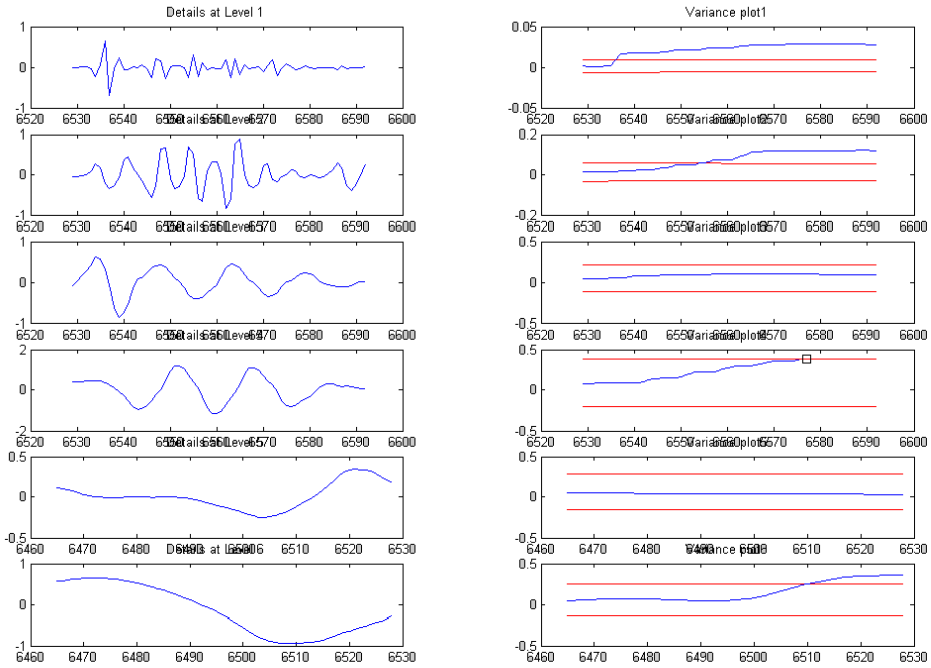

Figure 6.11. Real-Time Plot of CMP Process Dataset - III.

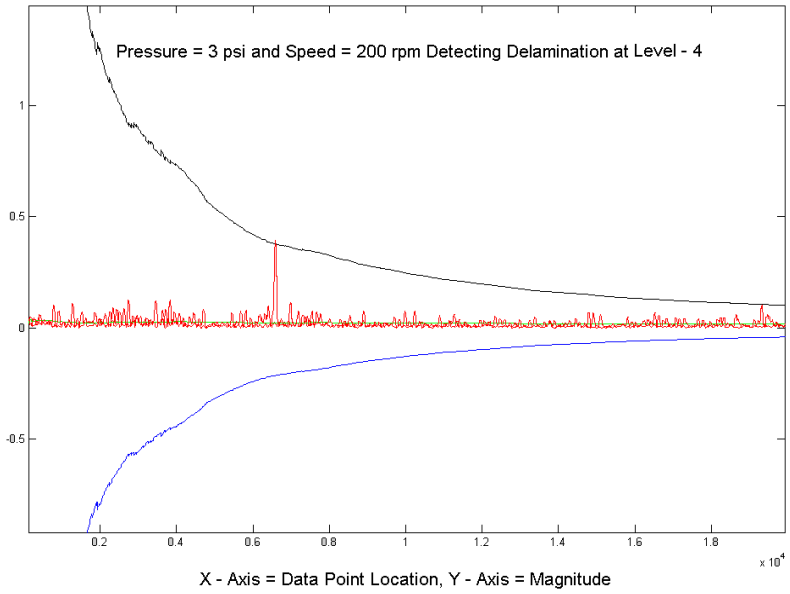

Figure 6.12. Offline Plot of CMP Process Dataset - III.

CHAPTER 7

CONCLUSIONS

7.1 Conclusions

Multiscale statistical monitoring of univariate and multivariate processes has been the focus of active research in the past few years and many promising results have appeared in the literature. Multiscale methods use several statistical tools such as wavelet decomposition, wavelet reconstruction, and charting. Almost all the existing multiscale methods are offline that do not enable real-time process monitoring to detect events of interest as and when they occur during the process. The online methodology proposed by Ganesan *et al.* [52] enables real-time monitoring but is computationally incapable of monitoring processes with high frequency of data generation. None of the existing offline or online multiscale methodologies use a statistical tool to identify the event of interest neither are they capable of displaying the process trends at all levels at the same time.

This research presents a real-time multiscale statistical monitoring methodology that uses real-time sensor data and a sequential testing procedure to detect undesirable process events. The methodology is computationally efficient and has been shown to match the rate of data generation for a CMP application. In addition to real time processing of sensor data, this methodology also displays the wavelet details and the SPRT chart for each level. This enables a visual means of observing the process condition along with the exact time and scale of an undesirable event when it occurs. The proposed methodology can be used as a useful tool for real time process monitoring and identification of undesirable events during various ultra

precision machining operations enabling significant reduction in costs by reducing the number of defectives and material wastage.

The real-time methodology was applied offline to a Chemical Mechanical Planarization (CMP) to detect delamination defect of low-k dielectric layers in a copper damascene CMP process by an extensive multiresolution study of Acoustic Emission (AE) signals. It was capable of detecting the start of the delamination in a process real-time by analyzing the data that represented a CMP process with delamination. No false alarm indicating the start of delamination was raised by the strategy. Thus the proposed real-time methodology provides an effective means to detect delamination of low-k dielectric layers during CMP processes. The length of the dyadic block is a very vital parameter for the effective operation of the proposed methodology. It is based upon the extent of decomposition required to extract relevant process related information at different levels from the input data. It has been found that as the levels of decomposition(j) increase, the block length increases by the relation 2^j, as a result increasing the computation time.

The selection of various significant design parameters of the Sequential Probability Ratio Test (SPRT) such as σ_0, σ_1, and α is extremely important. The method employed in selecting σ_0, σ_1 for every data block at all levels has been found to ensure a high sensitivity of the SPRT chart control limits. A low value of α is preferred to prevent from raising a false alarm on the occurrence of an undesirable event.

7.2 Research Extensions

In this research, the methodology developed was used to detect delamination defect during a Chemical Mechanical Planarization (CMP) process successfully. One could explore the utility of this methodology in detecting other CMP defects such as end point detection (EPD), thickness monitoring, and micro-scratches real-

time or any other manufacturing process for that matter. Though the methodology proposed in this research has outlined the steps involved in interfacing the moving block strategy with the manufacturing process, a lot of work remains to be done for its $in - situ$ implementation.

REFERENCES

[1] R. Ganesan. Wavelet based process monitoring:an application to cmp process. Master's thesis, University Of South Florida, 2002.

[2] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? (with discussion). *Journal of the Royal Statistical Society*, 57(2):301–369, 1995.

[3] A. G. Bruce and H. -Y. Gao. Waveshrink with firm shrinkage. *Statist. Sinica*, 4:855–874, 1997.

[4] B. R. Bakshi and M. N. Nounou. On-line multiscale filtering of random and gross errors without process models. *AIChE Journal*, 45(5):1041–1058, 1999.

[5] B. R. Bakshi and M. N. Nounou. Multiscale methods for denoising and compression. In B. Walczak, editor, *Wavelets in Chemistry*, volume 22 of *Data Handling in Science and Technology*, chapter 5, pages 119–150. Elsevier, P. O. Box 211, 1000 AE Amsterdam, Netherlands, 2000.

[6] B. R. Bakshi. Multiscale statistical process control and model-based denoising. In B. Walczak, editor, *Wavelets in Chemistry*, volume 22 of *Data Handling in Science and Technology*, chapter 17, pages 411–436. Elsevier, P. O. Box 211, 1000 AE Amsterdam, Netherlands, 2000.

[7] H. Hotelling. Multivariate quality control, illustrated by the air testing of sample bombsights. In C. Eisenhart, M. W. Hastay, and A. W. Wallis, editors, *Selected Techniques of Statistical Analysis for Science and Industrial Research and Production and Management Engineering*, pages 111–184, New York, 1947. McGraw-Hill Book Company.

[8] C. A. Lowry, W. H. Woodall, C. W. Champ, and S. E. Rigdon. A multivariate exponentially weighted moving average control chart. *Technometrics*, 34:46–53, 1992.

[9] S. S. Prabhu and G. C. Runger. Designing a multivariate EWMA control chart. *Journal of Quality Technology*, 29(1):8–15, 1997.

[10] Y. Meyer. *Wavelets: Algorithms and Applications*. SIAM, Philadelphia, 1993.

[11] S. G. Mallat. Multiresolution approximations and wavelet orthonormal bases. *Transactions of the American Mathematical Society*, 315(1):69–87, 1989.

[12] D. Morlet. Time-scale analysis of high-resolution signal-averaged surface ECG using wavelet transformation. *Proceedings of Computers and Cardiology*, pages 393–396, 1991.

[13] I. Daubechies. *Ten Lectures in Wavelets*. John Wiley, Philadelphia, 1992.

[14] A. Grossman and J. Morlet. Decomposition of Hardy functions into square integrable wavelet of constant shape. *SIAM J. Math. Anal.*, 15:723–736, 1984.

[15] S. G. Mallat and S. Zhong. Signal characterization from multiscale edges. Technical report 14, NYU, New York, 1991.

[16] Y. Wang. Jump and sharp detection by wavelets. *Biometrika*, 2:385–397, 1995.

[17] F. Song and S. Jutamulia. Wavelet transform and its use in edge detection. In B. Wash, editor, *Proceedings of SPIE - The International Society for Optical Engineering*, volume 4222. The Society, 2000.

[18] J. T. Y. Cheung and G. Stephanopoulous. Representation of process trends -I: A formal representation framework. *Computers Chem. Engg.*, 14:495–510, 1990.

[19] B. R. Bakshi and G. Stephanopoulos. Representation of process trends-III. Multi-scale extraction of trends from process data. *Computers Chem. Engg.*, 4:267–302, 1994.

[20] B. R. Bakshi and G. Stephanopoulos. Representation of process trends-IV. Induction of real-time patterns from operating data for diagnosis and supervisory control. *Computers Chem. Engg.*, 4:303–332, 1994.

[21] S. M. Alexander. Process monitoring, diagnosis and control. NSF Award abstract DMII #9820846, NSF, Arlington, VA 22230, 1999.

[22] N. T. Ibrahim and O. Rodriguez. Automated monitoring of microdrilling operations. *Trans. N. Am. MFG. Res. SME*, pages 205–210, 1992.

[23] N. T. Ibrahim, C. Meckdeci, O. Rodriguez, and B. Uragun. Monitoring drill conditions with wavelet based encoding and neural networks. *Intl. J. of Machine Tools and Manufacture*, 33(4):559–575, 1993.

[24] B. R. Bakshi. Multiscale analysis and modeling using wavelets. *Journal of Chemometrics*, 13:415–434, 1999.

[25] B. R. Bakshi. Multiscale PCA with application to multivariate statistical process monitoring. *AIChE Journal*, 44(7):1596–1610, 1998.

[26] R. Shao, F. Jia, E. B. Martin, and A. J. Morris. Wavelets and non-linear principal components analysis for process monitoring. *Control Engineering Practice*, 7:865–879, 1999.

[27] S. H. Fourie and P. De. Vaal. Advanced process monitoring using an on-line linear multiscale principal components analysis methodology. *Computers and Chemical Engineering*, 24:775–760, 2000.

[28] G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley Cambridge Press, Wellesley MA, 1996.

[29] H. B. Aradhye, B. R. Bakshi, R. A. Strauss, and J. F. Davis. Multiscale statistical process control using wavelets - theoretical analysis and properties. Technical report, Dept. of Chemical Engg., Ohio State University, Columbus, OH, 2001.

[30] P. Miller, R. E. Swanson, and C. E. Heckler. Contribution plots: A missing link in multivariate quality control. *Appl. Math. and Comp. Sci*, 8(4):775–792, 1998.

[31] G. Qi. Wavelet-based AE characterization of composite materials. *NDTE International*, 33:133–144, 2000.

[32] X. Li, S. Dong, and Z. Yuan. Discrete wavelet transform for tool breakage monitoring. *International Journal of Machine Tools and Manufacture*, 39:1935–1944, 1999.

[33] Q. Wang and F. Chu. Experimental determination of the rubbing location by means of acoustic emission and wavelet transform. *Journal of Sound and Vibration*, 248(1):91–103, 2001.

[34] D. Paul. Detection of change in process using wavelets. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pages 174–177, Philadelphia, PA, USA, 1994.

[35] X. Chen, J. Tang, and D. Dornfeld. Monitoring and analysis of ultra-precision metal cutting with acoustic emission. In *Proceedings of the ASME Dynamics Systems and Control Division*, volume DSC-58, pages 387–393, New York, 1996. ASME.

[36] A. Antoniadis. Wavelets in statistics: A review. *Journal of the Italian Statistical Society*, 6(2):97–144, 1999.

[37] A. Antoniadis. Smoothing noisy data with tapered coiflets series. *Scand. Journal of Statistics*, 23:313–330, 1996.

[38] A. A. Safavi, J. Chen, and J. A. Romagnoli. Wavelet-based density estimation and application to process monitoring. *AIChE Journal*, 43(5):1227–1238, 1997.

[39] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2):508–539, 1996.

[40] M. Vannucci. Non-parametric density estimation using wavelets: A review. Technical Report 95-26, ISDS, Duke University, Durham, NC.

[41] F. Abramovich, T. C. Bailey, and T. Sapatinas. Wavelet analysis and its statistical applications. *The Statistician - Journal of the Royal Statistical Society, Series D*, 49:1–29, 2000.

[42] A. Ikonomopoulos and A. Endou. Wavelet application in process monitoring. *Nuclear Technology*, 125:225–234, 1999.

[43] C. Rosen and J. A. Lennox. Multivariate and multiscale monitoring of wastewater treatment operation. *Water Resources*, 35(14):3402–3410, 2001.

[44] B. E. Stine, V. Mehrotra, D. S. Boning, J. E. Chung, and D. J. Ciplickas. A simulation methodology for assessing the impact of spatial/pattern dependent interconnect parameter variation on circuit performance. *IEDM Technical Digest*, pages 133–136, 1997.

[45] S. J. Fang, T. H. Smith, G. B. Shinn, J. A. Stefani, and D. S. Boning. Advanced process control in dielectric chemical mechanical polishing (CMP). In *Proc. of Chemical Mechanical Polish for ULSI Multilevel Interconnection Conference (CMP-MIC)*, Santa Clara, CA, USA, 1999.

[46] E. I. Hwang. Endpoint detection in CMP process: AE sensor. Technical report, Laboratory for Manufacturing Automation, University of California, Berkeley, CA, USA, 2001.

[47] A. Chang. Endpoint detection for CMP using acoustic emission. Technical report, Laboratory for Manufacturing Automation, University of California, Berkeley, CA, USA, 1999.

[48] A. Chang, J. Luo, E. I. Hwang, and D. A. Dornfeld. Process monitoring and development for chemical mechanical polishing. In *Proc. of the NSF 2001 Design, Service and Manufacturing Grantees and Research Conference*, Tampa, FL, USA, 2001. NSF.

[49] D. A. Dornfeld. Process monitoring and control for precision manufacturing. Technical report, Laboratory for Manufacturing Automation, University of California, Berkeley, CA, USA, 1999.

[50] A. Tanzawa, T. Igarashi, S. Matsuzaki, T. Suzuki, and K. Tokushige. Development of data logging system for chemical mechanical polishing and its application for process control. Technical report, 2001.

[51] D. Boning, W. Moyne, T. Smith, J. Moyne, R. Telfeyan, A. Hurwitz, S. Shellman, and J. Taylor. Run by run control of chemical mechanical polishing. *IEEE Trans. on Components, Packaging and Manufacturing Technology (C)*, 19(4):307–314, 1996.

[52] R. Ganesan, T. K. Das, A. K. Sikder, and A. Kumar. Wavelet-based identification of delamination defect in cmp (cu-low k) using nonstationary acoustic

emission signal. *IEEE Transactions on Semiconductor Manufacturing*, 16:677–685, 2003.

[53] R. Ganesan, T. K. Das, and V. Venkataraman. Wavelet based multiscale statistical process monitoring- Literature review and research extensions. *Submitted to IIE Transactions*, May 2002. Available: http://www.eng.usf.edu/~das.

[54] C. S. Burrus, R. A. Gopinath, and H. Guo. *Introduction to Wavelets and Wavelet Transform - A Primer*. Prentice-Hall, Upper Saddle River, NJ, 1998.

[55] S. G. Mallat. Multifrequency channel decompositions of images and wavelet models. *IEEE Transactions on Acoustic Speech and Signal Processing*, 37:2091–2110, 1989.

[56] A. Wald. *Sequential Analysis*. Dover, New York, 1947.

[57] P. G. Hoel. *Introduction to Mathematical Statistics*. Wiley, New York, 1984.

[58] D. Seigmund. *Sequential Analysis: Tests and Confidence Intervals*. Springer, New York, 1985.

[59] F. Pachano-Azuaje and T.K. Das. Simultaneous monitoring of mean and variance through optimally designed sprt charts. Kluwer Academic Publishers, 2001.

[60] S. P. Steigerwald, S. P. Murarka, and R. J. Gutmann. *Chemical Mechanical Planarization of microelectronic materials*. John Wiley Sons, New York, 1997.

[61] S. H. Li and R. O. Miller. *Chemical Mechanical Polishing in Silicon Processing*. Academic Press, San Diego, 2000.

[62] S. R. Runnels. Feature-scale fluid based erosion modeling for chemical mechanical polishing. *Journal of Electrochemical Society*, 141(7):1900–1904, 1994.

[63] S. R. Runnels and L. M. Eyman. Tribology analysis of chemical mechanical polishing. *Journal of Electrochemical Society*, 141(6):1689–1701, 1994.

[64] S. Sundarajan, D. G. Thakarta, D. W. Schwendeman, S. P. Muraka, and W. N. Gill. Two-dimensional wafer scale chemical mechanical planarization models based on lubrication theory and mass transport. *Journal of Electrochemical Society*, 146(2):761–766, 1999.

[65] C. -W. Liu, B. -T. Dai, W. -T. Tseng, and C. -F. Yeh. Modeling of wear mechanism during chemical mechanical polishing. *Journal of Electrochemical Society*, 143(2):776–821, 1996.

[66] F. G. Shi, B. Zhao, and S. -O. Wang. A new theory for CMP with soft pads. In *Proceedings of International Interconnect Technology Conference*, pages 73–75, San Francisco, CA, USA, 1998.

[67] O. G. Chekina, L. M. Keer, and H. Liang. Wear contact problems and modeling of chemical mechanical polishing. *Journal of Electrochemical Society*, 145(6):2100–2106, 1998.

[68] J. Luo and D. A. Dornfeld. Material removal mechanism in chemical mechanical polishing: theory and modeling. *IEEE Trans. on Semiconductor Manufacturing*, 14(2):112–133, May 2001.

[69] J. Luo. Material removal mechanism in chemical mechanical polishing. Technical report, Laboratory for Manufacturing Automation, University of California, Berkeley, CA, USA, 1999.

[70] J. Luo and D. A. Dornfeld. Effect of particle size distribution in chemical mechanical polishing: modeling and verification. Technical report, Laboratory for Manufacturing Automation, University of California, Berkeley, CA, USA, 2000.

[71] J. Luo and D. A. Dornfeld. Material removal saturation in chemical mechanical polishing with the abrasive weight concentration: effects of abrsive size and wafer-pad contact area. Technical report, Laboratory for Manufacturing Automation, University of California, Berkeley, CA, USA, 2000.

[72] Y. Moon. *Mechanical aspects of the material removal mechanism in chemical mechanical polishing CMP*. PhD thesis, University of California, Berkeley, CA, USA, 1999.

[73] E. I. Hwang. Endpoint detection in CMP process: review of current approaches. Technical report, Laboratory for Manufacturing Automation, University of California, Berkeley, CA, USA, 2001.