## DCU

# Analysis, Simulation and Design of Nonlinear RF Circuits 

## Tao Xu

A dissertation submitted in fulfilment of the requirements for the Degree of

## Doctor of Philosophy

Research Supervisor: Dr. Marissa Condon

School of Electronic Engineering<br>Dublin City University

June 2009

## Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: $\qquad$ (Candidate) ID No.: 56107625

Date: $\qquad$

This dissertation is a gift to my Dad.
I believe you are always looking at and proud of me in the other world.

## Acknowledgements

I would like to thank my Mum, Lanping Ma and wife, Miao He. Without your support, I would not have achieved this.

I want to thank my supervisor, Dr. Marissa Condon for her amazing supervision. She always gave me help whenever I needed it.

Dr. Zhipeng Ye is another important person. I am grateful to him for the helpful discussions about the delta-sigma modulators.

I would also like to thank Dr. Georgi Grahovski for his constant friendliness, help and advice.

I would like to thank Biji, John, Diana, Marie, Patrick and Conor for their help throughout the years. Also, I would like to thank the technical staff whose friendliness was so much appreciated.

## Contents

Contents ..... iv
List of Figures ..... ix
List of Tables ..... XV
1 Introduction ..... 1
2 Background to Computer Aided Design (CAD) Algorithms ..... 6
2.1 Numerical Integration Methods ..... 6
2.2 Euler Methods ..... 9
2.2.1 Forward Euler Method ..... 9
2.2.2 Backward Euler Method ..... 10
2.3 Trapezoidal Method ..... 11
2.4 Runge-Kutta Method ..... 12
2.5 Adams' Methods ..... 14
2.5.1 Adams-Bashforth Method ..... 14
2.5.2 Adams-Moulton method ..... 15
2.6 Stiff Systems and Stability Definitions ..... 15
2.7 Conclusions ..... 16
3 Padé-Based Algorithms for Numerical Integration of Ordinary Differen- tial Equations (ODE) ..... 18
3.1 Single-Step Padé-Based Algorithms ..... 19
3.1.1 Padé-Taylor Method ..... 19
3.1.2 Padé-Exponential Method ..... 20
3.1.3 Padé-Richardson Method ..... 22
3.1.4 Padé Approximation with Predictor-Corrector ..... 24
3.2 Multi-Step Padé-Based Algorithms ..... 25
3.2.1 Explicit Multi-Step Padé Approximation ..... 25
3.2.2 Implicit Multi-Step Padé Approximation ..... 28
3.3 Application and Validation ..... 28
3.3.1 Classic Test System ..... 28
3.3.2 MESFET Amplifier ..... 35
3.4 Conclusions ..... 37
4 Filon-Type Numerical Quadrature for Highly-Oscillatory Systems. ..... 38
4.1 Background to Highly Oscillatory Quadrature ..... 38
4.2 Filon-Type Numerical Approximation ..... 40
4.3 Example ..... 42
4.4 Numerical Results ..... 47
4.4.1 Amplitude Modulation ..... 48
4.4.2 Digital Modulation ..... 51
4.5 Comparison with the Envelope Simulation ..... 53
4.6 Conclusions ..... 57
5 Simulation Techniques to Capture the Locking Range of an Injection- Locked Frequency Divider ..... 58
5.1 Introduction to the Injection-Locked Frequency Divider (ILFD) ..... 58
5.2 Background to the LC-Oscillator Based ILFD ..... 59
5.2.1 Effect of Negative Resistance on LC Oscillators ..... 59
5.2.2 LC Oscillator with Cross-Coupled Negative Resistor ..... 60
5.3 Overview of the Existing Models of Injection-Locked Oscillators ..... 62
5.3.1 Adler's Model ..... 62
5.3.2 Miller's Model ..... 63
5.3.3 Verma's Model ..... 64
5.3.4 Xu's Model ..... 66
5.4 Existing Methods to Obtain the Locking Range of Oscillators ..... 73
5.4.1 Adler Approach ..... 73
5.4.2 Graphical Analysis of the Phase Characteristic ..... 77
5.4.3 Analysis with Harmonic Balance ..... 79
5.5 Novel Use of the Warped Multitime Partial Differential Equation (WaMPDE) to Determine the Locking Range ..... 83
5.5.1 The Use of the WaMPDE for Modelling the Injection-Locked Oscillator ..... 83
5.5.2 The Determination of the Limit of the Locking Range ..... 86
5.5.3 Determination of the Locking Range from Use of the WaMPDE in Conjunction with the Devil's Staircase ..... 91
5.6 A Novel Method Utilising the Multiple-Phase-Condition Envelope Fol- lowing Method (MPCENV) to Determine the Locking Range ..... 93
5.6.1 Background to the Envelope Following Method ..... 93
5.6.2 Multiple-Phase-Condition Envelope Following method ..... 94
5.7 Numerical Results and Experimental Validation ..... 97
5.7.1 Experimental Equipments Setup ..... 97
5.7.2 Numerical Results ..... 98
5.8 Conclusions ..... 101
6 Analysis and Design of Digital Delta-Sigma Modulators (DDSM) ..... 102
6.1 Background to Delta-Sigma Modulation ..... 102
6.2 Introduction of Digital Delta-Sigma Modulators ..... 106
6.3 State-of-the-art Architectures of DDSMs ..... 111
6.3.1 The Conventional DDSMs ..... 111
6.3.1.1 The Conventional MASH ..... 111
6.3.1.2 The Conventional EFM ..... 115
6.3.1.3 The Sequence Length of Conventional DDSMs ..... 117
6.3.2 HK-MASH DDSMs ..... 118
6.4 Mathematical Analysis of the MASH DDSMs employing Multi-Moduli ..... 122
6.4.1 The Suitability of the Multi-Modulus MASH-DDSM ..... 124
6.4.2 The Effect of the Multi-Moduli on the Modulator Sequence Length ..... 126
6.4.3 A Mathematical Analysis of Higher-Order MM-EFMs ..... 131
6.5 The Proposed Structure of the MASH DDSM ..... 134
6.6 The Simulation Results ..... 135
6.7 Conclusions ..... 138
7 Conclusions ..... 147
A $A$-stability and $L$-stability of Padé approximations ..... 150
B Richardson extrapolation and improvement of accuracy ..... 152
C MESFET amplifier details ..... 154
D Schematics of the diode rectifier circuit used in ADS ..... 159
E Models of DDSMs used in Simulink ..... 161
F List of relevant publications ..... 164
Bibliography ..... 166

# Analysis, Simulation and Design of Nonlinear RF Circuits 

Tao Xu


#### Abstract

The PhD project consists of two parts. The first part concerns the development of Computer Aided Design (CAD) algorithms for high-frequency circuits. Novel Padébased algorithms for numerical integration of ODEs as arise in high-frequency circuits are proposed. Both single- and multi-step methods are introduced. A large part of this section of the research is concerned with the application of Filon-type integration techniques to circuits subject to modulated signals. Such methods are tested with analog and digital modulated signals and are seen to be very effective. The results confirm that these methods are more accurate than the traditional trapezoidal rule and Runge-Kutta methods.


The second part of the research is concerned with the analysis, simulation and design of RF circuits with emphasis on injection-locked frequency dividers (ILFD) and digital delta-sigma modulators (DDSM). Both of these circuits are employed in fractional-N frequency synthesizers. Several simulation methods are proposed to capture the locking range of an ILFD, such as the Warped Multi-time Partial Differential Equation (WaMPDE) and the Multiple-Phase-Condition Envelope Following (MPCENV) methods. The MPCENV method is the more efficient and accurate simulation technique and it is recommended to obviate the need for expensive experiments. The Multi-stAge noise Shaping (MASH) digital delta-sigma modulator (DDSM) is simulated in MATLAB and analysed mathematically. A novel structure employing multimoduli, termed the MM-MASH, is proposed. The goal in this design work is to reduce the noise level in the useful frequency band of the modulator. The success of the novel structure in achieving this aim is confirmed with simulations.

## List of Figures

2.1 The integral from $t_{n}$ to $t_{n+1}$. ..... 11
3.1 The comparison of the simulation results from Adams Moulton (—) and exact analysis (**). ..... 30
3.2 The comparison of the simulation results from Runge-Kutta (-) and exact analysis (**). ..... 31
3.3 The comparison of the simulation results from Padé approximation (— ) and exact analysis (**). ..... 32
3.4 The comparison of the simulation results from Padé-based multi-step predictor-corrector method (-) and exact analysis (**). ..... 34
3.5 Schematics of the MESFET amplifier. ..... 36
3.6 The comparison of the simulation results from Adams-Moulton (—) and Padé method (**). ..... 36
4.1 Diode rectifier circuit. ..... 42
4.2 The exact solution of the amplitude-modulated equation with $I_{0}=1$, $C=1, R=1$ and $k=40$. ..... 48
4.3 The solution for the amplitude-modulated case with the trapezoidal rule. ..... 49
4.4 The solution for the amplitude-modulated equation with the third-order explicit Runge-Kutta method. ..... 50
4.5 The solution for the amplitude-modulated case using the Filon-type method, combined with waveform relaxation. ..... 50
4.6 The solution for the amplitude-modulated case with realistic values using the Filon-type method combined with waveform relaxation. ..... 51
4.7 The exact solution of the digitally modulated equation. ..... 52
4.8 A very small time sub-interval of Fig. 4.7 ..... 52
4.9 The solution of the digital-modulated equation by the trapezoidal rule. ..... 54
4.10 The solution of the digital-modulated equation by third-order explicit Runge-Kutta method. ..... 54
4.11 The solution of the digital-modulated equation by the Filon-type method, combined with waveform relaxation. ..... 55
4.12 A Modulated signal and its simulated time-varying spectrum [1]. ..... 56
5.1 Decaying impulse responses of a tank. ..... 59
5.2 Addition of negative resistance to cancel tank losses. ..... 60
5.3 Use of an active circuit to provide negative resistance. ..... 60
5.4 Cross-coupled CMOS LC Oscillator. ..... 61
5.5 Adler's model for the injection-locked oscillator. ..... 62
5.6 Miller-type model for the injection-locked oscillator ..... 64
5.7 General model for the injection-locked frequency divider. ..... 65
5.8 Circuit schematic. ..... 67
5.9 The simplified circuit. ..... 67
5.10 The schematic used in PSpice to capture the driving point characteristics. ..... 68
5.11 Driving point characteristics: Solid: Experimental; Dotted: Cubic Fit To Experiment; Dashed: PSpice. ..... 69
5.12 Model of the free-running oscillator. ..... 70
$5.13 I_{D}$ against $V_{G S}$. ..... 70
5.14 The loop enhanced model of the ILFD. ..... 71
5.15 Analog Multiplier as a phase detector. ..... 72
5.16 Vector diagram of the instantaneous voltages. ..... 74
5.17 Sample oscillator circuit [2]. ..... 74
5.18 Phase versus frequency for a simple tuned circuit. ..... 75
5.19 (a)Conceptual oscillator.(b) Frequency shift due to additional phase shift. (c) Open-loop characteristics.(d)Frequency shift by injection. ..... 77
5.20 Phasor diagram. ..... 78
5.21 Phasor diagram with maximum angle between $I_{\text {inj }}$ and $I_{T}$. ..... 79
5.22 General Circuit Model for ILFD. ..... 80
5.23 Relation between $\tau_{1}$, $t$ and $\phi_{1}(t)$. ..... 85
5.24 Steady state: Non-locked oscillator. ..... 88
5.25 Comparison of the norm of the Jacobian column when the oscillator is locked (continuous line) and non-locked (dashed line). ..... 88
5.26 Steady state: Locked oscillator. ..... 89
5.27 Local frequency and norm of Jacobian column (dashed line) vs. itera- tion number. ..... 90
5.28 Experimentally measured Devil's staircase diagram showing lockings at $\omega_{\mathrm{inj}} / \omega_{\mathrm{o}}=2$ and $\omega_{\mathrm{inj}} / \omega_{\mathrm{o}}=4$. ..... 91
5.29 Forward-Euler-based envelope-following method. ..... 93
5.30 Backward-Euler-based envelope-following method. ..... 94
5.31 Experiment setup ..... 97
5.32 The schematic of the LC-ILFD. ..... 98
5.33 The staircase obtained from simulation and experiment. ..... 99
5.34 The locking range of the LC-ILFD captured from experiments. ..... 101
6.1 An example of 4-bit PCM. ..... 103
6.2 Block diagram of the delta-modulation process. ..... 103
6.3 Derivation of Delta-Sigma Modulation from Delta Modulation. ..... 104
6.4 Block diagram of delta-sigma modulation. ..... 105
6.5 Spectrum of a signal. ..... 105
6.6 Spectrum of an oversampled signal. ..... 105
6.7 The Fractional-N Frequency Synthesizer. ..... 107
6.8 The architecture of an lth order single-loop delta-sigma modulator. ..... 108
6.9 Model of a dithered MASH DDSM. ..... 109
6.10 The power spectral density of a 9-bit MASH 1-1-1 DDSM [3]. ..... 110
6.11 MASH DDSM architecture. ..... 111
6.12 EFM1: First-order error-feedback modulator. ..... 113
6.13 The linear model of an EFM1. ..... 113
6.14 Higher-order error-feedback modulator. ..... 116
6.15 Linear model of higher-order error-feedback modulator. ..... 116
6.16 The modified EFM1 used in the HK-MASH. ..... 118
6.17 The linear model of the HK-EFM1. ..... 120
6.18 The structure of the HK-MASH. ..... 121
6.19 MM-EFM1: The modified first-order error-feedback modulator used in the MM-MASH. ..... 123
6.20 The MM-MASH consists of only MM-EFM1. ..... 125
6.21 Higher order error-feedback modulator. ..... 132
6.22 Structure of the 2nd order error-feedback modulator. ..... 132
6.23 The structure of the MM-MASH. ..... 134
6.24 The autocorrelation result for the 5-bit MM-MASH 1-1-1. ..... 136
6.25 The power spectral density of the dithered 5-bit conventional MASH DDSM and non-dithered 5-bit MM-MASH. ..... 139
6.26 The power spectral density of the dithered 6-bit conventional MASH DDSM and non-dithered 6-bit MM-MASH. ..... 139
6.27 The power spectral density of the dithered 7-bit conventional MASH DDSM and non-dithered 7-bit MM-MASH. ..... 140
6.28 The power spectral density of the dithered 8-bit conventional MASH DDSM and non-dithered 8-bit MM-MASH. ..... 140
6.29 The power spectral density of the dithered 9-bit conventional MASH DDSM and non-dithered 9-bit MM-MASH. ..... 141
6.30 The power spectral density of the dithered 10-bit conventional MASH DDSM and non-dithered 10-bit MM-MASH. ..... 141
6.31 The power spectral density of the dithered 11-bit conventional MASH DDSM and non-dithered 11-bit MM-MASH. ..... 142
6.32 The power spectral density of the 11-bit HK-MASH DDSM and 11-bit MM-MASH. ..... 142
6.33 The power spectral density of different structures of 3 rd order 5-bit MM-MASH. ..... 143
6.34 The power spectral density of different structures of 3 rd order 6 -bit MM-MASH. ..... 143
6.35 The power spectral density of different structures of 3 rd order 7-bit MM-MASH. ..... 144
6.36 The power spectral density of different structures of 3 rd order 8-bit MM-MASH. ..... 144
6.37 The power spectral density of different structures of 3 rd order 9-bit MM-MASH. ..... 145
6.38 The power spectral density of different structures of 3 rd order 10-bit MM-MASH. ..... 145

## LIST OF FIGURES

6.39 The power spectral density of different structures of 3rd order 11-bit MM-MASH. ..... 146
C. 1 Schematic of the MESFET amplifier. ..... 154
D. 1 The diode rectifier circuit with amplitude modulated input signal. ..... 159
D. 2 The diode rectifier circuit with digitally modulated input signal. ..... 160
E. 1 The model of EFM1 used in Simulink. ..... 161
E. 2 The model of HK-EFM1 used in Simulink. ..... 162
E. 3 The model of EFM2 used in Simulink. ..... 162
E. 4 The model of MASH 1-1-1 used in Simulink. ..... 163
E. 5 The model of MASH 1-2 used in Simulink. ..... 163

## List of Tables

1.1 The market value of the top 5 EDA companies. ..... 2
3.1 Total simulation time of different methods ..... 35
5.1 The value of the selected components used in both experiment and simulation. ..... 99
5.2 Locking range, where $\mathbf{M}, \mathbf{E}$ and $\mathbf{W}$ represent MPCENV, Experiment and WaMPDE ..... 100
6.1 The preferred initial conditions for the conventional DDSMs. ..... 117
6.2 The sequence length of the conventional DDSMs. ..... 117
6.3 The specifically-chosen a for various modulator word lengths. ..... 119
6.4 A comparison between the maximum sequence length of the conven- tional MASH and the HK-MASH. ..... 119
6.5 Some sample moduli of the 3 rd order MM-MASH. ..... 135
6.6 A comparison of the sequence lengths for the HK-MASH and MM-MASH. 137

## Chapter 1

## Introduction

Computer Aided Design (CAD) techniques are widely used in all branches of engineering from construction engineering to mechanical and electronic engineering. Construction and mechanical engineers draw both 2-Dimensional (2D) vector-based rendering and 3-Dimensional (3D) solid models with the aid of CAD software such as Auto CAD and Pro/E. Electronic engineers, on the other hand, use CAD software for for the design, analysis and optimisation of circuits and systems and also for drawing drafts such as Printed Circuit Boards (PCB) and circuit layouts. For electronic engineering applications, CAD software tools are termed Electronic Design Automation (EDA) tools. The earliest EDA tools were developed in Universities and are presented as open source. "Berkeley VLSI Tools Tarball" was one of the most popular software tools empoloyed by engineers to design VLSI in the 1970s [4]. In the 1980s, industry began developing EDA tools. A lot of traditional electronic companies began their research on EDA internally, such as Hewlett Packard, Tektronix and Intel. However, attracted by the bright future of EDA development, companies specifically for EDA development were also founded in this period. For example, Mentor Graphics was founded by managers from Tektronix, Daisy Systems was founded largely by designers from Intel, and Valid Logic Systems was founded by developers from Lawrence Livermore National Laboratory and Hewlett Packard [4]. Because of competition and
mergers over the the next 30 years, only a few EDA companies are still in existence and these companies occupy the whole market. For example, Cadence and Mentor Graphics share most of the market for physical IC design, Synopsys occupies the logic synthesis area and Agilent provides good high-frequency simulation software. The market value of the top five EDA businesses is shown in Table 1.1 [4]. Note that the EDA division is only part of Agilent Technologies. The market value ( $\$ 11$ billion) comes from all of its products such as EDA software (ADS), electronic measurement tools and life science applications.

Table 1.1: The market value of the top 5 EDA companies.

| Company | Location | Market Value (March 2009) |
| :---: | :---: | :---: |
| Agilent Technologies | Santa Clara, California | $\$ 11$ billion |
| Synopsys | Mountain View, California | $\$ 2550$ million |
| Cadence Design Systems | San Jose, California | $\$ 990$ million |
| Mentor Graphics | San Jose, California | $\$ 410$ million |
| Zuken Inc. | Yokohama, Japan | $\$ 149$ million |

The development and advancement of CAD tools is extremely important in today's electronic and RF industry. The complexity, diversity and level of integration of electronic circuits has grown exponentially in recent times and CAD tools need to be redeveloped for these new and multifarious applications. In this thesis, two areas of CAD development shall be addressed. The first is on the numerical integration techniques that are employed. The second is on the modelling approaches for fractional- N frequency synthesisers and the components of such synthesisers. The basic numerical integration techniques of the earliest CAD tools are totally inappropriate for the advanced modulation formats in current communication applications. Hence, the thesis shall address the development of novel approaches that are suited to modulated signals and highly stiff systems.

In Chapter 2, the background to numerical integration methods is introduced. Both of the advantages and disadvantages of these methods are described. Some important concepts are also explained in detail such as stability and stiffness.

In Chapter 3, novel Padé-based algorithms are proposed for the discrete-time integration of stiff non-linear differential equations. The basic single-step Padé-Taylor method and its form as a predictor corrector was introduced in [5]. However, in this thesis, a new condition is introduced to ensure that the method is always A- and Lstable. Furthermore, the use of a novel Padé-Exponential method and Richardson extrapolation are recommended for increasing efficiency and accuracy. Multi-step methods are also suggested. The efficacy of the methods is shown using two examples. The results are compared with those from the Adam-Moulton and Runge-Kutta techniques. The methods are suitable for application in any field of science requiring efficient and accurate numerical solution of stiff differential equations.

An effective Filon-type numerical integration scheme is introduced in Chapter 4. This novel technique is proposed for an efficient numerical solution of systems of highly oscillatory ordinary differential equations that arise in electronic systems subject to modulated signals. Olver [6] has shown that such a technique is beneficial if the system oscillates at a very high frequency. He also demonstrated that asymptotic expansion is helpful to bound the error to a small value. We extend his work and apply it to a rectifier circuit. The Filon-type method and waveform relaxation techniques are combined to solve nonlinear implicit systems of ODEs. The proposed method is compared with the traditional methods such as the trapezoidal rule and Runge-Kutta methods. This comparison shows that the proposed approach can be very effective when dealing with systems of highly oscillatory differential equations.

In many branches of science, the phenomenon of injection locking is of importance. This phenomenon occurs when the natural frequency of an oscillator changes to become identical to or an integer multiple of an external perturbing frequency. In
wireless communications, the phenomenon has been exploited for very beneficial purposes in applications such as frequency synthesis. In the feedback loop of a frequency synthesiser, a frequency pre-scaler is employed to divide the frequency by a fixed number. Injection Locked Frequency Dividers (ILFD) consume less power than static dividers [7] and hence are preferable for low-power wireless applications. Unfortunately, the bandwidth over which locking occurs for ILFDs is limited. However, they are usually employed in LC-VCO-based Phase Locked Loops (PLL) which have a limited tuning range. Consequently, the restriction on the bandwidth of the ILFDs is not an impediment to their usefulness. However, the ability to determine it to an adequate level of accuracy is an important requirement for the electronic design industry [7] [8]. The Devil's Staircase [9] was introduced as an experimental technique to measure the locking range of an ILFD. Since it requires expensive equipment and takes a long time, some analysis and simulation techniques were introduced to predict the locking range. For example, Ye introduced a method using expressions derived by Harmonic Balance analysis [8] to predict the locking range and Razavi [10] captured the locking range from phasor diagram.

In Chapter 5, several simulation techniques are proposed for the determination of the locking range of an Injection-Locked Frequency Divider (ILFD) or any general oscillator circuit. The first approach is based on the Warped Multi-time Partial Differential Equation (WaMPDE) [11]. This is a variant of the standard multi-time scale model [11]. For the determination of the locking range, three time scales are considered in the warped multi-time scale model. The first time scale is for the oscillator autonomous solution. The second time-scale is for the input signal to which the oscillator circuit synchronises when locking occurs. The third time scale is for the transient evolution of the system. The use of the warped multi-time scale model enables identification of the natural frequency of the ILFD which may then be compared with the input frequency to enable locking to be identified. The Multiple-Phase-Condition En-
velope Following (MPCENV) technique is another technique that is proposed for the prediction of the locking range. This method is based on the envelope following simulation. The simulation is repeated to determine the output frequencies of an ILFD corresponding to the increasing input frequencies. The Devil's Staircase [9], a plot of $\omega_{\text {inj }} / \omega_{\mathrm{o}}$ against $\omega_{\mathrm{inj}}$, can be created from the obtained data. Then the locking range may be measured from the staircase.

As with the ILFD, the digital delta-sigma modulator (DDSM) is also employed in the feedback of an Fractional- $N$ Frequency Synthesizers [12] [13]. However, it works as the controller of a static frequency divider in order to ensure that the division ratio is the required fractional value. In Chapter 6, the behaviour of the DDSMs shall be analysed mathematically. DDSMs have unwanted quantization noise in the desired frequency band. The goal is therefore to find a technique or a structure for DDSMs that reduces this noise. Multi-stAge noise Shaping (MASH) digital delta-sigma modulators (DDSM) shall be investigated with the goal being to obtain a structure that has the best noise performance.

It has been proved that a longer sequence length could reduce the noise effect in useful band [3]. Several methods are proposed to maximise the sequence length. Borkowski [14] obtains the maximum sequence length for the conventional DDSMs by setting the initial condition of registers. Hosseini [3] introduced a new structure to maximise the sequence length. We propose a novel architecture to further lengthen the sequence length [15] [16]. The multi-moduli technique is employed in this structure. Hence, it is termed the MM-MASH. Simulations shall confirm its superior noise performance to existing structures.

## Chapter 2

## Background to Computer Aided Design (CAD) Algorithms

### 2.1 Numerical Integration Methods

Electronic systems are modelled using ordinary differential equations (ODEs), differential algebraic equations (DAEs) or partial differential equations (PDEs). Thus the aim of circuit simulation is to find the numerical solution to these systems of equations. To start with, consider the following ODE:

$$
\begin{equation*}
y^{\prime}=\frac{d y}{d t}=f(t, y(t)) \tag{2.1}
\end{equation*}
$$

where $f(t, y(t))$ is a function of time $t$, and $y$ is a variable such as capacitor voltage or inductor current. Newton, Leibniz and Euler found that there may be more than one solution that satisfies (2.1) [17]. Thus an initial value condition is required to make the solution unique [18]:

$$
\begin{equation*}
y\left(t_{0}\right)=y_{0} . \tag{2.2}
\end{equation*}
$$

If there are $n$ variables, the system may be expressed as

$$
\begin{array}{cc}
y_{1}^{\prime}=f_{1}\left(t, y_{1}, y_{2}, \ldots, y_{n}\right), & y_{1}\left(t_{0}\right)=y_{10} \\
y_{2}^{\prime}=f_{2}\left(t, y_{1}, y_{2}, \ldots, y_{n}\right), & y_{2}\left(t_{0}\right)=y_{20}  \tag{2.3}\\
\vdots & \vdots \\
y_{n}^{\prime}=f_{n}\left(t, y_{1}, y_{2}, \ldots, y_{n}\right), & y_{n}\left(t_{0}\right)=y_{n 0}
\end{array}
$$

(2.3) is usually written as

$$
\begin{equation*}
y^{\prime}(t)=f(t, y(t)), \quad y\left(t_{0}\right)=y_{0} \tag{2.4}
\end{equation*}
$$

where $y$ is a vector-valued function and $t$ is a scaler:

$$
\begin{align*}
& y=\left(y_{1}, y_{2}, \ldots, y_{n}\right)^{\mathrm{T}}  \tag{2.5}\\
& f=\left(f_{1}, f_{2}, \ldots, f_{n}\right)^{\mathrm{T}} . \tag{2.6}
\end{align*}
$$

Hence, the PDEs for $n$ variables are transformed to a system of ODEs.
In order to find the numerical approximation to (2.4), the differential equations are solved in a step by step manner and the techniques employed are termed step-by-step methods or difference methods. With this approach, a sequence of $y(t)$ along the $t$ coordinate are produced. The time sequence is defined as:

$$
\begin{gather*}
t_{1}=t_{0}+h \\
t_{2}=t_{0}+2 h \\
\vdots  \tag{2.7}\\
t_{n}=t_{0}+n h \\
t_{n+1}= \\
t_{0}+(n+1) h .
\end{gather*}
$$

or alternatively, in the form:

$$
\begin{equation*}
t_{n+1}=t_{n}+h \tag{2.8}
\end{equation*}
$$

where $h$ is the time-step size, $n \in\{0,1, \ldots, N-1\}$ and $N$ is the number of steps in the algorithm. A small time-step size results in a bigger $N$, which decreases the simulation speed. Consequently, there has been extensive research into the development of methods that achieve a balance between accuracy and efficiency for numerically integrating ODEs or PDEs. In this Chapter, for the purposes of this explanation, a fixed step size

### 2.1 Numerical Integration Methods

shall be assumed. Then the approximate value of $y(t)$ at different time states are

$$
\begin{gather*}
y_{0}=y\left(t_{0}\right) \\
y_{1}=y\left(t_{1}\right) \\
\vdots \\
y_{n}=y\left(t_{n}\right) \\
y_{n+1}=y\left(t_{n+1}\right) . \tag{2.9}
\end{gather*}
$$

where $y(t)$ represents all the variables in (2.1), $y_{n}$ refers to the value of $y(t)$ at time $t_{n}$, and $y_{0}$ is the initial value as introduced above.

There are two types of errors introduced in the numerical calculation: round-off error and truncation error. The round-off error results from the limit of the computer representation of a number. For example, an infinite decimal fraction, $\pi$, may be approximated as 3.14. On the other hand, the truncation error is caused by the algorithm itself. For instance, the most accurate Taylor series expansion has an infinite number of terms, while the computer calculation has a finite number of terms.

The numerical methods to solve ODEs and PDEs can be divided into two groups, explicit methods and implicit methods. An explicit method is of the form:

$$
\begin{equation*}
y_{n+1}=f\left(t_{n}, y_{n}, t_{n-1}, y_{n-1}, \ldots, t_{n-k}, y_{n-k}\right) \tag{2.10}
\end{equation*}
$$

where $k \in\{0,1, \ldots, n\}$. It uses the state of the system at the current time, $t_{n}$ and $y_{n}$, and at previous time steps to define the system state at the later time $y_{n+1}$. The right hand side can be explicitly calculated. On the other hand, an implicit method involves $y_{n+1}$ in both the left and right hand sides. It is defined as:

$$
\begin{equation*}
y_{n+1}=f\left(t_{n+1}, y_{n+1}, t_{n}, y_{n}, \ldots, t_{n-k}, y_{n-k}\right) \tag{2.11}
\end{equation*}
$$

The advantage of implicit methods is that they permit a larger time-step size owing to improved stability qualities. However, their evaluation is more difficult.

Explicit methods and implicit methods can be further subdivided into single-step and multi-step methods. Single-step methods require only one previous value of the function $f()$, i.e., $y_{n}$, to calculate $y_{n+1}$. On the contrary, multi-step methods use several previous values such as $y_{n}, y_{n-1}$ and $y_{n-2}$.

The concept of stability is an important property of the numerical approximation. BIBO stability is what concerns us in this thesis. A bounded signal is any signal whose absolute value never exceeds a finite positive value, $B \in \Re$. If the output is bounded for all bounded inputs to the system, the system is termed BIBO stable system.

In this Chapter, the explicit and implicit numerical methods that are later used in this thesis for comparative purposes are briefly described. Detailed derivations and descriptions are given in numerous reference text-books e.g. [17] [18].

### 2.2 Euler Methods

### 2.2.1 Forward Euler Method

The Euler method is a first-order single-step explicit algorithm for numerically integrating the ordinary differential equations (ODEs).

Recall the ODE (2.1) introduced in Section 2.1:

$$
\begin{equation*}
y^{\prime}(t)=f(t, y(t)) \tag{2.12}
\end{equation*}
$$

where

$$
\begin{gather*}
y\left(t_{0}\right)=y_{0} \\
y\left(t_{1}\right)=y_{1}  \tag{2.13}\\
\vdots \\
y\left(t_{n}\right)=y_{n}
\end{gather*}
$$

The time sequence is defined as $t_{0}, t_{1}=t_{0}+h, t_{2}=t_{0}+2 h, \ldots, t_{n}=t_{0}+n h$ and $h$ is the time-step size. The differential $y^{\prime}(t)$ is approximated as in [19]:

$$
\begin{equation*}
y^{\prime}(t)=\frac{y(t+h)-y(t)}{h} . \tag{2.14}
\end{equation*}
$$

Substituting (2.14) into (2.12), it is obtained:

$$
\begin{equation*}
y(t+h)=y(t)+h f(t, y(t)) \tag{2.15}
\end{equation*}
$$

where $t$ is assumed as the current time $t_{n}$. Thus (2.15) can be also expressed as

$$
\begin{equation*}
y_{n+1}=y_{n}+h f\left(t_{n}, y_{n}\right) \tag{2.16}
\end{equation*}
$$

### 2.2.2 Backward Euler Method

The Backward Euler method is derived in a similar manner to the Forward Euler method. Instead of (2.14), the differential is expressed by

$$
\begin{equation*}
y^{\prime}(t)=\frac{y(t)-y(t-h)}{h} . \tag{2.17}
\end{equation*}
$$

Substitute (2.17) into (2.12), it yields

$$
\begin{equation*}
y(t)=y(t-h)+h f(t, y(t)) . \tag{2.18}
\end{equation*}
$$

It may also be expressed in the form:

$$
\begin{equation*}
y_{n}=y_{n-1}+h f\left(t, y_{n}\right) \tag{2.19}
\end{equation*}
$$

In order to compare it with the Forward Euler method, the Backward Euler method is usually expressed as:

$$
\begin{equation*}
y_{n+1}=y_{n}+h f\left(t_{n+1}, y_{n+1}\right) . \tag{2.20}
\end{equation*}
$$

It is an implicit method and therefore, it is more stable than the Forward Euler method.

### 2.3 Trapezoidal Method

Sometimes the Trapezoidal method is also called the Trapezoidal Rule. The solution of the ODE (2.12) is obtained by approximating the area under the curve using a trapezoid.

$$
\begin{equation*}
y_{n+1}=y_{n}+\int_{t_{n}}^{t_{n+1}} f(t, y) d t \tag{2.21}
\end{equation*}
$$

where

$$
\begin{align*}
y_{n} & =y\left(t_{n}\right)  \tag{2.22}\\
y_{n+1} & =y\left(t_{n+1}\right) . \tag{2.23}
\end{align*}
$$



Figure 2.1: The integral from $t_{n}$ to $t_{n+1}$.

As seen in Fig. 2.1, the area of the trapezoid is obtained by:

$$
\begin{align*}
\int_{t_{n}}^{t_{n+1}} f(t, y) d t & =\left(t_{n+1}-t_{n}\right) \frac{f\left(t_{n+1}\right)+f\left(t_{n}\right)}{2} \\
& =\frac{h}{2}\left(f\left(t_{n+1}\right)+f\left(t_{n}\right)\right) \tag{2.24}
\end{align*}
$$

where $h=t_{n+1}-t_{n}$ is the time-step size. Hence, the approximation of $y(t)$ is obtained as:

$$
\begin{equation*}
y_{n+1}=y_{n}+\frac{h}{2}\left(f\left(t_{n+1}\right)+f\left(t_{n}\right)\right) . \tag{2.25}
\end{equation*}
$$

Again, the Trapezoidal Rule is an implicit method.

### 2.4 Runge-Kutta Method

The Runge-Kutta method was developed around 1900 by the German mathematicians C. Runge and M.W. Kutta [20]. The basic idea of the Runge-Kutta method is that if the slopes at several points between $t_{n}$ and $t_{n+1}$ are found, the average value of them is assumed as the slope at $t_{n}$.

$$
\begin{equation*}
y_{n+1}=y_{n}+h \cdot(\text { average_slope }) . \tag{2.26}
\end{equation*}
$$

Mathematically, (2.26) is expressed as:

$$
\begin{equation*}
y_{n+1}=y_{n}+h \sum_{i=1}^{s} b_{i} k_{i} \tag{2.27}
\end{equation*}
$$

where $k_{i}$ represent the slopes at the points in the interval between $y_{n}$ and $y_{n+c_{s}}$. They are defined as:

$$
\begin{align*}
k_{1} & =f\left(t_{n}, y_{n}\right)  \tag{2.28}\\
k_{2} & =f\left(t_{n}+c_{2} h, y_{n}+a_{21} h k_{1}\right)  \tag{2.29}\\
k_{3} & =f\left(t_{n}+c_{3} h, y_{n}+a_{31} h k_{1}+a_{32} h k_{2}\right)  \tag{2.30}\\
& \vdots  \tag{2.31}\\
k_{s} & =f\left(t_{n}+c_{s} h, y_{n}+a_{s 1} h k_{1}+a_{s 2} h k_{2}+\cdots+a_{s, s-1} h k_{s-1}\right)
\end{align*}
$$

There are a lot of choices for the number of stages $s$ and the coefficients $a_{i j}, b_{p}$ and $c_{q}$, where $1 \leq j<i \leq s, p \in\{1,2, \ldots, s\}$ and $q \in\{2,3, \ldots, s\}$. To obtain specific values for the parameters, $y_{n+1}$ is expanded in powers of $h$ such that it agrees with the Taylor
series expansion of the solution of the differential equation to a specified number of terms.

The Forward Euler method is the simplest one-stage Runge-Kutta with the corresponding coefficients:

$$
\begin{align*}
s & =1  \tag{2.32}\\
b_{1} & =1 \tag{2.33}
\end{align*}
$$

Then the state $y_{n+1}$ is obtained as:

$$
\begin{equation*}
y_{n+1}=y_{n}+h k_{1} . \tag{2.34}
\end{equation*}
$$

The classical fourth-order Runge-Kutta method (RK4) is the most popular one. It is:

$$
\begin{equation*}
y_{n+1}=y_{n}+\frac{h}{6}\left(k_{1}+2 k_{2}+2 k_{3}+k_{4}\right) \tag{2.35}
\end{equation*}
$$

where

$$
\begin{align*}
k_{1} & =f\left(t_{n}, y_{n}\right)  \tag{2.36}\\
k_{2} & =f\left(t_{n}+\frac{h}{2}, y_{n}+\frac{h}{2} k_{1}\right)  \tag{2.37}\\
k_{3} & =f\left(t_{n}+\frac{h}{2}, y_{n}+\frac{h}{2} k_{2}\right)  \tag{2.38}\\
k_{4} & =f\left(t_{n}+h, y_{n}+h k_{3}\right) \tag{2.39}
\end{align*}
$$

$k_{1}$ is the slope at the beginning of the interval;
$k_{2}$ is the slope at the midpoint of the interval. The slope $k_{1}$ is used to determine the value of $y$ at the point $t_{n}+\frac{h}{2}$ using Euler's method;
$k_{3}$ is again the slope at the midpoint, but now the slope $k_{2}$ is used to determine the $y$-value;
$k_{4}$ is the slope at the end of the interval, with its $y$-value determined using $k_{3}$. Then, the average slope is determined as:

$$
\begin{equation*}
\text { slope }=\frac{k_{1}+2 k_{2}+2 k_{3}+k_{4}}{6} \tag{2.40}
\end{equation*}
$$

Thus, $y_{n+1}$ is determined from the previous state $y_{n}$, the time-step $h$ and an estimated slope.

### 2.5 Adams' Methods

### 2.5.1 Adams-Bashforth Method

The Adams-Bashforth method [21] is an explicit multi-step method. The major advantage of the Adams-Bashforth method over the Runge-Kutta methods is that only one unknown function needs to be evaluated for each step. All the other functions such as $f\left(t_{k-1}, y_{k-1}\right)$ and $y_{k-1}$ have been obtained from the previous time-step.

The solution to (2.12) may be written as

$$
\begin{equation*}
y\left(t_{n+1}\right)=y\left(t_{n}\right)+\int_{t_{n}}^{t_{n+1}} f(t, y) d t \tag{2.41}
\end{equation*}
$$

The function $f$ is replaced by a Lagrange interpolation polynomial [22] at $k$ points $t_{n}, t_{n-1}, \ldots, t_{n-k+1}:$

$$
\begin{align*}
f_{\text {poly }}(t)= & f_{n}+\left(t-t_{n}\right) f\left[t_{n}, t_{n-1}\right] \\
& +\left(t-t_{n}\right)\left(t-t_{n-1}\right) f\left[t_{n}, t_{n-1}, t_{n-2}\right]+\cdots \\
& +\left(t-t_{n}\right)\left(t-t_{n-1}\right) \cdots\left(t-t_{n-k+2}\right) f\left[t_{n}, t_{n-1}, \ldots, t_{n-k+1}\right] \tag{2.42}
\end{align*}
$$

where

$$
\begin{equation*}
f\left[t_{n}, t_{n-1}, \ldots, t_{n-k+1}\right]=\frac{1}{t-t_{n-k+1}}\left(f\left[t_{n}, t_{n-1}, \ldots, t_{n-k+2}\right]-f\left[t_{n-1}, \ldots, t_{n-k+1}\right]\right) . \tag{2.43}
\end{equation*}
$$

As an example, the derivation of the 2nd-order Adams-Bashforth method, i.e., $k=$ 2, is shown as below. First, the second-order Lagrangian polynomial is obtained

$$
\begin{equation*}
f_{\text {poly }}(t)=-\frac{1}{h}\left(t-t_{i}\right) f\left(t_{i-1}, y\left(t_{i-1}\right)\right)+\frac{1}{h}\left(t-t_{i-1}\right) f\left(t_{i}, y\left(t_{i}\right)\right) . \tag{2.44}
\end{equation*}
$$

Substituting (2.44) into (2.41) and integrating yields
$y\left(t_{i+1}\right)=y\left(t_{i}\right)+\left[-\frac{1}{h}\left(t^{2} / 2-t_{i} t\right) f\left(t_{i-1}, y\left(t_{i-1}\right)\right)+\frac{1}{h}\left(t^{2} / 2-t_{i-1} t\right) f\left(t_{i}, y\left(t_{t}\right)\right)\right]_{t_{i}}^{t_{i+1}}$.

Then the final expression for the 2nd-order Adams-Bashforth method is obtained by inserting the limits.

$$
\begin{equation*}
y_{i+1}=y_{i}+\frac{h}{2}\left[3 f\left(t_{i}, y_{i}\right)-1 f\left(t_{i-1}, y_{i-1}\right)\right] . \tag{2.46}
\end{equation*}
$$

### 2.5.2 Adams-Moulton method

The Adams-Moulton method is an implicit multi-step method, which follows from the Adams-Bashforth method. The key difference between them is that with the AdamsMoulton method, the function $f(t, y)$ is replaced by the polynomial at the $(k+1)$ points $t_{n+1}, t_{n}, \ldots, t_{n-k+1}$. Then the polynomial function is

$$
\begin{align*}
f_{\text {poly }}(t)= & f_{n+1}+\left(t-t_{n+1}\right) f\left[t_{n+1}, t_{n}\right] \\
& +\left(t-t_{n+1}\right)\left(t-t_{n}\right) f\left[t_{n+1}, t_{n}, t_{n-1}\right]+\cdots \\
& +\left(t-t_{n+1}\right)\left(t-t_{n}\right) \cdots\left(t-t_{n-k+2}\right) f\left[t_{n+1}, t_{n}, \ldots, t_{n-k+1}\right] \tag{2.47}
\end{align*}
$$

where

$$
\begin{equation*}
f\left[t_{n+1}, t_{n}, \ldots, t_{n-k+1}\right]=\frac{1}{t_{n+1}-t_{n-k+1}}\left(f\left[t_{n+1}, t_{n}, \ldots, t_{n-k+2}\right]-f\left[t_{n}, \ldots, t_{n-k+1}\right]\right) \tag{2.48}
\end{equation*}
$$

The value of $f\left(t_{n+1}, y_{n+1}\right)$ may be predicted using the Adams-Bashforth method before the evaluation of (2.47). This combined procedure is termed a predictor-corrector method.

### 2.6 Stiff Systems and Stability Definitions

A system of differential equations is said to be stiff when it is the interval of absolute stability rather than a requirement to keep the truncation error small that constrains the
step-size. It generally arises when components of the solution of the differential equation system undergo changes of a similar magnitude over very different time intervals. Consequently, any ODE solver in an RF simulation package must be able to handle stiff systems in order to avoid the numerically instability and inefficiency.

In order to analyse the stability qualities of a numerical method, the methods are applied to the test equation:

$$
\begin{equation*}
y^{\prime}=\lambda y \tag{2.49}
\end{equation*}
$$

where $\lambda$ is a constant. The solution is

$$
\begin{equation*}
y_{n+1}=\phi(\lambda) y_{n} \tag{2.50}
\end{equation*}
$$

where $\phi$ is a function of $\lambda$. The numerical method is said to be A-stable, if the solution is stable for all $\lambda<0$, i.e., the solution satisfies (??) for $\forall \lambda<0$ [23]. A numerical method is L-stable if it satisfies the two conditions given below [24]:

1. It is $A$-stable.
2. It satisfies $\lim _{\lambda \rightarrow-\infty} \phi=0$.

### 2.7 Conclusions

Since a small time-step size results in expensive and time-consuming computations, many CAD techniques have been developed to achieve a balance between stability and efficiency. The majority of circuit simulators such as SPICE [25] use implicit numerical integration techniques with adaptive time stepping. These methods have the advantages of good stability and accuracy control properties, but a solution to a set of non-linear algebraic equations is required at each time-step and this can be computationally expensive. An alternate approach is to employ predictor-corrector methods. The traditional predictor-corrector methods such as the Adams Moulton method employ polynomial extrapolation to provide an estimate of the solution at the next time
step. The estimate is subsequently corrected using one or several iterations of an implicit formula. However, for stiff problems, the maximum allowable time-step that can be used with the traditional predictor-corrector techniques may be unacceptably small. Thus, novel numerical methods utilising the Padé approximation will be introduced in Chapter 3. With these methods, a larger time-step size $h$ is permitted without deceasing the accuracy.

## Chapter 3

## Padé-Based Algorithms for Numerical Integration of Ordinary Differential Equations (ODE)

As mentioned in the previous chapter, there is an on-going need for improved numerical integration algorithms for solving stiff differential systems. To this end, several novel ODE solvers are proposed in this Chapter. In particular, several Padé-based algorithms [5] [26] [27] are proposed.

Recall the ODE system:

$$
\begin{equation*}
\frac{d y}{d t}=f(t, y(t)) \tag{3.1}
\end{equation*}
$$

where

$$
\begin{equation*}
y_{0}=y\left(t_{0}\right) . \tag{3.2}
\end{equation*}
$$

In order to solve (3.1), a sequence of Padé approximations to $y_{i}(t)$ are provided:

$$
\begin{equation*}
y_{i}(t)=\frac{\sum_{j=0}^{m} a_{j} h^{j}}{\sum_{k=0}^{n} b_{k} h^{k}}, m<n \tag{3.3}
\end{equation*}
$$

where $b_{0}=1$ and $y_{i}(t)$ is the $i$ th element of the $y(t)$ vector. The order of this system is defined as $m+n$.

The Padé approximation (3.3) can be represented in the form of an $m+n$th order polynomial:

$$
\begin{equation*}
y_{i}(t)=c_{0}+c_{1} h+c_{2} h^{2}+\cdots+c_{m+n} h^{m+n} . \tag{3.4}
\end{equation*}
$$

Equating (3.3) and (3.4),

$$
\begin{equation*}
\sum_{i=0}^{m+n} c_{i} h^{i}=\frac{\sum_{j=0}^{m} a_{j} h^{j}}{\sum_{k=0}^{n} b_{k} h^{k}} \tag{3.5}
\end{equation*}
$$

where $i \in\{0,1, \ldots, m+n\}, j \in\{0,1, \ldots, m\}$ and $k \in\{0,1, \ldots, n\}$. The coefficients $a_{j}$ and $b_{k}$ are obtained by equating the coefficients of $h$ in (3.5). Once the coefficients $a_{j}$ and $b_{k}$ are obtained, the value for $y_{i}$ at the next time step is obtained from (3.3).

The Padé-based algorithm is advanced in time by using the solution at time $t$ as the initial condition for the next time-step. The use of Padé functions of the form of (3.3) with $m<n$ results in an $A$-stable and $L$-stable method as shown in the Appendix A. The central element of the Padé algorithms is the determination of $c_{i}$. This shall be discussed in what follows.

### 3.1 Single-Step Padé-Based Algorithms

### 3.1.1 Padé-Taylor Method

To determine the coefficients $c_{i}$, consider a Taylor series for $y(t)$. If (3.1) has an unique solution $y(t)$, the $p$ th order Taylor series of $y(t)$ about $t_{n}$ is

$$
\begin{equation*}
y(t)=y\left(t_{n}\right)+\left(t-t_{n}\right) y^{\prime}\left(t_{n}\right)+\frac{1}{2!}\left(t-t_{n}\right)^{2} y^{\prime \prime}\left(t_{n}\right)+\cdots+\frac{1}{p!}\left(t-t_{n}\right)^{p} y^{(p)}\left(t_{n}\right) \tag{3.6}
\end{equation*}
$$

where $t$ is in the required interval $[0, T]$. By assuming that $t=t_{n+1}$ and $h=t_{n+1}-t_{n}$, (3.6) is then in the form of

$$
\begin{equation*}
y(t)=y\left(t_{n}\right)+h y^{\prime}\left(t_{n}\right)+\frac{h^{2}}{2!} y^{\prime \prime}\left(t_{n}\right)+\cdots+\frac{h^{p}}{p!} y^{(p)}\left(t_{n}\right) \tag{3.7}
\end{equation*}
$$

Substituting (3.1) into (3.7) yields

$$
\begin{equation*}
y_{n+1}=y_{n}+h f\left(t_{n}, y_{n}\right)+\frac{h^{2}}{2!} f^{\prime}\left(t_{n}, y_{n}\right)+\cdots+\frac{h^{p}}{p!} f^{(p-1)}\left(t_{n}, y_{n}\right) \tag{3.8}
\end{equation*}
$$

By comparing (3.8) and (3.4), the coefficients $c_{i}$ are obtained as

$$
\begin{align*}
& c_{0}= y_{n} \\
& c_{1}= f\left(t_{n}, y_{n}\right) \\
& c_{2}= \frac{f^{\prime}\left(t_{n}, y_{n}\right)}{2!}  \tag{3.9}\\
& \vdots \\
& c_{m+n}= \frac{f^{(p-1)}\left(t_{n}, y_{n}\right)}{p!}
\end{align*}
$$

Substituting (3.9) into (3.5), the coefficients $a_{j}$ and $b_{k}$ are found.

### 3.1.2 Padé-Exponential Method

In this method, the solution of (3.1) is formed as the composition of a Padé approximation and an exponential function as below:

$$
\begin{equation*}
y_{i}\left(t_{k+1}\right)=\frac{\sum_{i=0}^{m} a_{i} h^{i}}{\sum_{j=0}^{n} b_{j} h^{j}}+d_{1} e^{d_{2} h} \tag{3.10}
\end{equation*}
$$

where $b_{0}=1, h=t_{k+1}-t_{k}, m<n$ and $d_{2}<0$. The rationale for this choice is that the exponential term extracts the fast-varying part of the solution and that a lowerorder Padé expression suffices for the slow-varying part of the solution. The method is $A$-stable and gives an exact solution for the test function $e^{-\lambda t}, \lambda>0$, if $m<n$ and $d_{2}<0$. Let $y_{i}\left(t_{k+1}\right)$ be approximated as the sum of a polynomial and an exponential function

$$
\begin{equation*}
y_{i}\left(t_{k+1}\right)=P_{k}\left(t_{k+1}\right)+d_{1} e^{d_{2} h} . \tag{3.11}
\end{equation*}
$$

where

$$
\begin{equation*}
P_{k}\left(t_{k+1}\right)=c_{0}+c_{1} h+c_{2} h^{2}+\cdots+c_{L} h^{k} \tag{3.12}
\end{equation*}
$$

Substituting (3.12) into (3.11) yields

$$
\begin{align*}
y_{i}\left(t_{k+1}\right) & =P E\left(t_{k+1}\right) \\
& =c_{0}+c_{1} h+c_{2} h^{2}+\cdots+c_{L} h^{L}+d_{1} e^{d_{2} h} \tag{3.13}
\end{align*}
$$

To determine the coefficients, a system of $L+3$ algebraic equations is formed from $y(t)$ and its derivatives $f^{(r)}(t)$, which is shown as below

$$
\begin{align*}
y_{i}(t) & =P E(t) \\
f_{i}(t) & =P E^{\prime}(t) \\
f_{i}^{\prime}(t) & =P E^{\prime \prime}(t)  \tag{3.14}\\
\vdots & \\
f_{i}^{(L+1)}(t) & =P E^{(L+2)}(t)
\end{align*}
$$

From (3.14), the $L+3$ unknowns in (3.13) are obtained, i.e., $c_{0}, c_{1}, \ldots, c_{L}$ and $d_{1}, d_{2}$. From these, the coefficients $a_{i}$ and $b_{j}$ in (3.10) can be found from (3.5).

As an example, the first-order Padé-Exponential method is described here, i.e., $m=0, n=1$ and $L=m+n=1$. Then (3.10) and (3.13) become

$$
\begin{align*}
& y_{i}\left(t_{k+1}\right)=\frac{a_{0}}{1+b_{1} h}+d_{1} e^{d_{2} h}  \tag{3.15}\\
& y_{i}\left(t_{k+1}\right)=c_{0}+c_{1} h+d_{1} e^{d_{2} h} \tag{3.16}
\end{align*}
$$

Thus by solving the equation

$$
\begin{equation*}
\frac{a_{0}}{1+b_{1} h}=c_{0}+c_{1} h \tag{3.17}
\end{equation*}
$$

it is obtained

$$
\begin{align*}
& a_{0}=c_{0}  \tag{3.18}\\
& b_{1}=-\frac{c_{1}}{c_{0}} . \tag{3.19}
\end{align*}
$$

Now (3.14) is in the form

$$
\begin{align*}
& y_{i}\left(t_{k+1}\right)=c_{0}+c_{1} h+d_{1} e^{d_{2} h} \\
& f_{i}\left(t_{k+1}\right)=c_{1}+d_{1} d_{2} e^{d_{2} h} \\
& f_{i}^{\prime}\left(t_{k+1}\right)=d_{1} d_{2}^{2} e^{d_{2} h}  \tag{3.20}\\
& f_{i}^{\prime \prime}\left(t_{k+1}\right)=d_{1} d_{2}^{3} e^{d_{2} h} .
\end{align*}
$$

Hence, the solution of (3.18), (3.19) and (3.20) at $h=0$ is

$$
\begin{align*}
d_{2} & =\frac{f_{i}^{\prime \prime}\left(t_{k+1}\right)}{f_{i}^{\prime}(t)}  \tag{3.21}\\
d_{1} & =\frac{f_{i}^{\prime}\left(t_{k+1}\right)}{d_{2}^{2}}  \tag{3.22}\\
a_{0} & =y_{i}\left(t_{k+1}\right)-d_{1}  \tag{3.23}\\
b_{1} & =-\frac{f_{i}\left(t_{k+1}\right)-d_{1} d_{2}}{y_{i}\left(t_{k+1}\right)-d_{1}} \tag{3.24}
\end{align*}
$$

Finally, the the value for $y_{i+1}$ is obtained by substituting (3.21)-(3.24) into (3.15).

### 3.1.3 Padé-Richardson Method

In the Padé-Richardson method, one of the methods introduced above, i.e., the PadéTaylor method or the Padé-Exponential method, is combined with Richardson extrapolation [28]. With this method, the accuracy is improved without having to evaluate higher-order derivatives of $f$ in equation (3.1).

Let the initial approximation of the solution of (3.1) be given by $y\left(t_{k+1}\right)$.

$$
\begin{equation*}
y_{e x}\left(t_{k+1}\right)=y\left(t_{k+1}\right)+O(h) \tag{3.25}
\end{equation*}
$$

If the error $O(h)$ is approximated by a polynomial of order $L$

$$
\begin{equation*}
O(h)=c h^{L}+O\left(h^{L+1}\right) \tag{3.26}
\end{equation*}
$$

(3.25) becomes

$$
\begin{equation*}
y_{e x}\left(t_{k+1}\right)=y\left(t_{k+1}\right)+c h^{L}+O\left(h^{L+1}\right) \tag{3.27}
\end{equation*}
$$

By evaluating (3.27) with two different time-steps, $h$ and $h / 2$, the approximation of $y\left(t_{k+1}\right)$ is expressed by two equations:

$$
\begin{align*}
& y_{e x}\left(t_{k+1}\right)=y_{h}\left(t_{k+1}\right)+c h^{L}+O\left(h^{L+1}\right)  \tag{3.28}\\
& y_{e x}\left(t_{k+1}\right)=y_{h / 2}\left(t_{k+1}\right)+c\left(\frac{h}{2}\right)^{L}+O\left(h^{L+1}\right) . \tag{3.29}
\end{align*}
$$

Multiplying (3.29) by $2^{L}$ gives

$$
\begin{equation*}
2^{L} y_{e x}\left(t_{k+1}\right)=2^{L} y_{h / 2}\left(t_{k+1}\right)+c h^{L}+O\left(h^{L+1}\right) \tag{3.30}
\end{equation*}
$$

Subtracting (3.28) from (3.30) , it is obtained

$$
\begin{equation*}
\left(2^{L}-1\right) y_{e x}\left(t_{k+1}\right)=2^{L} y_{h / 2}\left(t_{k+1}\right)-y_{h}\left(t_{k+1}\right) . \tag{3.31}
\end{equation*}
$$

Hence,

$$
\begin{equation*}
y_{e x}\left(t_{k+1}\right)=\frac{2^{L} y_{h / 2}\left(t_{k+1}\right)-y_{h}\left(t_{k+1}\right)}{2^{L}-1} \tag{3.32}
\end{equation*}
$$

Suppose that the two approximations, $y_{h}\left(t_{k+1}\right)$ and $y_{h / 2}\left(t_{k+1}\right)$ are obtained with a Padé method of order $L$, the Padé approximation for the solution at $t_{k+1}$ can be improved by forming:

$$
\begin{equation*}
\bar{y}_{P}\left(t_{k+1}\right)=\frac{2^{L} y_{h / 2}\left(t_{k+1}\right)-y_{h}\left(t_{k+1}\right)}{2^{L}-1} . \tag{3.33}
\end{equation*}
$$

The improved estimate in equation (3.33) combines two $L$ th-order expansions and is thus computationally more expensive than a single $L$ th-order expansion. For example, consider a first-order Padé method. If Richardson extrapolation is employed with two first-order methods, then the total computational effort is equivalent to a second-order expansion around a single point. (This is shown in the Appendix B for the specific case of the exponential function). However, the results in the next section show that the Richardson's Extrapolation in conjunction with Padé-based methods gives better results than a basic second-order expansion for the same time step. The conclusion is that a much larger time step may be employed to achieve a specific accuracy tolerance.

Thus, savings are obtained from a computational viewpoint by utilising the Richardson extrapolation as although two function evaluations are required at each time step, there are fewer time steps involved.

### 3.1.4 Padé Approximation with Predictor-Corrector

With the predictor-corrector method, the accuracy of the Pade approximation is further increased and it is not necessary to calculate the higher-order derivatives. The basis to form a corrector stage for the single-step method is as follows: a Padé approximation of order $p$ matches the first $p+1$ coefficients (time-domain moments) of a Taylor series expansion. It also provides additional terms. Considering the test function $e^{-t}$. Its Taylor series about $t=0$ is:

$$
\begin{equation*}
e^{-t}=1-h+\frac{h^{2}}{2}-\frac{h^{3}}{6}+\cdots \tag{3.34}
\end{equation*}
$$

A first-order Padé approximation for $e^{-t}$ about $t=0$ is given by:

$$
\begin{equation*}
y_{a p p}=\frac{1}{1+h} . \tag{3.35}
\end{equation*}
$$

This function matches the first two coefficients of a Taylor series expansion for $e^{-t}$. It also produces additional terms, the first of which is:

$$
\begin{equation*}
T_{3_{-} P}=h^{2} . \tag{3.36}
\end{equation*}
$$

However, the correct third coefficient in a Taylor series expansion for $e^{-t}$ as seen in equation (3.34) is:

$$
\begin{equation*}
T_{3-T}=\frac{h^{2}}{2} . \tag{3.37}
\end{equation*}
$$

Now, a corrector is chosen so as to match $T_{3}$ for the particular test function, $y=e^{-t}$, without requiring a higher-order derivative. So for this test case, the corrector would be:

$$
\begin{equation*}
T_{\text {corr }}=-\frac{h^{2}}{2}=f(0) \frac{h^{2}}{2} . \tag{3.38}
\end{equation*}
$$

where $f(0)=-1$. Bearing this in mind, the choice of corrector for a general singlestep first-order Padé method is selected as follows:

$$
\begin{equation*}
\hat{y}_{i}\left(t_{k+1}\right)=\bar{y}_{i}\left(t_{k+1}\right)+\frac{1}{2} h^{2} \bar{f}\left(t_{k+1}\right) . \tag{3.39}
\end{equation*}
$$

$\hat{y}_{i}\left(t_{k+1}\right)$ is the corrected estimate of $y_{i}\left(t_{k+1}\right) . \bar{f}\left(t_{k+1}\right)$ is the estimate of the first derivative of $y_{i}\left(t_{k+1}\right)$. For higher-order methods, a similar procedure is followed:

$$
\begin{equation*}
\hat{y}_{i}\left(t_{k+1}\right)=\bar{y}_{i}\left(t_{k+1}\right)+\frac{1}{\alpha} h^{p+1} \bar{f}^{(p-1)}\left(t_{k+1}\right) . \tag{3.40}
\end{equation*}
$$

where $1 / \alpha$ is the amount required to correct the $(p+2)$ th coefficient in a $p$ th order Padé approximation for $e^{-t}$ about $t=0$.

### 3.2 Multi-Step Padé-Based Algorithms

### 3.2.1 Explicit Multi-Step Padé Approximation

To derive a multi-step formula, recall again:

$$
\begin{equation*}
y\left(t_{k+1}\right)=y\left(t_{k}\right)+\int_{t_{k}}^{t_{k+1}} f(y(t), t) d t \tag{3.41}
\end{equation*}
$$

In this case, a Padé approximation is used to interpolate the function, $f$, through $m+$ $n+1$ interpolation points starting from $t_{i}$ and finishing at $t_{k} \cdot m+n$ is the order of the Padé approximation:

$$
\begin{equation*}
f(y, t)=\frac{\sum_{j=0}^{m} p_{j}\left(t-t_{i}\right)^{j}}{1+\sum_{j=1}^{n} q_{j}\left(t-t_{i}\right)^{j}} . \tag{3.42}
\end{equation*}
$$

The function $f(y, t)$ can also be approximated by an $(m+n)$ th order polynomial as below

$$
\begin{equation*}
f(y, t)=c_{0}+c_{1}\left(t-t_{i}\right)+c_{2}\left(t-t_{i}\right)^{2}+\cdots+c_{m+n}\left(t-t_{i}\right)^{m+n} . \tag{3.43}
\end{equation*}
$$

Then the values of $f(y, t)$ at the different times are

$$
\begin{align*}
& f_{t_{i}}= \sum_{j=0}^{m+n} c_{j}\left(t_{i}-t_{i}\right)^{j} \\
& f_{t_{i+1}}= \sum_{j=0}^{m+n} c_{j}\left(t_{i+1}-t_{i}\right)^{j} \\
& \vdots  \tag{3.44}\\
& f_{t_{k}}= \sum_{j=0}^{m+n} c_{j}\left(t_{k}-t_{i}\right)^{j}
\end{align*}
$$

The coefficients $c_{j}$ are obtained by solving (3.44). Then the coefficients $p_{j}$ and $q_{j}$ are determined by equating the Padé approximation in equation (3.42) with the series expansion in equation (3.43):

$$
\begin{equation*}
\frac{\sum_{j=0}^{m} p_{j}\left(t-t_{i}\right)^{j}}{1+\sum_{j=1}^{m} q_{j}\left(t-t_{i}\right)^{j}}=\sum_{j=0}^{m+n} c_{j}\left(t-t_{i}\right)^{j} \tag{3.45}
\end{equation*}
$$

where $t=\left\{t_{i}, t_{i+1}, \ldots, t_{k}\right\}$.
Replacing $f$ in equation (3.41) by the expression in equation (3.42) and performing the integration analytically yields a multi-step formula for $y\left(t_{k+1}\right)$. This is an explicit formula for $y\left(t_{k+1}\right)$. For example, if a first-order Pade approximation is employed with a time step of $h=t_{k+1}-t_{k}$, the resulting explicit formula is:

$$
\begin{equation*}
y\left(t_{k+1}\right)=y\left(t_{k}\right)+\int_{t_{k}}^{t_{k+1}} \frac{p_{0}}{1+q_{1}\left(t-t_{k-1}\right)} d t \tag{3.46}
\end{equation*}
$$

In this case, $m=0$ and $n=1$.

$$
\begin{equation*}
y\left(t_{k+1}\right)=y\left(t_{k}\right)+\int_{h}^{2 h} \frac{p_{0}}{1+q_{1} t} d t \tag{3.47}
\end{equation*}
$$

which results in

$$
\begin{equation*}
y\left(t_{k+1}\right)=y\left(t_{k}\right)+\frac{p_{0}}{q_{1}}\left(\log \left(1+2 h q_{1}\right)-\log \left(1+h q_{1}\right)\right) \tag{3.48}
\end{equation*}
$$

The first-order polynomial approximation is:

$$
\begin{align*}
f(y, t) & =c_{0}+c_{1}\left(t-t_{0}\right) \\
& =c_{0}+c_{1} t . \tag{3.49}
\end{align*}
$$

Thus $f$ at different times are

$$
\begin{align*}
& f\left(t_{0}\right)=c_{0}+c_{1} t_{0}=c_{0}  \tag{3.50}\\
& f\left(t_{1}\right)=c_{0}+c_{1} t_{1}=c_{0}+c_{1} h . \tag{3.51}
\end{align*}
$$

Then the coefficients $c_{j}$ are obtained:

$$
\begin{align*}
& c_{0}=f\left(t_{0}\right)  \tag{3.52}\\
& c_{1}=\frac{f\left(t_{1}\right)-f\left(t_{0}\right)}{h} . \tag{3.53}
\end{align*}
$$

Substituting (3.52) and (3.53) into (3.45), when the order is 1 , yields

$$
\begin{equation*}
\frac{p_{0}}{1+q_{1} t}=f\left(t_{0}\right)+\frac{f\left(t_{1}\right)-f\left(t_{0}\right)}{h} t . \tag{3.54}
\end{equation*}
$$

The values of $p_{0}$ and $q_{1}$ are obtained by solving (3.54):

$$
\begin{align*}
& p_{0}=f\left(t_{0}\right)  \tag{3.55}\\
& q_{1}=-\frac{f\left(t_{1}\right)-f\left(t_{0}\right)}{f\left(t_{0}\right) h} . \tag{3.56}
\end{align*}
$$

By substituting (3.55) and (3.56) into (3.48), the Padé approximation to (3.1) is expressed as:

$$
\begin{align*}
y\left(t_{k+1}\right)= & y\left(t_{k}\right) \\
& -\frac{f^{2}\left(t_{0}\right) h}{f\left(t_{1}\right)-f\left(t_{0}\right)}\left(\log \left(1+2 \frac{f\left(t_{0}\right)-f\left(t_{1}\right)}{f\left(t_{0}\right)}\right)-\log \left(1+\frac{f\left(t_{0}\right)-f\left(t_{1}\right)}{f\left(t_{0}\right)}\right)\right) . \tag{3.57}
\end{align*}
$$

### 3.2.2 Implicit Multi-Step Padé Approximation

In a similar manner to the derivation of the Adam's Moulton predictor-corrector method, a corrector for the multi-step method in equations (3.41)-(3.57) may be formed by interpolating $f$ through points including $t_{k+1}$. When the resultant formula for $f(t, y)$ is inserted in equation (3.41), the outcome is an implicit formula for $y\left(t_{k+1}\right)$. The key difference from the explicit multi-step Padé approximation is that the implicit multi-step Padé approximation uses the following equations to find $c_{j}$ :

$$
\begin{align*}
& f_{t_{i+1}}= \sum_{j=0}^{m+n} c_{j}\left(t_{i+1}-t_{i}\right)^{j} \\
& \vdots  \tag{3.58}\\
& f_{t_{k}}= \sum_{j=0}^{m+n} c_{j}\left(t_{k}-t_{i}\right)^{j} \\
& f_{t_{k+1}}= \sum_{j=0}^{m+n} c_{j}\left(t_{k+1}-t_{i}\right)^{j}
\end{align*}
$$

In order to solve (3.58), $f_{t_{k+1}}$ has to be predicted first. If the function $f$ is nonlinear, then a nonlinear solver is required to solve for $y\left(t_{k+1}\right)$ and this step can be computationally expensive. Hence, rather than solving the nonlinear equation set directly, the value for $y\left(t_{k+1}\right)$ in the right-hand side of the implicit formula is replaced with its estimate from use of the explicit predictor formula-the explicit multi-step Padé approximation. This is the standard procedure in predictor-corrector methods such as the Adams-Moulton method.

### 3.3 Application and Validation

### 3.3.1 Classic Test System

The following well-known classic equation system [29] is chosen as the first illustrative example. Its analytical solution is known, equations (3.63) and (3.64) are an exact
reference solution. In addition, since it constitutes a stiff system of differential equations, the accuracy and efficiency of new methods can be thoroughly investigated. The system equations are:

$$
\begin{align*}
& \frac{d u}{d t}=998 u+1998 v  \tag{3.59}\\
& \frac{d v}{d t}=-999 u-1999 v \tag{3.60}
\end{align*}
$$

where

$$
\begin{align*}
& u_{0}=1  \tag{3.61}\\
& v_{0}=1 . \tag{3.62}
\end{align*}
$$

The exact analytical solution is given by:

$$
\begin{align*}
& u(t)=4 e^{-t}-3 e^{-1000 t}  \tag{3.63}\\
& v(t)=-2 e^{-t}+3 e^{-1000 t} \tag{3.64}
\end{align*}
$$

Figure 3.1 shows, superimposed on the analytical solution, the result computed with the standard third-order Adams-Moulton predictor-corrector method for a stepsize of 0.001 s . Figure 3.2 illustrates the simulation results from the third-order RungeKutta method. Figure 3.3 shows the corresponding result computed with the basic third-order Padé single-step method of equation (3.3). As evidenced by these results, the new technique is in fact more accurate than the standard Adams-Moulton and Runge-Kutta method of the same order. To compare methods, the root mean squared (RMS) error is evaluated. The third-order Adams-Moulton technique yields an error of $1.53 \times 10^{-1}$ for the given step size. The third-order Runge-Kutta method has an error of $6.65 \times 10^{-2}$ for the same step size. The basic third-order Padé single-step method yields an error of $4.94 \times 10^{-2}$ for the same step size.

When the combined third-order Padé and exponential function method is used, the error is dramatically reduced to $3.93 \times 10^{-12}$. When a first-order Padé and exponential
function is employed the error is $1.45 \times 10^{-4}$. Thus, the addition of the exponential term enables a specific error tolerance to be met with a lower-order Padé function.


Figure 3.1: The comparison of the simulation results from Adams Moulton (一) and exact analysis ( ${ }^{* *}$ ).

Finally, Richardson extrapolation is applied to the basic third-order Padé method. The error now is $1.4 \times 10^{-3}$. This goal of the Richardson extrapolation method is to eliminate the higher-order errors. When the method is applied to the third-order Padé


Figure 3.2: The comparison of the simulation results from Runge-Kutta (一) and exact analysis (**).


Figure 3.3: The comparison of the simulation results from Padé approximation (—) and exact analysis ( ${ }^{*}$ ).
method, the aim is to eliminate fourth-order errors. However, the basic fourth-order Taylor series approximation results in an error of $2.01 \times 10^{-2}$. Thus, the Richardson extrapolation achieves greater accuracy and consequently, would enable the use of a significantly larger time-step for the same level of accuracy as a basic fourth-order method. Although use of Richardson extrapolation requires an extra function evaluation at each time-step, the time-step chosen can be considerably greater than that for the basic method to achieve comparable accuracy. Thus, the total number of function evaluations for a simulation is reduced. For example, if the time-step size is doubled when the Richardson extrapolation is applied to the basic third-order Padé method, the error is $1.3 \times 10^{-2}$.

All of the Padé-based methods result in an RMS error that is less than the AdamsMoulton method. This means that for a fixed accuracy tolerance, a larger step size can be employed if the new methods are used as the ODE solver.

Figure 3.4 shows the result computed with the Padé-based multi-step predictorcorrector method of Section 3.2. Again, a step size of $h=0.001 \mathrm{~s}$ is employed. The RMS error when first-order Padé approximants are used is $2.1 \times 10^{-2}$. Thus, Padébased predictor-corrector methods are effective and efficient when compared to the standard Adams Moulton method.


Figure 3.4: The comparison of the simulation results from Padé-based multi-step predictor-corrector method (-) and exact analysis (**).

The total simulation time of different methods, which simulate the example system (3.59) and (3.60) for 0.2 s with the same time-step size $h=0.001 \mathrm{~s}$, is illustrated in Table 3.1. The Padé-Exponential method costs the least time, while the 3rd-Order AdamMoulton method is the slowest one. Note that the third-order Runge-Kutta method is not accurate enough, though it is faster than the Padé-Taylor method.

Table 3.1: Total simulation time of different methods

| Numerical Integration Methods | Simulation Time (s) |
| :---: | :---: |
| Padé-Taylor | 0.017 |
| Padé-Exponential | 0.0016 |
| Padé-Exponential-Predictor-Corrector | 0.0037 |
| 3rd-Order Adam-Moulton | 0.041 |
| 3rd-Order Runge-Kutta | 0.009 |

### 3.3.2 MESFET Amplifier

As a second example to further confirm the utility of the Padé-based methods, Figure 3.5 shows a simple single-ended MESFET amplifier. The amplifier is described by ten non-linear differential equations which are stiff in nature. The complete details of the amplifier circuit and the model employed in it are given in the Appendix C. Figure 3.6 shows a short segment of the output voltage $\mathrm{v}(\mathrm{t})$ obtained with the fourth-order Adams-Moulton predictor-corrector. The input is a 2 GHz cosine wave.

A time-step of $\sim 0.12$ ps was required to solve the given system of equations with this technique. Figure 3.6 also shows the corresponding result obtained with the new Padé technique of Section 3.1.1 and the single-step corrector of Section 3.1.4. The new technique permits the use of a step-size ten times larger than the Adams Moulton technique, i.e., 1.2 ps , for a comparable level of accuracy. This constitutes a remarkable saving in computing time.


Figure 3.5: Schematics of the MESFET amplifier.


Figure 3.6: The comparison of the simulation results from Adams-Moulton (一) and Padé method (**).

### 3.4 Conclusions

Several Padé-based algorithms are proposed to improve the accuracy and efficiency when solving the stiff systems. The Padé-Taylor method is the basic one which obtains the parameters by comparison with the Taylor series. The Padé-exponential method is used to increase the efficiency if the system has exponential components. The Richardson method is applied to Padé method to avoid the computation of higher-order derivatives. The predictor-corrector method is also used to further reduce the error. The simulation results for the examples in question show that Padé method is better than the Adams-Moulton technique in terms of both accuracy and efficiency.

## Chapter 4

## Filon-Type Numerical Quadrature for Highly-Oscillatory Systems.

### 4.1 Background to Highly Oscillatory Quadrature

High-frequency signals abound in Radio Frequency (RF) communication systems. This is a consequence of modulation: the imposition of a lower-frequency information signal onto a high-frequency carrier. The goal is to enable antennae of a manageable size to be employed for information transmission. Antennae of an impractical size would be required if modulation was not performed. In RF communication systems, signals in the MHz frequency range and higher are common. Furthermore, nonlinearities abound in RF transmission systems owing to the presence of solid-state amplifiers, mixers and so on [30].

Most RF systems involve a linear part and a nonlinear part with the linear part due to the presence of linear resistors, inductors and capacitors and the nonlinear part due to amplifiers, mixers or nonlinear and controlled resistors and capacitors. A typical equation system is of the form:

$$
\begin{align*}
& y^{\prime}=A y+g(t, y)  \tag{4.1}\\
& y_{0}=y(0) \tag{4.2}
\end{align*}
$$

### 4.1 Background to Highly Oscillatory Quadrature

where $t \geq 0$. The general solution of (4.1) may be written in the form:

$$
\begin{equation*}
y(t)=e^{t A} y_{0}+\int_{0}^{t} e^{A(t-\tau)} g(\tau, y(\tau)) d \tau \tag{4.3}
\end{equation*}
$$

The recent explosion of developments in the RF and telecommunications industry has put pressure on circuit designers for faster simulations, faster designs and faster product output and the existing Computer Aided Design (CAD) tools have struggled to keep pace. In addition, the growing complexity of the modulation formats is rendering software tools unacceptably slow and consequently, unsatisfactory. There is therefore, an urgent need for a complete revamp and update of the fundamental numerical processes within these CAD packages taking into account the modern developments and formats.

Most numerical simulators of electric and electronic circuits, such as SPICE [25], as well as general-purpose solvers of ordinary differential equations (ODEs), like those in MATLAB, use either multi-step or Runge-Kutta methods. This is perfectly adequate for a great majority of ODEs in applications, yet falls woefully short for systems subject to modulated signals or RF oscillators. In this setting, traditional quadrature approaches can necessitate the use of minute step-sizes with the consequent outcome of great inefficiencies and often impractical simulation times.

Modulation is the process whereby information is transmitted at a high frequency to enable antennae of practical dimensions to be employed. In amplitude modulation (AM), the information signal (envelope) has a low frequency content relative to the carrier, $A_{m} \sin \omega_{m} t$. The resultant amplitude-modulated signal is therefore $A_{c}(1+$ $\left.m \sin \omega_{m} t\right) \sin \omega_{c} t$, where $m=A_{m} / A_{c}$. Another variation of amplitude modulation is the Double-Sided Suppressed Carrier $A_{m} \sin \omega_{m} t \sin \omega_{c} t . \omega_{c}=2 \pi f_{c}$ is the carrier frequency in $\mathrm{rad} / \mathrm{s}$ and $A_{c}$ is its amplitude. In general AM, the information signal is in the kHz range, while the carrier signal is in the MHz or GHz range.

In digital modulation, the information to be transmitted is a sequence of ones and zeros, termed bits. The amplitude, frequency or phase of a carrier signal is varied dependent on the bit value. For example, in Binary Phase Shift Keying (BPSK) the modulated signal is of the form $b(t)=\cos \left(\omega_{c} t\right) x(t)$, where $x(t)$ is $+A$ or $-A$ if a ' 1 ' or a ' 0 ' bit is to be transmitted, respectively. In digital technologies involving more complex formats, such as EDGE, the information/envelopes have bandwidths in the kHz range, while the carriers are 800 MHz and 1800 MHz . For evaluation of a bit error ratio of such an RF transmission, several information envelope time periods are required, but the step size is governed by the underlying carrier frequency which is significantly higher than the envelope frequency.

This chapter will address the issue of simulating ODEs involving very high frequencies and widely varying frequency ranges using Filon-type methods, applied in tandem with exponential integrators. This results in increased efficiencies for systems involving signal of widely varying frequencies. Similar techniques have been already investigated for different models of highly oscillatory ODEs in [31].

### 4.2 Filon-Type Numerical Approximation

The following equation

$$
\begin{equation*}
F[f]=\int_{a}^{b} f(t) G(t, \omega) d t \tag{4.4}
\end{equation*}
$$

is assumed as a highly oscillatory integral: in particular, the oscillator $G$ oscillates rapidly when the oscillatory parameter $\omega$ is large, i.e., $\omega \gg 1$, while $f$ itself is nonoscillatory. A typical example of an oscillator is $G(x, \omega)=e^{i \omega g(x)}$ where $g$ is some given smooth function. The calculation of (4.4) for large $\omega$ by classical methods (e.g. Gaussian quadrature) is prohibitively expensive, but such integrals can be calculated with relative ease using Filon-type integrators [32].

The idea behind Filon-type methods is to replace the function $f$ in (4.4) by a polynomial interpolation. Specifically, let

$$
\begin{equation*}
a=c_{1}<c_{2}<\cdots<c_{q}=b \tag{4.5}
\end{equation*}
$$

be given nodes, each $c_{k}$ has a multiplicity $m_{k} \geq 1$. Thus, a polynomial $p$ of degree $r=\sum_{k=1}^{q} m_{k}-1$ is constructed as:

$$
\begin{equation*}
p^{(j)}\left(c_{k}\right)=f^{(j)}\left(c_{k}\right) \tag{4.6}
\end{equation*}
$$

where $j=0,1, \ldots, m_{k}-1$ and $k=1,2, \ldots, q$. The Filon-type method for the integral (4.4) is

$$
\begin{equation*}
Q_{F}[f]=\int_{a}^{b} p(t) G(t, \omega) d t \tag{4.7}
\end{equation*}
$$

It is based on the assumption that the integral (4.4) can be calculated explicitly for function $f$.

The interpolating polynomial $p$ can be written explicitly in the form:

$$
\begin{equation*}
p(t)=\sum_{k=1}^{q} \sum_{j=0}^{m_{k}-1} \alpha_{k, j}(t) f^{(j)}\left(c_{k}\right) \tag{4.8}
\end{equation*}
$$

where each $\alpha_{k, j}$ is the cardinal polynomial of Hermite interpolation, $\alpha_{k, j}^{(j)}\left(c_{k}\right)=1$, otherwise $\alpha_{k, j}^{(i)}\left(c_{l}\right)=0$ for all $i=0,1, \ldots, m_{l}-1$, and $l=1,2, \ldots, q$. Therefore (4.7) can be written in the form

$$
\begin{equation*}
Q_{F}[f]=\sum_{k=1}^{q} \sum_{j=0}^{m_{k}-1} b_{k, j}(\omega) f^{(j)}\left(c_{k}\right) \tag{4.9}
\end{equation*}
$$

where

$$
\begin{equation*}
b_{k, j}(\omega)=\int_{a}^{b} \alpha_{k, j}(t) G(t, \omega) d t \tag{4.10}
\end{equation*}
$$

Then the suitable nodes and multiplicities are chosen by the Filon-type methods so as to increase the accuracy and efficiency.

### 4.3 Example

Here, the diode rectifier circuit shown in Fig. 4.1 is chosen as an example to verify the accuracy and efficiency of the Filon-type numerical quadrature, because its model is in the form of (4.1).


Figure 4.1: Diode rectifier circuit.

The governing equation for the circuit in Fig. 4.1 is

$$
\begin{equation*}
C \frac{d v(t)}{d t}+\frac{v(t)}{R}=I_{0}\left[e^{k(b(t)-v(t))}-1\right] \tag{4.11}
\end{equation*}
$$

where $C$ is the capacitor value, $R$ is the resistor value, $I_{0}$ and $k$ are constants, $b(t)$ is the input signal and the unknown $v(t)$ is the voltage.

In order to modify (4.11) to be in the form of the variation constants formula (4.3), each side of (4.11) is multiplied by $e^{-A t}$, where $A=-1 / R C$.

$$
\begin{equation*}
e^{t / R C} \frac{d v(t)}{d t}+e^{t / R C} \frac{v(t)}{R C}=e^{t / R C} \frac{I_{0}}{C}\left[e^{k(b(t)-v(t))}-1\right] . \tag{4.12}
\end{equation*}
$$

Since the left side

$$
\begin{equation*}
e^{t / R C} \frac{d v(t)}{d t}+e^{t / R C} \frac{v(t)}{R C}=\frac{d}{d t}\left(e^{t / R C} v(t)\right) \tag{4.13}
\end{equation*}
$$

(4.12) becomes

$$
\begin{equation*}
\frac{d}{d t}\left(e^{t / R C} v(t)\right)=e^{t / R C} \frac{I_{0}}{C}\left[e^{k(b(t)-v(t))}-1\right] . \tag{4.14}
\end{equation*}
$$

If (4.14) is integrated from $t_{n}$ to $t_{n+1}$, it is obtained

$$
\begin{equation*}
\left.e^{t / R C} v(t)\right|_{t_{n}} ^{t_{n+1}}=-\left.R I_{0} e^{t / R C}\right|_{t_{n}} ^{t_{n+1}}+\frac{I_{0}}{C} \int_{t_{n}}^{t_{n+1}} \exp \left(\frac{t}{R C}-k v(t)+k b(t)\right) d t \tag{4.15}
\end{equation*}
$$

Thus

$$
\begin{align*}
e^{t_{n+1} / R C} v_{n+1}-e^{t_{n} / R C} v_{n}= & R I_{0}\left[e^{t_{n} /(R C)}-e^{t_{n+1} /(R C)}\right] \\
& +\frac{I_{0}}{C} \int_{t_{n}}^{t_{n+1}} \exp \left(\frac{t}{R C}-k v(t)+k b(t)\right) d t . \tag{4.16}
\end{align*}
$$

By rearranging (4.16) with the time-step $h=t_{n+1}-t_{n}$, we get

$$
\begin{align*}
v_{n+1} & =v_{n} e^{-h /(R C)}+R I_{0}\left[e^{-h /(R C)}-1\right]+\frac{I_{0}}{C} \int_{t_{n}}^{t_{n+1}} \exp \left(\frac{t-t_{n+1}}{R C}-k v(t)+k b(t)\right) d t \\
& =v_{n} e^{-h /(R C)}+R I_{0}\left[e^{-h /(R C)}-1\right]+\frac{I_{0}}{C} \Gamma[g] \tag{4.17}
\end{align*}
$$

where

$$
\begin{align*}
& g(t)=\exp \left(\frac{t-t_{n+1}}{R C}-k v(t)+k b(t)\right)  \tag{4.18}\\
& \Gamma[g]=\int_{t_{n}}^{t_{n+1}} g(t) d t \tag{4.19}
\end{align*}
$$

If $b(t)=\varepsilon \cos \vartheta(t)$, then the integral in (4.17) is of the form

$$
\begin{equation*}
\Gamma[g]=\int_{t_{n}}^{t_{n+1}} f(t, v(t)) e^{z \cos \vartheta(t)} d t \tag{4.20}
\end{equation*}
$$

where

$$
\begin{align*}
z & =k \varepsilon  \tag{4.21}\\
f(t, v(t)) & =\exp \left[\frac{t-t_{n+1}}{R C}-k v(t)\right] . \tag{4.22}
\end{align*}
$$

From [33],

$$
\begin{align*}
& e^{z \sin \vartheta}=\mathrm{I}_{0}(z)+2 \sum_{m=0}^{\infty}(-1)^{m} \mathrm{I}_{2 m+1}(z) \sin (2 m+1) \vartheta+2 \sum_{m=1}^{\infty}(-1)^{m} \mathrm{I}_{2 m}(z) \cos 2 m \vartheta  \tag{4.23}\\
& e^{z \cos \vartheta}=\mathrm{I}_{0}(z)+2 \sum_{m=1}^{\infty} \mathrm{I}_{m}(z) \cos m \vartheta \tag{4.24}
\end{align*}
$$

where $\mathrm{I}_{m}$ is the $m$ th modified Bessel function.
Consequently, the integral $\Gamma[g]$ can be written in the form

$$
\begin{align*}
\Gamma[g] & =\int_{t_{n}}^{t_{n+1}} f(t, v(t)) e^{z \cos \vartheta(t)} d t \\
& =\mathrm{I}_{0}(z) \int_{t_{n}}^{t_{n+1}} f(t, v(t)) d t+2 \sum_{m=1}^{\infty} \mathrm{I}_{m}(z) \int_{t_{n}}^{t_{n+1}} f(t, v(t)) \cos (m \vartheta(t)) d t \tag{4.25}
\end{align*}
$$

thereby expressing it as an infinite sum of integrals. However, the highly oscillatory integrals on the right of (4.25) are amenable to very rapid and efficient numerical calculation with Filon-type methods [32].

Applying the Filon-type method (4.7) to (4.25), it follows that:

$$
\begin{equation*}
\Gamma_{F}[g]=\mathrm{I}_{0}(z) \int_{t_{n}}^{t_{n+1}} p(t) d t+2 \sum_{m=1}^{\infty} \mathrm{I}_{m}(z) \int_{t_{n}}^{t_{n+1}} p(t) \cos (m \vartheta(t)) d t \tag{4.26}
\end{equation*}
$$

where $p(t)$ interpolates the function $f(t, v(t))$.
As shown in [33]:

$$
\begin{align*}
& \mathrm{I}_{m}(z)=\mathrm{i}^{-m} \mathrm{~J}_{m}(i z)  \tag{4.27}\\
& \mathrm{J}_{m}(z)=\frac{\left(\frac{e z}{2 m}\right)^{m}}{\sqrt{2 \pi m}} \tag{4.28}
\end{align*}
$$

it is true that

$$
\begin{equation*}
\mathrm{I}_{m}(z) \sim \frac{1}{\sqrt{2 \pi m}}\left(\frac{e z}{2 m}\right)^{m} \tag{4.29}
\end{equation*}
$$

### 4.3 Example

where $z \in \Re$ and $m \gg 1$. Thus the infinite series in (4.26) converges very rapidly such that

$$
\begin{equation*}
\Gamma_{F}[g]=\mathrm{I}_{0}(z) \int_{t_{n}}^{t_{n+1}} p(t) d t+2 \sum_{m=1}^{N} \mathrm{I}_{m}(z) \int_{t_{n}}^{t_{n+1}} p(t) \cos (m \vartheta(t)) d t \tag{4.30}
\end{equation*}
$$

for a relatively small value of $N$.
Due to the fact that the unknown $v(t)$ features inside the integral sign-the ODE is nonlinear. Hence, the variation of constants formula (4.17) is implicit. To this end, waveform relaxation (WR) is employed to solve it. Several waveform relaxation methods have been developed in the last few decades, see for example [34] [35] [36].

The standard form of WR for (4.3) is

$$
\begin{align*}
& y^{[0]}(t)=y_{0} \\
& y^{[s]}(t)=e^{t A} y_{0}+\int_{0}^{t} e^{A(t-\tau)} g\left(\tau, y^{[s-1]}(\tau)\right) d \tau \tag{4.31}
\end{align*}
$$

where $s=1,2,3, \ldots$.
Combining the Filon-type quadrature with (4.31) results in the following iterative
scheme, that is executed for each time step.

$$
\begin{align*}
v_{n+1}^{[0]}= & v_{0} \\
& \tilde{v}^{[0]}(t)=\frac{t_{n+1}-t}{h} v_{n}+\frac{t-t_{n}}{h} v_{n+1}^{[0]}\left(\equiv v_{n}\right) \\
v_{n+1}^{[1]}= & v_{n} e^{-h / R C}+R I_{0}\left[e^{-h / R C}-1\right]+\frac{I_{0}}{C} \int_{t_{n}}^{t_{n+1}} e^{k b(t)} p\left(t, \tilde{v}^{[0]}(t)\right) d t \\
& \tilde{v}^{[1]}(t)=\frac{t_{n+1}-t}{h} v_{n}+\frac{t-t_{n}}{h} v_{n+1}^{[1]} \\
v_{n+1}^{[2]}= & v_{n} e^{-h / R C}+R I_{0}\left[e^{-h / R C}-1\right]+\frac{I_{0}}{C} \int_{t_{n}}^{t_{n+1}} e^{k b(t)} p\left(t, \tilde{v}^{[1]}(t)\right) d t \\
& \tilde{v}^{[2]}(t)=\frac{t_{n+1}-t}{h} v_{n}+\frac{t-t_{n}}{h} v_{n+1}^{[2]} \\
& \vdots \\
v_{n+1}^{[s]}= & v_{n} e^{-h / R C}+R I_{0}\left[e^{-h / R C}-1\right]+\frac{I_{0}}{C} \int_{t_{n}}^{t_{n+1}} e^{k b(t)} p\left(t, \tilde{v}^{[s-1]}(t)\right) d t \\
& \tilde{v}^{[s]}(t)=\frac{t_{n+1}-t}{h} v_{n}+\frac{t-t_{n}}{h} v_{n+1}^{[s]} \tag{4.32}
\end{align*}
$$

Thus, the Filon nodes are $c_{1}=t_{n}$ and $c_{2}=t_{n+1}$, with unit multiplicities at both points: the linear polynomial $p$ agrees with $f\left(t, \tilde{v}^{[i]}(t)\right)$ at $t=t_{n}$ and $t_{n+1}$. The iteration is terminated once $\left|v_{n+1}^{[r]}-v_{n+1}^{[r-1]}\right|<t o l$, hence we let $v_{n+1}=v_{n+1}^{[r]}$.

The next example to consider is the more complicated case of analogue amplitude modulation, whereby $b(t)=\sin \left(\omega_{1} t\right) \sin \left(\omega_{2} t\right)$. The first step is to rewrite $b(t)$ in the form

$$
\begin{equation*}
b(t)=\frac{1}{2}\left[\cos \left(\tilde{\omega}_{1} t\right)-\cos \left(\tilde{\omega}_{2} t\right)\right] \tag{4.33}
\end{equation*}
$$

where

$$
\begin{align*}
& \tilde{\omega}_{1}=\omega_{1}-\omega_{2}  \tag{4.34}\\
& \tilde{\omega}_{2}=\omega_{1}+\omega_{2} . \tag{4.35}
\end{align*}
$$

This enables the integral $\Gamma[g]$ to be expressed as a product of two infinite series,

$$
\begin{align*}
\Gamma[g]= & \int_{t_{n}}^{t_{n+1}} f(t, v(t)) \exp \left(\frac{1}{2} z \cos \tilde{\omega}_{1} t-\frac{1}{2} z \cos \tilde{\omega}_{2} t\right) d t \\
= & {\left[\mathrm{I}_{0}\left(\frac{1}{2} z\right)\right]^{2} \int_{t_{n}}^{t_{n+1}} f(t, v(t)) d t } \\
& +2 \mathrm{I}_{0}\left(\frac{1}{2} z\right) \sum_{m=1}^{\infty} \mathrm{I}_{m}\left(\frac{1}{2} z\right) \int_{t_{n}}^{t_{n+1}} f(t, v(t)) \cos \left(m \tilde{\omega}_{1} t\right) d t \\
& +2 \mathrm{I}_{0}\left(\frac{1}{2} z\right) \sum_{l=1}^{\infty}(-1)^{l} \mathrm{I}_{l}\left(\frac{1}{2} z\right) \int_{t_{n}}^{t_{n+1}} f(t, v(t)) \cos \left(l \tilde{\omega}_{2} t\right) d t \\
& +4 \sum_{m=1}^{\infty} \sum_{l=1}^{\infty}(-1)^{l} \mathrm{I}_{m}\left(\frac{1}{2} z\right) \mathrm{I}_{l}\left(\frac{1}{2} z\right) \int_{t_{n}}^{t_{n+1}} f(t, v(t)) \cos \left(m \tilde{\omega}_{1} t\right) \cos \left(l \tilde{\omega}_{2} t\right) d t \tag{4.36}
\end{align*}
$$

Following the same procedure as before, the Filon-type method corresponding to (4.37) is

$$
\begin{align*}
\Gamma[g]= & {\left[\mathrm{I}_{0}\left(\frac{1}{2} z\right)\right]^{2} \int_{t_{n}}^{t_{n+1}} p(t) d t+2 \mathrm{I}_{0}\left(\frac{1}{2} z\right) \sum_{m=1}^{\infty} \int_{t_{n}}^{t_{n+1}} p(t) \cos \left(m \tilde{\omega}_{1} t\right) d t } \\
& +2 \mathrm{I}_{0}\left(\frac{1}{2} z\right) \sum_{l=1}^{\infty}(-1)^{l} \mathrm{I}_{l}\left(\frac{1}{2} z\right) \int_{t_{n}}^{t_{n+1}} p(t) \cos \left(l \tilde{\omega}_{2} t\right) d t \\
& +4 \sum_{m=1}^{\infty} \sum_{l=1}^{\infty}(-1)^{l} \mathrm{I}_{m}\left(\frac{1}{2} z\right) \mathrm{I}_{l}\left(\frac{1}{2} z\right) \int_{t_{n}}^{t_{n+1}} p(t) \cos \left(m \tilde{\omega}_{1} t\right) \cos \left(l \tilde{\omega}_{2} t\right) d t \tag{4.37}
\end{align*}
$$

where $p(t)$ is the interpolating polynomial of $f$.

### 4.4 Numerical Results

The proposed method will be illustrated with two input signals, one is amplitude modulated and the second is digitally modulated. The values of the parameters are specif-
ically chosen to emphasise both the highly oscillatory nature of the responses and the important features of the algorithm presented before.

### 4.4.1 Amplitude Modulation

For this case, $b(t)=\sin \omega_{1} t \sin \omega_{2} t$, with $\omega_{1}=100 \mathrm{rad} / \mathrm{s}$ and $\omega_{2}=10^{7} \mathrm{rad} / \mathrm{s}$. There is a difference of five orders of magnitude between the two oscillation scales. The values are set as: $I_{0}=1 A, C=1 f, R=1 \Omega$ and $k=40$.


Figure 4.2: The exact solution of the amplitude-modulated equation with $I_{0}=1$, $C=1, R=1$ and $k=40$.

Figure 4.2 depicts the 'exact' solution of the ODE-(this was formed using a numerical solution with an exceedingly small step size). It is clear that the solution is a slow-varying wave and the extraordinarily large frequency $\omega_{2}$ is not visible. Nonetheless, the presence of high oscillation is enough to render traditional numerical methods inefficient. In Figs. 4.3-4.5 we exhibit the outcome of integrating the ODE in the interval $[0,0.2]$ with the constant step size $h=2.5 \times 10^{-4}$ with three numerical methods.

In each case, the numerical solution is denoted by solid line while, for comparison, the exact solution from Fig. 4.2 features as a dashed line.

The standard approach to increase accuracy is to use a higher-order method. To this end, the trapezoidal rule is of order 2, while the Runge-Kutta method employed is of order 3. The improvement in accuracy is tangible but neither method gives satisfactory results.

In comparison, the Filon-type method from Fig. 4.5, using just linear interpolation, produces a result which is visually hardly distinguishable from the exact solution.

It should be noted that the waveform relaxation converged fairly rapidly and letting $N=7$ in (4.37) our simulations produced perfectly satisfactory results.


Figure 4.3: The solution for the amplitude-modulated case with the trapezoidal rule.


Figure 4.4: The solution for the amplitude-modulated equation with the third-order explicit Runge-Kutta method.


Figure 4.5: The solution for the amplitude-modulated case using the Filon-type method, combined with waveform relaxation.

The values for the first example shown above are selected to emphasise various qualities of the Filon-method. Here, an example with realistic values will be implemented. $I_{0}=10^{-12} A, C=10^{-6} f, R=100 \Omega$ and $k=1 / 0.0259$. The input signal is the same as the one in the first example, i.e., $b(t)=\sin \omega_{1} \sin \omega_{2} t$, with $\omega_{1}=100$ $\mathrm{rad} / \mathrm{s}$ and $\omega_{2}=10^{7} \mathrm{rad} / \mathrm{s}$. As seen from Fig. 4.6, the results from the Filon-method and exact solution match well.


Figure 4.6: The solution for the amplitude-modulated case with realistic values using the Filon-type method combined with waveform relaxation.

### 4.4.2 Digital Modulation

As an example of digital modulation, we have taken $b(t)=x(t) \cos \omega \pi t$, where $\omega=$ $4 \times 10^{9} \mathrm{rad} / \mathrm{s}$, while $x(t)$ is an alternating sequence of +1 and -1 with a bit period of 133 ns. ${ }^{1}$ This is an example for the Binary Phase Shift Keying modulation technique.

We let $I_{0}=100 A, C=10^{-4} f, R=1 \Omega$ and $k=1$.

[^0]

Figure 4.7: The exact solution of the digitally modulated equation.


Figure 4.8: A very small time sub-interval of Fig. 4.7

The exact solution of the digitally-modulated equation with the above parameters is displayed in Fig. 4.7 and 4.8. Fig. 4.8 zooms in on a small time window to show that the solution exhibits exceedingly rapid, small-amplitude oscillations. Such oscillations are invisible in less detailed plots but, defeat traditional ODE solvers or further decrease the step size.

In Figs. 4.9-4.11, we display the errors committed by three numerical methods, all with a constant step size $h=2.5 \times 10^{-10}$, applied to the digital-modulation equation: the (second-order) trapezoidal rule, the standard explicit, third-order three-stage Runge-Kutta method and the Filon-type method with piecewise-linear approximation and waveform relaxation. The reason for the choice of the minute step size, roughly of the order of magnitude of $O\left(\omega^{-1}\right)$, is that traditional methods require it. Even taking such a tiny step size, it is evident how the numerical solution (solid line) rapidly departs from the exact solution, denoted by a dashed line. All accuracy is lost in even such a short interval. In comparison, the Filon-type method produces an outcome visually indistinguishable from the exact solution. Of course, with Filon-type method, an accurate solution would be obtained with a larger step size: in methods designed using asymptotic principles the size of the step plays a minor role insofar as accuracy is concerned.

This Chapter presents a preliminary study into an alternative technique for numerical integration suitable for systems subjected to high-frequency signals. Numerical examples confirm the theoretical expectations of the significant potential of Filon-type methods.

### 4.5 Comparison with the Envelope Simulation

As stated in Chapter 1, Advanced Design System (ADS) from Angilent Technologies is the EDA tool which is often employed for the simulation of high frequency circuits.


Figure 4.9: The solution of the digital-modulated equation by the trapezoidal rule.


Figure 4.10: The solution of the digital-modulated equation by third-order explicit Runge-Kutta method.


Figure 4.11: The solution of the digital-modulated equation by the Filon-type method, combined with waveform relaxation.

Thus the diode rectifier circuit shown in Fig. 4.1 is simulated with the envelope simulator in ADS in order to compare it in terms of accuracy and efficiency to the Filon-type method.

The Envelope Transient Harmonic Balance technique (ETHB) [37] [38] is used in the envelope simulator. This is a mixed-mode simulation technique which performs the analysis in both the time- and the frequency-domain [1]. Briefly, the RF carrier signal is represented in the frequency domain and the modulation envelope is represented in the time domain.

Consider the fast oscillation signal shown as below:

$$
\begin{equation*}
u(t)=\sum_{k} U_{k}(t) e^{j 2 \pi f_{k} t} \tag{4.38}
\end{equation*}
$$

where $U_{k}(t)$ and $f_{k}$ represent the amplitudes and frequencies of the $k$ th period of the envelope and $t$ is the time scale. Note that $U_{k}(t)$ must vary slowly relative to $f_{k}$ because when the bandwidth of is greater than $f_{k} / 2$ then the sidebands of adjacent harmonics
begin to overlap [39]. If there is an $f_{0}$, which satisfies $f_{k}=K f_{0}$, then $u(t)$ is one tone signal, and the corresponding analysis is a single carrier ETHB envelope simulation. Otherwise, $u(t)$ is a multi-tone signal, and the corresponding analysis is a multi-carrier ETHB envelope simulation. As shown in Fig. 4.12 the output of the ETHB envelope simulation is a time-varying spectrum for time points $t_{1}, t_{2}, \ldots$. The HB analysis is performed at each time points $t_{i}$. Also, the spectrum is converted to time domain waveforms in each clock cycle.


Figure 4.12: A Modulated signal and its simulated time-varying spectrum [1].

The circuit in Section 4.4 is simulated in ADS using the same values for the capacitance, resistance and input signal. The schematic is shown in Appendix D. However, ADS is NOT able to do the simulation with the same time-step stated in Section 4.4.1 and 4.4.2 as used in Filon-type method. The simulation can only be done in ADS when the time-step size is decreased to a suitable value. This means that the Filon-type method is more efficient than the ADS ETHB envelope simulation for the selected system.

### 4.6 Conclusions

We have presented a preliminary study into an alternative techniques for numerical integration of systems subject to high-frequency signals. We applied the basic theory to a rectifier diode circuit. Two cases are considered - when the input signal to the circuit is amplitude modulated and when it is digital modulated. Numerical results confirm the theoretical expectations of the significant potential of Filon-type methods in this setting. The simulation of the frequency-modulated signals will be addressed in future research.

## Chapter 5

## Simulation Techniques to Capture the Locking Range of an Injection-Locked Frequency Divider

### 5.1 Introduction to the Injection-Locked Frequency Divider (ILFD)

Phase-locked loops (PLL) are widely used in wireless communication applications such as frequency synthesis. The frequency divider (FD) is used in the feedback of a frequency synthesizer in order to divide the frequency by a fixed number. Frequency dividers are categorized into two groups, i.e. injection-locked (ILFD) and static [40] frequency dividers. Static dividers offer a wide-bandwidth but with a high power consumption. On the contrary, ILFDs consume less power but at the expense of a narrow locking range. One of the drawbacks to the use of injection locking as a method of frequency division is that there is a limited input bandwidth (locking range) over which frequency division occurs. Therefore, the locking range is one of the most important concerns in designing ILFDs. Hence, an efficient and accurate method for its determination from simulation is essential. Locking occurs when the oscillator locks into or tracks the phase and frequency of an injected signal. The output frequency is either equal to the input frequency or a submultiple or harmonic of it. Hence, the term "fre-
quency divider". In this work, the LC-oscillator based ILFD shall be considered so the first section of this chapter describes it. Then several existing modelling approaches for injection-locked oscillators are presented. The chapter proceeds to review some techniques for determining the locking ranges of oscillators. Finally, some novel approaches and numerical results for determination of locking ranges are presented.

### 5.2 Background to the LC-Oscillator Based ILFD

### 5.2.1 Effect of Negative Resistance on LC Oscillators

The concept of negative resistance is used to understand the operation of an oscillator. Figure 5.1 shows a model of a simple tank circuit.


Figure 5.1: Decaying impulse responses of a tank.

Consider the tank circuit being stimulated by a current impulse. The tank responds with decaying oscillatory behaviour because, in every cycle, some energy is lost in the form of heat dissipated in the resistor $R_{p}$. Now suppose that a resistor equal to $-R_{p}$ is placed in parallel with the tank, then since $R_{p} \|-R_{p}=\infty$, the tank oscillates indefinitely, as illustrated in Fig. 5.2.

Thus, if a circuit exhibiting a negative resistance is placed in parallel with the lossy tank, then the combination may oscillate. Such a topology is called a one-port oscillator. In practice, a negative resistance is provided by using an active circuit, as


Figure 5.2: Addition of negative resistance to cancel tank losses.
shown in Fig. 5.3. The following quote is taken from [41], which remarks on the confusing decision of whether an oscillator is of the negative resistance type or not:
"As a final comment, it should be clear that many (if not all) oscillators may be considered as negative resistance oscillators since, from the point of view of the tank, the active elements cancel the loss due to the finite $Q$ of the resonators. Hence, whether to call an oscillator a "negative resistance" oscillator type is actually more a philosophical decision than anything fundamental."


Figure 5.3: Use of an active circuit to provide negative resistance.

### 5.2.2 LC Oscillator with Cross-Coupled Negative Resistor

Due to their relatively good phase noise performance, ease of implementation, and differential operation, cross-coupled inductance-capacitance (LC) oscillators play an important role in high-frequency circuit design [42]. This topology uses both NMOS


Figure 5.4: Cross-coupled CMOS LC Oscillator.
and PMOS cross-coupled pairs to provide the negative resistance necessary to cancel the losses in the tank circuit.

As shown in Fig. 5.4, two cross-coupled CMOS inverters, consisting of $M 1$ to $M 4$, are used to generate the negative resistance needed to cancel the losses in the tank circuit. The resonant circuit comprising of $R, L$ and $C$ is connected in parallel with the negative resistance. The differential output of the oscillator is available between the nodes labeled $V_{\text {out } 1}$ and $V_{\text {out } 2}$. A current source is used to provide a bias current $I_{o}$ for the transistors.

### 5.3 Overview of the Existing Models of Injection-Locked Oscillators

### 5.3.1 Adler's Model

Adler [2] introduced the mechanism of injection-locking for a small injected signal. Figure 5.5 [43] shows a simple, popular model which can describe the injection-locked oscillator.


Figure 5.5: Adler's model for the injection-locked oscillator.

In this model, the two inputs $v_{i}$ and $v_{o}$ simply add before being operated on by the nonlinearity $f$. The nonlinearity is required both for amplitude stability and to enable frequency mixing. The linear filter $H(j \omega)$ rejects all frequencies far from the frequency of the free-running oscillator $\omega_{o}$.

The expression for the nonlinear device $f$ is derived as a polynomial series, which is:

$$
\begin{equation*}
f\left(v_{i}+v_{o}\right)=\sum_{m=0}^{\infty} a_{m} \cdot\left(v_{i}+v_{o}\right)^{m} \tag{5.1}
\end{equation*}
$$

By setting $v_{o}=V_{o} \cos \left(\omega_{o} t+\varphi\right)$ and $v_{i}=V_{i} \cos \left(\omega_{i} t\right)$, the products of $f\left(v_{i}+v_{o}\right)$ are obtained as:

$$
\begin{align*}
\text { products } & =f\left(v_{i}+v_{o}\right) \\
& =\sum_{m=0}^{\infty} \sum_{n=0}^{\infty} K_{m, n} \cos \left(n \omega_{i} t\right) \cos \left(m \omega_{o} t+n \varphi\right) \tag{5.2}
\end{align*}
$$

For some integers $m$ and $n$, when $\left|m \omega_{o} \pm n \omega_{i}\right|=\omega_{o}$, the corresponding output terms in (5.2) will exist at $\omega_{o}$, the frequency of the free-running oscillator. The output of $f$ has a phase shift with respect to the input. For oscillation, Barkhausen's criterion must be satisfied. Hence, the phase shift contributed by $H(j \omega)$ must adjust so that the net phase around the loop is $2 k \pi$. If the input frequency is too large, then this is not possible and locking fails. The locking range is the set of frequencies for which locking is possible.

### 5.3.2 Miller's Model

The injection-locked oscillator (ILO) can be modeled with a generalized Miller-type model [44] [45], as shown in Fig. 5.6. The filter $H(j \omega)$ is assumed to reject all frequencies except $\omega_{o}$.

As seen from Fig. 5.6, there are two memoryless nonlinear functions, $f\left(v_{o}\right)$ and $g\left(v_{i}\right)$, in this model:

$$
\begin{align*}
& f\left(v_{o}\right)=\sum_{n=0}^{\infty} a_{m} \cdot v_{o}^{m}  \tag{5.3}\\
& g\left(v_{i}\right)=\sum_{n=0}^{\infty} b_{n} \cdot v_{i}^{n} \tag{5.4}
\end{align*}
$$



Figure 5.6: Miller-type model for the injection-locked oscillator

Set $v_{o}=V_{o} \cos \left(\omega_{o} t+\varphi\right)$ and $v_{i}=V_{i} \cos \left(\omega_{i} t\right)$. The output of the mixer is then obtained by a product of Fourier series for $f$ and $g$.

$$
\begin{align*}
\text { product } & =f \cdot g \\
& =\left[\sum_{n=0}^{\infty} B_{n} \cdot \cos \left(n \omega_{i} t\right)\right] \cdot\left[\sum_{m=0}^{\infty} A_{m} \cdot \cos \left(m \omega_{o} t+m \varphi\right)\right], \tag{5.5}
\end{align*}
$$

where the coefficients $B_{n}$ are functions of the input amplitude $V_{i}$ only, while $A_{m}$ are functions of the output amplitude $V_{o}$. We can then determine which products lie at $\omega_{o}$. This model has its limitations when predicting the sub- and superharmonic injection locking. In particular, the expression for the products is not accurate enough because the coefficients $A_{m}$ and $B_{n}$ may be functions of both $V_{i}$ and $V_{o}$ in practice.

### 5.3.3 Verma's Model

Verma [46] introduced a more general model for the injection-locked frequency divider shown in the block diagram in Fig. 5.7.

Here, $f\left(v_{i}, v_{o}\right)$ is assumed as a memoryless nonlinear function of both $v_{i}$ and $v_{o}$. The linear filter $H(j \omega)$ filters out all the frequencies other than $\omega_{o}$. The expression for $f$ is given as:

$$
\begin{equation*}
f\left(v_{i}, v_{o}\right)=\sum_{m=0}^{\infty} a_{m}\left(v_{i}\right) \cdot v_{o}^{m} \tag{5.6}
\end{equation*}
$$



Figure 5.7: General model for the injection-locked frequency divider.

Using a Taylor series expansion of $v_{i}$ around a $d c$ point $V_{d c}, f$ is obtained as:

$$
\begin{equation*}
f\left(v_{i}, v_{o}\right)=\sum_{n=0}^{\infty} \frac{\left(v_{i}-V_{d c}\right)^{n}}{n!} \cdot\left[\left.\sum_{m=0}^{\infty} \frac{\partial^{n}}{\partial v_{i}^{n}} a_{m}\left(v_{i}\right)\right|_{v_{i}=V_{d c}} v_{o}^{m}\right] . \tag{5.7}
\end{equation*}
$$

It is assumed that the magnitude of the injection is weak compared to the static bias point, i.e. $v_{i}$ is close to $V_{d c}$. In this case, only the terms with $n<2$ in Eq. (5.7) are considered.

$$
\begin{equation*}
f\left(v_{i}, V_{o}\right) \cong \sum_{m=0}^{\infty} a_{m}\left(V_{d c}\right) \cdot v_{o}^{m}+\left(v_{i}-V_{d c}\right) \cdot\left[\sum_{m=0}^{\infty} \widetilde{a}_{m}\left(V_{d c}\right) \cdot v_{o}^{m}\right] \tag{5.8}
\end{equation*}
$$

where

$$
\begin{equation*}
\tilde{a}_{m}\left(V_{d c}\right)=\left.\frac{\partial}{\partial_{v_{i}}} a_{m}\left(v_{i}\right)\right|_{v_{i}=V_{d c}} . \tag{5.9}
\end{equation*}
$$

The coefficients $a_{m}$ and their derivatives can be determined either from the analytical form of $f$ or extracted by measuring the effect of slight perturbations on the nonlinearity about the bias point $V_{d c}$.

If both $v_{i}$ and $v_{o}$ are sinusoidal and of the form $v_{i}=V_{d c}+V_{i} \cos \left(\omega_{i} t\right)$ and $v_{o}=$ $V_{o} \cos \left(\omega_{o} t+\varphi\right)$, the full output of (5.7) can be written as:

$$
\begin{equation*}
f\left(v_{i}, v_{o}\right)=\sum_{n=0}^{\infty} \frac{\left(V_{i} \cos \left(\omega_{i} t\right)\right)^{n}}{n!} \cdot\left[\left.\sum_{m=0}^{\infty} \frac{\partial^{n}}{\partial v_{i}^{n}} a_{m}\right|_{v_{i}=V_{d c}}\left(V_{o} \cos \left(\omega_{o} t+\varphi\right)\right)^{m}\right] . \tag{5.10}
\end{equation*}
$$

The terms may be regrouped to express the bracketed quantity in (5.10) as a sum of harmonics of $\omega_{o}$

$$
\begin{equation*}
f\left(v_{i}, v_{o}\right)=\sum_{n=0}^{\infty} \frac{\left(V_{i} \cos \left(\omega_{i} t\right)\right)^{n}}{n!} \cdot\left[\left.\sum_{m=0}^{\infty} \frac{\partial^{n}}{\partial v_{i}^{n}} A_{m}\right|_{v_{i}=V_{d c}} \cos \left(m \omega_{o} t+m \varphi\right)\right] \tag{5.11}
\end{equation*}
$$

where the coefficient $A_{m}$ is a function of all of the $a_{m}$ and $V_{o}$. Assuming weak injection and simplifying ( $n<2$ )

$$
\begin{equation*}
f\left(v_{i}, v_{o}\right)=\sum_{m=0}^{\infty} A_{m} \cos \left(m \omega_{o} t+m \varphi\right)+\frac{1}{2} \sum_{m=0}^{\infty} V_{i} \tilde{A}_{m} \cos \left[\left(m \omega_{o} \pm \omega_{i}\right) t+m \varphi\right] \tag{5.12}
\end{equation*}
$$

where

$$
\begin{equation*}
\tilde{A}_{m}=\left.\frac{\partial A_{m}}{\partial v_{i}}\right|_{v_{i}=V_{d c}} \tag{5.13}
\end{equation*}
$$

Note that the first term in (5.12) represents the nonlinearity for a free-running oscillator while the second term shows the mixer products due to the presence of the injected signal.

### 5.3.4 Xu's Model

Xu's model [7] is a specialised model for an LC-oscillator based injection-locked frequency divider (LC-ILFD). This model proves that the LC-ILFD behaves like a PhaseLocked Loop (PLL) when it is locked.

The circuit schematic of the LC-ILFD under consideration has been described in [47], [48], and is shown in Fig. 5.8.

It can be simplified to the ideal circuit shown in Fig. 5.9. $N_{\mathrm{R}}$ represents the nonlinear resistor comprising the components in Fig. 5.8 other than $R_{\mathrm{s}}, L$ and $C$.

When there is no injected signal, the state equations that describe this circuit are:

$$
\begin{align*}
C \frac{d V_{\mathrm{C}}}{d t} & =I_{\mathrm{L}}-f\left(V_{\mathrm{C}}\right)  \tag{5.14}\\
L \frac{d I_{\mathrm{L}}}{d t} & =-I_{\mathrm{L}} R_{\mathrm{S}}-V_{\mathrm{C}} \tag{5.15}
\end{align*}
$$



Figure 5.8: Circuit schematic.


Figure 5.9: The simplified circuit.
where $f\left(V_{\mathrm{C}}\right)$ is the driving point characteristic of the non-linear resistor which can be obtained by experiment and/or by Spice simulation. The schematic used in PSpice to capture the driving point characteristics is illustrated in Fig. 5.10. The RLC tank is replaced by a DC source, whose value is increased from -9 V to 9 V during the simulation.


Figure 5.10: The schematic used in PSpice to capture the driving point characteristics.

The driving-point characteristics derived from Spice simulations and experiments are shown in Fig. 5.11, together with a cubic approximation. The cubic approximation is:

$$
\begin{equation*}
f\left(V_{\mathrm{C}}\right)=a V_{\mathrm{C}}+b V_{\mathrm{C}}^{2}+c V_{\mathrm{C}}^{2} \tag{5.16}
\end{equation*}
$$

where

$$
\begin{equation*}
c=-\frac{a}{V_{\mathrm{DD}}^{2}} \tag{5.17}
\end{equation*}
$$

and the parameter $a\left(V_{\mathrm{GS}}\right)$ is a function of the bias voltage of the tail transistor $M_{6}$, shown in Fig. 5.8. Note that the second harmonic term $b V_{\mathrm{C}}^{2}$ is so small as to be ignored.


Figure 5.11: Driving point characteristics: Solid: Experimental; Dotted: Cubic Fit To Experiment; Dashed: PSpice.

Thus the cubic approximation (5.16) is expressed in the form:

$$
\begin{equation*}
f\left(V_{\mathrm{C}}\right)=a V_{\mathrm{C}}-\frac{a}{V_{\mathrm{DD}}^{2}} V_{\mathrm{C}}^{3} \tag{5.18}
\end{equation*}
$$

This free-running oscillator can be viewed as a feedback system; its model is shown in Fig. 5.12. In this model, the RLC tank is represented by its transfer function $H(j \omega)$.

When a signal is injected into the circuit shown in Fig. 5.8, the parameter $a$ is modified to:

$$
\begin{equation*}
a\left(V_{\mathrm{GS}}+v_{\mathrm{GS}}\right) \approx \hat{a}\left(V_{\mathrm{GS}}\right)+\left.\frac{d \hat{a}}{d V_{\mathrm{GS}}}\right|_{V_{\mathrm{GS}}} v_{\mathrm{GS}} \tag{5.19}
\end{equation*}
$$

where $V_{\mathrm{GS}}$ is the bias voltage of $M_{6}$.
By varying $V_{G S_{M 6}}$, a number of different driving point characteristics are found and thus a number of different cubic expressions for $f\left(V_{\mathrm{C}}\right)$ [48]. Then several values for $A$ are extracted and plotted as shown in Fig. 5.13. The slope around the bias point, $\frac{d \hat{a}}{d V_{\mathrm{GS}}}$, can be estimated analytically after fitting a quadratic to the $I_{D_{M 6}}-V_{G S_{M 6}}$ characteristic.


Figure 5.12: Model of the free-running oscillator.


Figure 5.13: $I_{D}$ against $V_{G S}$.

Substituting (5.19) into (5.18), the governing equations for the ILFD are obtained as:

$$
\begin{align*}
C \frac{d V_{\mathrm{C}}}{d t} & =I_{\mathrm{L}}-\left(A+d a V_{\mathrm{in}}\right) V_{\mathrm{C}}+\frac{A+d a V_{\mathrm{in}}}{V_{\mathrm{DD}}^{2}} V_{\mathrm{C}}^{3}  \tag{5.20}\\
L \frac{d I_{\mathrm{L}}}{d t} & =-I_{\mathrm{L}} R_{\mathrm{S}}-V_{\mathrm{C}} \tag{5.21}
\end{align*}
$$

where $A$ and $d a$ are used to represent $\hat{a}\left(V_{\mathrm{GS}}\right)$ and $\left.\frac{d \hat{a}}{d V_{\mathrm{GS}}}\right|_{V_{\mathrm{GS}}}$ in (5.19), respectively.
Equation (5.20) may be rearranged by collecting the factors $A$ and $d a V_{\text {in }}$, to obtain:

$$
\begin{equation*}
C \frac{d V_{\mathrm{C}}}{d t}=I_{\mathrm{L}}-A\left(V_{\mathrm{C}}-\frac{V_{\mathrm{C}}^{3}}{V_{\mathrm{DD}}^{2}}\right)-d a V_{\mathrm{in}}\left(V_{\mathrm{C}}-\frac{V_{\mathrm{C}}^{3}}{V_{\mathrm{DD}}^{2}}\right) . \tag{5.22}
\end{equation*}
$$

The difference between (5.14) and (5.22) is the additional feedback terms $-d a V_{\mathrm{in}}\left(V_{\mathrm{C}}-\frac{V_{C}^{3}}{V_{\mathrm{DD}}^{2}}\right)$. The injection-locked system described by (5.22) can be modeled as shown in Fig. 5.14.

As evident from Fig. 5.14, the negative feedback is composed of two loops in this case. Both paths pass through the nonlinear device but the inner loop has a gain of $A$


Figure 5.14: The loop enhanced model of the ILFD.

### 5.3 Overview of the Existing Models of Injection-Locked Oscillators

and the outer loop has a gain of $d a$. The nonlinear device is governed by the expression:

$$
\begin{equation*}
v_{\mathrm{nd}}=1-\frac{v_{\mathrm{o}}^{2}}{V_{\mathrm{DD}}^{2}} . \tag{5.23}
\end{equation*}
$$

The input signal, $v_{\mathrm{in}}$, is first multiplied by the output of the outer loop to form $I_{\mathrm{inj}}$. $I_{\mathrm{inj}}$ then combines with the output of the inner loop and the output is passed through the RLC tank. This model is specific to an LC-ILFD unlike previous models [46]. The RLC tank in the oscillator can be expressed as:

$$
\begin{equation*}
H(j \omega)=\frac{R_{\mathrm{S}}+j \omega L}{1+j \omega R_{\mathrm{S}} C+(j \omega)^{2} L C} . \tag{5.24}
\end{equation*}
$$



Figure 5.15: Analog Multiplier as a phase detector.
As seen in Fig. 5.15, if the two inputs of the analog multiplier are $A \cos \omega t$ and $B \cos (\omega t+\phi)$, the output can be expressed as

$$
\begin{equation*}
A B \cos (\omega t) \cos (\omega t+\phi)=\frac{A B}{2}(\cos (\phi)-\cos (2 \omega t+\phi)) . \tag{5.25}
\end{equation*}
$$

The average value of the output is

$$
\begin{equation*}
\overline{A B \cos (\omega t) \cos (\omega t+\phi)}=\frac{A B}{2} \cos (\phi) . \tag{5.26}
\end{equation*}
$$

Then the phase difference between the two input signals can be obtained. Thus an analog multiplier acts as a phase detector [41]. It can be concluded that the injectionlocked frequency divider is equivalent from a mathematical perspective to a dualloop phase-locked-loop (PLL) [7], when the injected frequency approximates the freerunning oscillator frequency.

### 5.4 Existing Methods to Obtain the Locking Range of Oscillators

### 5.4.1 Adler Approach

This method was described in [2]. It is assumed that the free-running oscillation gives an angular frequency $\omega_{0}$. The injected signal has the frequency $\omega_{1}$ which forces a beat frequency:

$$
\begin{equation*}
\Delta \omega_{0}=\omega_{0}-\omega_{1} . \tag{5.27}
\end{equation*}
$$

Because of the effect of the injected signal, the output frequency changes to $\omega$, and the instantaneous beat frequency is termed as:

$$
\begin{align*}
\Delta \omega & =\frac{d \alpha}{d t}  \tag{5.28}\\
& =\omega-\omega_{1} \tag{5.29}
\end{align*}
$$

where as shown in Fig. 5.16, $\alpha$ is the angle between the input and output signal in the excited system.

As seen in Fig. 5.16 and Fig. 5.17, $E_{1}$ and $E_{g}$ are the amplitude of the input and output signals in the excited system, $E$ is the amplitude of the output signal of the free-running circuit. If the amplitude of the injected signal is much smaller than that of the free-running output signal, i.e. $E_{1} \ll E$, then:

$$
\begin{equation*}
\varphi=\frac{E_{1} \sin (-\alpha)}{E}=-\frac{E_{1}}{E} \sin \alpha \tag{5.30}
\end{equation*}
$$



Figure 5.16: Vector diagram of the instantaneous voltages.


Figure 5.17: Sample oscillator circuit [2].


Figure 5.18: Phase versus frequency for a simple tuned circuit.
Figure 5.18 shows the $\varphi$ versus $\omega$ curve. The slope of it is

$$
\begin{equation*}
A=\frac{d \varphi}{d \omega} . \tag{5.31}
\end{equation*}
$$

The angle between the free-running output signal and the resultant signal can be calculated as:

$$
\begin{equation*}
\varphi=A\left(\omega-\omega_{0}\right) . \tag{5.32}
\end{equation*}
$$

Now rearrange (5.32) and substitute (5.27) and (5.29):

$$
\begin{align*}
\varphi & =A\left(\omega-\omega_{0}\right) \\
& =A\left[\left(\omega-\omega_{1}\right)-\left(\omega_{0}-\omega_{1}\right)\right] \\
& =A\left(\Delta \omega-\Delta \omega_{0}\right) . \tag{5.33}
\end{align*}
$$

After substituting (5.30) on the left side and (5.28) on the right side,

$$
\begin{equation*}
-\frac{E_{1}}{E} \sin \alpha=A\left(\frac{d \alpha}{d t}-\Delta \omega_{0}\right) . \tag{5.34}
\end{equation*}
$$

Then

$$
\begin{equation*}
\frac{d \alpha}{d t}=-\frac{E_{1}}{E} \cdot \frac{1}{A} \sin \alpha+\Delta \omega_{0} \tag{5.35}
\end{equation*}
$$

The quality factor $Q$ is defined as in [41]:

$$
\begin{equation*}
Q=\frac{\varphi}{2} \cdot \frac{\omega_{0}}{\omega-\omega_{0}} . \tag{5.36}
\end{equation*}
$$

Merging (5.32) and (5.36),

$$
\begin{equation*}
A=\frac{2 Q}{\omega_{0}} \tag{5.37}
\end{equation*}
$$

Substituting (5.37) into (5.35), yields:

$$
\begin{equation*}
\frac{d \alpha}{d t}=-\frac{E_{1}}{E} \cdot \frac{\omega_{0}}{2 Q} \sin \alpha+\Delta \omega_{0} \tag{5.38}
\end{equation*}
$$

If the system is in steady state, $d \alpha / d t$ must be zero. Hence,

$$
\begin{equation*}
0=-\frac{E_{1}}{E} \cdot \frac{\omega_{0}}{2 Q} \sin \alpha+\Delta \omega_{0} \tag{5.39}
\end{equation*}
$$

or

$$
\begin{equation*}
\sin \alpha=2 Q \frac{E}{E_{1}} \cdot \frac{\Delta \omega_{0}}{\omega_{0}} \tag{5.40}
\end{equation*}
$$

Because $\sin \alpha$ can only assume values between +1 and -1 , the right side of (5.40) must be within this range.

$$
\begin{equation*}
\left|2 Q \frac{E}{E_{1}} \cdot \frac{\Delta \omega_{0}}{\omega_{0}}\right| \leq 1 \tag{5.41}
\end{equation*}
$$

So the locking range is

$$
\begin{equation*}
\left|\Delta \omega_{0}\right| \leq \frac{E_{1}}{E} \cdot \frac{\omega_{0}}{2 Q} \tag{5.42}
\end{equation*}
$$

For an RLC tank, the $Q$ factor may be expressed as:

$$
\begin{equation*}
Q=R \sqrt{\frac{C}{L}} \tag{5.43}
\end{equation*}
$$

Hence, in an RLC circuit, the locking range is:

$$
\begin{equation*}
\left|\Delta \omega_{0}\right| \leq \frac{E_{1}}{E} \cdot \frac{\omega_{0}}{2 R} \cdot \sqrt{\frac{L}{C}} \tag{5.44}
\end{equation*}
$$

### 5.4.2 Graphical Analysis of the Phase Characteristic

In this second method, the locking range is determined graphically, as described in [10]. For example, consider Fig. 5.19, the frequency of the free-running oscillator is defined as $\omega_{0}=1 / \sqrt{L_{1} C_{1}}$. (Neglecting all of the parasitics). As illustrated in Fig. 5.19(b), when a phase shift $\phi_{0}$ is inserted into the system, the oscillator no longer oscillates at the frequency $\omega_{0}$. Thus, as shown in Fig. 5.19(c), the output frequency deviates from $\omega_{0}$ to a new value $\omega_{1}$ so as to cancel the effect of $\phi_{0}$ and satisfy Barkhausen's criterion. If a current source $I_{i n j}$ is added to the drain of $M_{1}$, as shown in Fig. 5.19(d), and if the frequency of $I_{i n j}$ is within the locking range, the output frequency of the system will be at the injected frequency $\omega_{i n j}$. The oscillator is then injection-locked.


Figure 5.19: (a)Conceptual oscillator.(b) Frequency shift due to additional phase shift. (c) Open-loop characteristics.(d)Frequency shift by injection.


Figure 5.20: Phasor diagram.

In order to determine the locking range, the phasor diagram for the oscillator subject to an external signal is shown in Fig. 5.20. $\phi_{0}$ is the angle between the free-running output signal $I_{o s c}$ and the resultant output signal of the injected system $I_{T}$. $\theta$ denotes the phase difference between $I_{o s c}$ and $I_{i n j}$. The relationship between them can be expressed as:

$$
\begin{align*}
\sin \phi_{0} & =\frac{I_{i n j}}{I_{T}} \sin \theta  \tag{5.45}\\
& =\frac{I_{i n j} \sin \theta}{\sqrt{I_{o s c}^{2}+I_{i n j}^{2}+2 I_{o s c} I_{i n j} \cos \theta}} \tag{5.46}
\end{align*}
$$

The maximum value of $\sin \phi_{0}$ is

$$
\begin{equation*}
\sin \phi_{0, \max }=\frac{I_{i n j}}{I_{o s c}} \tag{5.47}
\end{equation*}
$$

when the condition

$$
\begin{equation*}
\cos \theta=-\frac{I_{i n j}}{I_{o s c}} \tag{5.48}
\end{equation*}
$$

is achieved.
As shown in Fig. 5.21, the angle between $I_{T}$ and $I_{i n j}$ has to be $\pi / 2$ so as to satisfy the condition (5.48). Hence, the maximum angle between $I_{i n j}$ and $I_{o s c}$ is $\pi / 2+\phi_{0}$.


Figure 5.21: Phasor diagram with maximum angle between $I_{i n j}$ and $I_{T}$.
A second-order parallel tank consisting of $\mathrm{L}, \mathrm{C}$, and $R_{P}$ exhibits a phase shift of

$$
\begin{equation*}
\alpha=\frac{\pi}{2}-\tan ^{-1}\left(\frac{L \omega}{R_{P}} \cdot \frac{\omega_{0}^{2}}{\omega_{0}^{2}-\omega^{2}}\right) . \tag{5.49}
\end{equation*}
$$

Since $\omega_{0}^{2}-\omega^{2} \approx 2 \omega_{0}\left(\omega_{0}-\omega\right), L \omega / R_{P}=1 / Q$, and $\pi / 2-\tan ^{-1} x=\tan ^{-1}\left(x^{-1}\right)$,

$$
\begin{equation*}
\tan \alpha \approx \frac{2 Q}{\omega_{0}}\left(\omega_{0}-\omega\right) \tag{5.50}
\end{equation*}
$$

From Fig. 5.21, $\tan \phi_{0}=I_{i n j} / I_{T}$ and $I_{T}=\sqrt{I_{o s c}^{2}-I_{i n j}^{2}}$. It follows that

$$
\begin{equation*}
\omega_{0}-\omega_{i n j}=\frac{\omega_{0}}{2 Q} \cdot \frac{I_{i n j}}{I_{o s c}} \cdot \sqrt{\frac{1}{1-\frac{I_{i j j}^{2}}{I_{o s c}^{2}}}} \tag{5.51}
\end{equation*}
$$

If the maximum value of $\omega_{0}-\omega_{i n j}$ is set as $\omega_{L}$, the overall locking range is obtained as $\pm \omega_{L}$ around $\omega_{0}$. Here $\omega_{L}$ is "one-sided" locking range.

Note that the locking range depends on the amplitude of the injected signal, $I_{i n j}$. If $I_{i n j}$ increases, the phase difference between $I_{o s c}$ and $I_{i n j}$ must become greater so as to maintain the angle between $I_{T}$ and $I_{o s c}$ at $\phi_{0}$.

### 5.4.3 Analysis with Harmonic Balance

In this section, the locking range is obtained by analysing an ILFD using Harmonic Balance [8]. The general model of an ILFD is shown in Fig 5.22. $f\left(v_{i n j}, v_{o}\right)$ represents


Figure 5.22: General Circuit Model for ILFD.
the nonlinear resistor while $H(j \omega)$ is the RLC tank.
The first step is to define the input and output voltages of the ILFD [49]:

$$
\begin{gather*}
v_{i n j}=\frac{V_{i n j}}{2}\left[e^{j\left(\omega_{\mathrm{inj} j} t+\phi\right)}+e^{-j\left(\omega_{\mathrm{inj} \mathrm{j}} t+\phi\right)}\right]  \tag{5.52}\\
v_{o}=\frac{V_{o}}{2}\left[e^{j \omega_{\mathrm{out}} t}+e^{-j \omega_{\mathrm{out}} t}\right] \tag{5.53}
\end{gather*}
$$

A third-order approximation is derived for the nonlinear block $f\left(v_{i n j}, v_{o}\right)$

$$
\begin{equation*}
f_{\left(v_{i n j}, v_{o}\right)}=a\left(v_{i n j}\right) v_{o}+b\left(v_{i n j}\right) v_{o}^{2}+c\left(v_{i n j}\right) v_{o}^{3} \tag{5.54}
\end{equation*}
$$

where the parameters $a, b$ and $c$ are the functions of $v_{i n j}$ which are shown as below:

$$
\begin{align*}
& a\left(v_{i n j}\right)=A+d a v_{i n j}, \\
& b\left(v_{i n j}\right)=B+d b v_{i n j},  \tag{5.55}\\
& c\left(v_{i n j}\right)=C+d c v_{i n j} .
\end{align*}
$$

Then by substituting equations (5.52), (5.53) and (5.55) into equation (5.54), it is
obtained:

$$
\begin{align*}
f_{o}\left(v_{i n j}, v_{o}\right)= & \left(A+\mathrm{d} a v_{\mathrm{inj}}\right) v_{o}+\left(B+\mathrm{d} b v_{\mathrm{inj}}\right) v_{o}^{2}+\left(C+\mathrm{d} c v_{\text {inj }}\right) v_{o}^{3} \\
= & \left(A+\mathrm{d} a \frac{V_{\text {inj }}}{2}\left[e^{j\left(\omega_{\mathrm{inj}} t+\phi\right)}+e^{-j\left(\omega_{\mathrm{inj}} t+\phi\right)}\right]\right) \times \frac{V_{o}}{2}\left[e^{j \omega_{\mathrm{out}} t}+e^{-j \omega_{\mathrm{out}} t}\right] \\
& +\left(B+\mathrm{d} b \frac{V_{\text {inj }}}{2}\left[e^{j\left(\omega_{\mathrm{inj}} t+\phi\right)}+e^{-j\left(\omega_{\mathrm{inj}} t+\phi\right)}\right]\right) \\
& \times \frac{V_{o}^{2}}{4}\left[e^{2 j \omega_{\mathrm{out}} t}+2+e^{-2 j \omega_{\mathrm{out}} t}\right] \\
& +\left(C+\mathrm{d} c \frac{V_{\text {inj }}}{2}\left[e^{j\left(\omega_{\mathrm{inj}} t+\phi\right)}+e^{-j\left(\omega_{\mathrm{inj}} t+\phi\right)}\right]\right) \\
& \times \frac{V_{o}^{3}}{8}\left[e^{3 j \omega_{\text {out }} t}+3 e^{j \omega_{\text {out }} t}+3 e^{-j \omega_{\mathrm{out}} t}+e^{-3 j \omega_{\text {out }} t}\right] . \tag{5.56}
\end{align*}
$$

From the many terms generated in (5.56), the terms of interest are those that fall close to $\omega_{\text {out }}$ as the resonator $H(j \omega)$ attenuates the components far from $\omega_{\text {out }}$. The bandpass filter response can be expressed as [50]

$$
\begin{equation*}
H\left(j \omega_{0}\right)=\frac{R_{L} H_{0}}{1+j 2 Q \frac{\delta \omega}{\omega_{0}}} \tag{5.57}
\end{equation*}
$$

where $\omega_{0}$ is the free-running frequency, $\delta \omega$ is the one-sided locking range defined as

$$
\begin{equation*}
\delta \omega=\omega_{\text {out }}-\omega_{0} \tag{5.58}
\end{equation*}
$$

$Q$ is the quality factor of the free-running $L C$ tank, which is given by

$$
\begin{equation*}
Q=\frac{\omega_{0} L}{R_{L}} \tag{5.59}
\end{equation*}
$$

$H_{0}$ is the coefficient to make the term $R_{L} H_{0}$ represent the net resistance across the $L C$ tank. It can be found by using impedance transformation:

$$
\begin{equation*}
H_{0}=Q^{2}+1 \tag{5.60}
\end{equation*}
$$

The following components at $\omega_{\text {out }}$ are generated. From $a v_{o}$ :

$$
\begin{equation*}
\frac{A V_{o}}{2}\left[e^{j\left(\omega_{\mathrm{out}} t\right)}+e^{-j\left(\omega_{\text {out }} t\right)}\right] \tag{5.61}
\end{equation*}
$$

From $b v_{o}^{2}$ :

$$
\begin{equation*}
\frac{\mathrm{d} b V_{i n j} V_{o}^{2}}{8}\left(2 e^{j\left(\omega_{\mathrm{inj}} t+\phi\right)}+e^{j\left(-\omega_{\mathrm{inj}} t-\phi+2 \omega_{\mathrm{out}} t\right)}+2 e^{-j\left(\omega_{\mathrm{inj} j} t+\phi\right)}+e^{j\left(\omega_{\mathrm{inj}} t+\phi-2 \omega_{\mathrm{out}} t\right)}\right) \tag{5.62}
\end{equation*}
$$

From $c v_{o}^{3}$ :

$$
\begin{equation*}
\frac{3 C V_{o}^{3}}{8}\left(e^{j \omega_{\mathrm{out}} t}+e^{-j \omega_{\mathrm{out}} t}\right) \tag{5.63}
\end{equation*}
$$

Next, these components pass through the resonator back to the input, which results in

$$
\begin{array}{r}
\frac{R_{L} H_{o}}{1+j\left(\frac{2 Q \delta \omega}{\omega_{0}}\right)}\left[\frac{A V_{o}}{2} e^{j \omega_{\text {out }} t}+\mathrm{d} b \frac{V_{\text {inj }} V_{\text {out }}^{2}}{8}\left(2 e^{j\left(\omega_{i n j} t+\phi\right)}+e^{j\left(2 \omega_{\text {out }} t-\omega_{\text {inj }} t-\phi\right)}\right)+\right. \\
\left.+\frac{3 C V_{o}^{3}}{8} e^{j \omega_{\text {out }} t}\right]=\frac{V_{o}}{2} e^{j \omega_{\text {out }} t} \tag{5.64}
\end{array}
$$

and

$$
\begin{array}{r}
\frac{R_{L} H_{o}}{1+j\left(\frac{2 Q \delta \omega_{0}}{\omega}\right)}\left[\frac{A V_{o}}{2} e^{-j \omega_{\mathrm{out}} t}+\mathrm{d} b \frac{V_{i n j} V_{o}^{2}}{8}\left(2 e^{-j\left(\omega_{\mathrm{in} \mathrm{j}} t+\phi\right)}+e^{j\left(-2 \omega_{\mathrm{out}} t+\omega_{\text {in } \mathrm{i}} t+\phi\right)}\right)+\right. \\
\left.\quad+\frac{3 C V_{o}^{3}}{8} e^{-j \omega_{\mathrm{out}} t}\right]=\frac{V_{o}}{2} e^{-j \omega_{\text {out }} t}(5.65)
\end{array}
$$

where $\frac{R_{L} H_{o}}{1+j\left(\frac{2 \chi^{\circ} \omega}{\left(\omega_{0}\right)}\right.}$ models the frequency selective block (an RLC tank in this analysis).
Then $e^{j \omega_{\text {out }} t}$ is eliminated from both sides of Eq. (5.64) ${ }^{1}$ and one gets:

$$
\begin{equation*}
\frac{R_{L} H_{o}}{1+j\left(\frac{2 Q \delta \omega}{\omega_{0}}\right)}\left[\frac{A V_{o}}{2}+\mathrm{d} b \frac{V_{\text {inj }} V_{\text {out }}^{2}}{4} e^{j \phi}+\mathrm{d} b \frac{V_{\text {inj }} V_{\text {out }}^{2}}{8} e^{-j \phi}+\frac{3 C V_{\text {out }}^{3}}{8}\right]=\frac{V_{o}}{2} \tag{5.66}
\end{equation*}
$$

Since elimination of the exponential terms on both sides of Eq. (5.64) and Eq. (5.65) gives the same result, in the following analysis, only the terms $e^{j \omega_{\text {out }} t}$ are considered.
$e^{j \phi}$ is expanded and the top and bottom of the left-hand side of the equation are multiplied by the quantity $\left[1-j 2 Q \frac{\delta \omega}{\omega_{0}}\right]$.

The resulting equations for the real and imaginary are as follows:
Real Part:

$$
\begin{equation*}
\left(\frac{R_{L} H_{o}}{1+\left(\frac{2 Q \delta \omega}{\omega_{0}}\right)^{2}}\right) \times\left[\frac{A V_{o}}{2}+\frac{3 C V_{o}^{3}}{8}+\frac{\mathrm{d} b V_{i n j} V_{o}^{2}}{8}\left(3 \cos \phi+\sin \phi\left(\frac{2 Q \delta \omega}{\omega_{0}}\right)\right)\right]=\frac{V_{o}}{2} \tag{5.67}
\end{equation*}
$$

[^1]Imaginary Part:

$$
\begin{align*}
\left(\frac{R_{L} H_{o}}{1+\left(\frac{2 Q \delta \omega}{\omega_{0}}\right)^{2}}\right)\left(\frac{-2 Q \delta \omega}{\omega_{0}}\right) \times & {\left[\frac{A V_{o}}{2}+\frac{3 C V_{o}^{3}}{8}+\frac{3 \mathrm{~d} b V_{i n j} V_{o}^{2}}{8} \cos \phi\right]+} \\
& +\left(\frac{R_{L} H_{o}}{1+\left(\frac{2 Q \delta \omega}{\omega_{0}}\right)^{2}}\right) \frac{\mathrm{d} b V_{i n j} V_{o}^{2}}{8} \sin \phi=0 \tag{5.68}
\end{align*}
$$

The real and imaginary equations may be solved simultaneously to obtain the output voltage $V_{o}$ and the locking range $\delta \omega$. Using the equation for the imaginary part, we can get the normalized locking range as follows:

$$
\begin{equation*}
\frac{2 Q \delta \omega}{\omega_{0}}=\frac{\mathrm{d} b V_{i n j} V_{o} \sin \phi}{4 A+3 C V_{o}^{2}+3 \mathrm{~d} b V_{i n j} V_{o} \cos \phi} \tag{5.69}
\end{equation*}
$$

When $\phi=90^{\circ}$, the maximum normalized locking range is:

$$
\begin{equation*}
\frac{2 Q \delta \omega}{\omega_{0}}=\frac{d b V_{i n j} V_{o}}{4 A+3 C V_{o}^{2}} \tag{5.70}
\end{equation*}
$$

From this equation, it can be seen clearly that the locking range can be enlarged by increasing $V_{i n j}$ and $V_{o}$.

Following a similar method, the locking range for other division ratios can be calculated as well. For example, if the division ratio is $2, e^{j \omega_{\text {inj }} t}=e^{2 j \omega_{\text {out }} t}$.

### 5.5 Novel Use of the Warped Multitime Partial Differential Equation (WaMPDE) to Determine the Locking Range

### 5.5.1 The Use of the WaMPDE for Modelling the Injection-Locked Oscillator

Consider a fairly general nonlinear circuit which is described by:

$$
\begin{equation*}
\dot{x}(t)=f(x(t))+b(t) \tag{5.71}
\end{equation*}
$$

### 5.5 Novel Use of the Warped Multitime Partial Differential Equation (WaMPDE) to Determine the Locking Range

where $b(t)$ is the excitation vector, $x(t)$ are the state variables and $f$ is a nonlinear function. Note that (5.71) follows from (2.1) except that the input signal $b(t)$ is separated out.

The ( $\mathrm{p}+1$ )-dimensional multivariate variables are defined as:

$$
\begin{equation*}
x(t)=\hat{x}\left(\tau_{1}, \tau_{2}, \ldots, \tau_{p}, t\right) \tag{5.72}
\end{equation*}
$$

where the different time scales are assumed as the functions of real time:

$$
\begin{gather*}
\tau_{1}=\tau_{1}(t) \\
\tau_{2}=\tau_{2}(t)  \tag{5.73}\\
\ldots \\
\tau_{p}=\tau_{p}(t) .
\end{gather*}
$$

Differentiating $x$ with respect to $t$, one gets:

$$
\begin{equation*}
\frac{d x}{d t}=\frac{d \hat{x}}{d \tau_{1}} \frac{d \tau_{1}}{d t}+\frac{d \hat{x}}{d \tau_{2}} \frac{d \tau_{2}}{d t}+\cdots+\frac{d \hat{x}}{d \tau_{p}} \frac{d \tau_{p}}{d t}+\frac{d \hat{x}}{d t} . \tag{5.74}
\end{equation*}
$$

Then the warped frequencies are introduced:

$$
\begin{equation*}
\omega_{i}=\frac{d \tau_{i}}{d t} \tag{5.75}
\end{equation*}
$$

where $i \in\{1,2, \ldots, p\}$.
Hence, (5.74) is rewrited to

$$
\begin{equation*}
\frac{d x}{d t}=\omega_{1}(t) \frac{d \hat{x}}{d \tau_{1}}+\omega_{2}(t) \frac{d \hat{x}}{d \tau_{2}}+\cdots+\omega_{p}(t) \frac{d \hat{x}}{d \tau_{p}}+\frac{d \hat{x}}{d t} . \tag{5.76}
\end{equation*}
$$

Then the $(p+1)$-dimensional WaMPDE [11] corresponding to (5.71) is:

$$
\begin{equation*}
\sum_{i=1}^{p}\left(\omega_{i}(t) \frac{\partial \hat{x}}{\partial \tau_{i}}\right)+\frac{\partial \hat{x}}{\partial t}=f(\hat{x})+\hat{b}\left(\tau_{1}, \ldots, \tau_{p}, t\right) \tag{5.77}
\end{equation*}
$$

where $\tau_{1}, \ldots, \tau_{p}$ correspond to the warped time scales and $t$ is the real time-scale. $\hat{x}$ and $\hat{b}$ are multivariate functions of the $p+1$ time variables. Once the solution of (5.77) is found, (5.71) can be solved from (5.78) and (5.79).

### 5.5 Novel Use of the Warped Multitime Partial Differential Equation (WaMPDE) to Determine the Locking Range

Equation (5.72) can be represented in the form:

$$
\begin{equation*}
x(t)=\hat{x}\left(\phi_{1}(t), \phi_{2}(t), \ldots, \phi_{p}(t), t\right) \tag{5.78}
\end{equation*}
$$

where

$$
\begin{equation*}
\phi_{i}(t)=\int_{0}^{t} \omega_{i}\left(\tau_{i}\right) d \tau_{i} \tag{5.79}
\end{equation*}
$$

The relationship between $\tau_{1}, t$ and $\phi_{1}(t)$ for the case with $p=1$ is shown in Fig. 5.23.


Figure 5.23: Relation between $\tau_{1}, t$ and $\phi_{1}(t)$.

In the case of the ILFD, three dimensions are required. The equation becomes:

$$
\begin{equation*}
\omega_{\mathrm{o}}(t) \frac{\partial \hat{x}}{\partial \tau_{1}}+\omega_{\mathrm{inj}} \frac{\partial \hat{x}}{\partial \tau_{2}}+\frac{\partial \hat{x}}{\partial t}=f(\hat{x})+b(t) \tag{5.80}
\end{equation*}
$$

where $\tau_{1}$ and $\tau_{2}$ are the free-running oscillation time scale and the injected signal time scale, respectively, while $t$ denotes the real time. Note that both $\tau_{1}$ and $\tau_{2}$ are warped time scales to enable the slow variation of the local and injected frequency. $\omega_{\mathrm{o}}$ is output frequency of the oscillator and $\omega_{\mathrm{inj}}$ is the injected frequency.

The five-point centered difference formula [51] is used for the derivatives with respect to the warped time variables.

$$
\begin{equation*}
\frac{\partial \hat{x}_{k}}{\partial \tau_{i}} \approx \frac{x_{k-2}-8 x_{k-1}+8 x_{k+1}-x_{k+2}}{12 h_{i}} \tag{5.81}
\end{equation*}
$$

### 5.5 Novel Use of the Warped Multitime Partial Differential Equation (WaMPDE) to Determine the Locking Range

where $h_{i}$ is the time-step in the $\tau_{i}$ scale and $k$ denotes the index in the same time scale. The Backward Euler (BE) rule is used in the $t$ time scale for the transient evolution. One useful characteristic of the BE rule is that it introduces artificial damping if the time step is large. This helps to attenuate undesired oscillations along the $t$ dimension.

One phase condition is required here to obtain the local frequency. This is solved by making the derivative of one of the state variables equal to zero:

$$
\begin{equation*}
\frac{\partial \hat{x}_{0}}{\partial \tau_{1}}=0 \tag{5.82}
\end{equation*}
$$

The derivative of the solution has to be continuous for the state variable on which the phase condition is imposed. Then, the phase condition is guaranteed to be satisfied. When (5.81) is substituted into (5.80) and the BE rule is used, the resulting nonlinear algebraic equations are solved in conjunction with (5.82) using Newton's method [51] [52].

There are two methods introduced below, which utilise the WaMPDE to determine the locking range. One of them determines the limit of the locking range directly [53]. However, it has a relatively long simulation time. The second method [54] is proposed to improve this problem. It produces the Devil's staircase first, which will be introduced in Section 5.5.3. Then the locking range is measured from the staircase.

### 5.5.2 The Determination of the Limit of the Locking Range

The lower and upper limit of the locking range are determined separately. Here, the procedure to obtain the lower limit is described [53], and the upper limit can be determined with the similar method. To be specific, the work is concerned with the accurate determination of the acquisition locking range. To illustrate the important points of the method, results relating to a forced Van der Pol oscillator [53] are shown in Figs. 5.24-5.27.

### 5.5 Novel Use of the Warped Multitime Partial Differential Equation (WaMPDE) to Determine the Locking Range

The process to obtain the lower limit is started by running a transient simulation with the following initial condition: the free-running oscillation is set along $\tau_{1}$ and no oscillations along $\tau_{2}$. The frequency of the excitation is set to a value known not to lock the oscillator. (5.80) along with the boundary conditions is then solved for several time-steps along $t$ until the steady-state condition is achieved. Each step along $t$ involves the calculation of a bi-dimensional steady-state problem with a grid with 15 points in each dimension, $\tau_{1}$ and $\tau_{2}$. In order to minimize the number of time steps necessary to achieve the steady-state condition (along $t$ ), the time step size in this dimension is exponentially increased. When the difference between two steps along $t$ is small enough, the analysis is switched from a 3-D transient analysis to a 2-D steady-state analysis by making the derivatives with respect to $t$ equal to zero. Fig. 5.24 shows steady state solution when the oscillator is not locked. This procedure results in an autonomous solution along $\tau_{1}$, if it exists. At each time-step, but especially when solving for the steady-state condition, the Jacobian matrix that is employed in Newton's method is noted. If the column that corresponds to the local frequency is near zero, this implies the system is becoming independent of the local frequency because there are no more oscillations along $\tau_{1}$. Thus, an autonomous solution along $\tau_{1}$ is no longer possible and only the forced oscillations are present. Locking has happened. Note that at this condition the Jacobian matrix is singular and so the Newton iterations are stopped. Thus, the locking condition can also be observed in Fig. 5.25. The norm of the Jacobian column approaches zero if the oscillator is locked. The steady state solution when the oscillator is locked is illustrated in Fig. 5.26.

After the steady-state solution is obtained (assuming no locking), the input frequency ( $\omega_{i n}$ ) is progressively incremented. For each increment, a transient analysis is performed and followed by a steady-state analysis. If the oscillator is deemed locked, further analysis may be necessary to determine the starting point of the locking range to a certain degree of accuracy. The bisection method [51] is used in this work. There-

### 5.5 Novel Use of the Warped Multitime Partial Differential Equation (WaMPDE) to Determine the Locking Range



Figure 5.24: Steady state: Non-locked oscillator.


Figure 5.25: Comparison of the norm of the Jacobian column when the oscillator is locked (continuous line) and non-locked (dashed line).

### 5.5 Novel Use of the Warped Multitime Partial Differential Equation (WaMPDE) to Determine the Locking Range



Figure 5.26: Steady state: Locked oscillator.
fore, the frequency step by which $\omega_{i n}$ is incremented is halved and the procedure is re-started with the initial conditions from the last frequency point at which there was no locking. This progressive halving of the frequency step increment is repeated until the lower limit of the locking range is attained to a certain tolerance.

Figure 5.27 shows the variation of the forcing frequency and the norm of the Jacobian column as a function of the iteration number. If no locking is detected, the frequency is increased. Each time that locking is detected, a lower frequency is tried and the initial conditions are reset to the last known oscillatory solution. This is the reason for the peaks observed in the norm of the Jacobian column.

### 5.5 Novel Use of the Warped Multitime Partial Differential Equation (WaMPDE) to Determine the Locking Range



Figure 5.27: Local frequency and norm of Jacobian column (dashed line) vs. iteration number.

### 5.5.3 Determination of the Locking Range from Use of the WaMPDE in Conjunction with the Devil's Staircase

The Devil's Staircase [9] is a method to visualize the locking range of an ILFD. Normally, the ILFD is considered as an oscillator with an injected external signal. In order to plot a Devil's Staircase, the frequencies of the oscillator and the injected signal must be varied relative to each other. In practice, it is easier and more accurate to adjust the injected frequency, $\omega_{\mathrm{inj}}$, automatically. The output frequency of the ILFD, $\omega_{\mathrm{o}}$, is then the only unknown variable. The Devil' Staircase [9] is obtained by plotting $\omega_{\text {inj }} / \omega_{\text {o }}$ against $\omega_{\text {inj }}$, as shown in Fig. 5.28.

The locking range can be measured from the Devil's Staircase diagram [9]. From the staircase diagram in Fig. 5.28, it is clear that there are lockings (flat regions) at division ratios of 2 and 4, as predicted experimentally in [7] for the LC-oscillator based ILFD.


Figure 5.28: Experimentally measured Devil's staircase diagram showing lockings at $\omega_{\mathrm{inj}} / \omega_{\mathrm{o}}=2$ and $\omega_{\mathrm{inj}} / \omega_{\mathrm{o}}=4$.

### 5.5 Novel Use of the Warped Multitime Partial Differential Equation (WaMPDE) to Determine the Locking Range

From the experimental results on an ILFD performed in [9] and as shown in Fig. 5.28, it is noted that the relationship between $\omega_{\mathrm{inj}} / \omega_{\mathrm{o}}$ and $\omega_{\mathrm{inj}}$ is approximately linear between the locking intervals [54] (the ILFD locks at multiples of its natural frequency - the $n$th locking range is described by $\omega_{\mathrm{inj}} / \omega_{\mathrm{o}}=n$ ). During the locking intervals, the slope is obviously zero. Consequently, two simulations are performed with two carefully selected input frequencies. The two selected $\omega_{\text {inj }}$ are known not to lock the ILFD and to be below the lower limit of the particular $n$th locking range. In other words, (5.80) is solved twice to obtain the values of $\omega_{\mathrm{o}}$ corresponding to two values of the manually picked $\omega_{\mathrm{inj}}$. From this, an estimate of the start of the $n$th locking range can be obtained. For instance, the start of the divide-by-two locking range is when $\omega_{\mathrm{inj}} / \omega_{\mathrm{o}}=2$. The slope of the line connecting the two points below the start is determined as:

$$
\begin{equation*}
m_{\mathrm{below}}=\frac{\left(\frac{\omega_{\mathrm{inj} 2}}{\omega_{\mathrm{o} 2}}\right)-\left(\frac{\omega_{\mathrm{inj} 1}}{\omega_{\mathrm{o} 1}}\right)}{\omega_{\mathrm{inj} 2}-\omega_{\mathrm{inj} 1}} . \tag{5.83}
\end{equation*}
$$

The slope can be also calculated by

$$
\begin{equation*}
m_{\text {below }}=\frac{2-\left(\frac{\omega_{\mathrm{inj1}}}{\omega_{\mathrm{ol}}}\right)}{\omega_{\mathrm{start}}-\omega_{\mathrm{inj} 1}} . \tag{5.84}
\end{equation*}
$$

where $\omega_{\text {start }}$ is the lower limit of the locking interval when $\omega_{\mathrm{inj}} / \omega_{\mathrm{o}}=2$. Hence, the lower limit of the locking range is

$$
\begin{equation*}
\omega_{\text {start }}=\omega_{\text {inj } 1}+\frac{2-\left(\frac{\omega_{\text {inj } 1}}{\omega_{\mathrm{ol}}}\right)}{m_{\text {below }}} \tag{5.85}
\end{equation*}
$$

The upper limit of the locking range can be obtained with a similar procedure. In this case, the two input frequencies are selected to determine the slope between the divide by 2 and the divide by 4 locking ranges.

$$
\begin{equation*}
2 \omega_{n}<\omega_{\text {inj } 1}<\omega_{\text {inj } 2}<4 \omega_{n} \tag{5.86}
\end{equation*}
$$

where $\omega_{n}$ is the natural frequency of the oscillator.

### 5.6 A Novel Method Utilising the Multiple-Phase-Condition Envelope Following Method (MPCENV) to Determine the Locking Range

### 5.6.1 Background to the Envelope Following Method

The circuit solution is assumed to be composed of fast oscillations whose amplitude and frequency vary much more slowly than the oscillations themselves. Let the period of the fast oscillation be $T$. In the case of oscillators, this will vary slowly. Let $T_{\text {env }}$ be the envelope time-step over which the response of the system can be extrapolated.


Figure 5.29: Forward-Euler-based envelope-following method.

The Forward-Euler-based envelope-following method is illustrated in Fig. 5.29. Firstly, an accurate transient simulation is performed from $x_{0}$ for one cycle of the fast oscillation $T$. This brings one to point $x_{1}$. If there are $(m-1) T$ cycles between $x_{1}$ and $x_{\text {end }}$, the following equation is obtained:

$$
\begin{equation*}
\frac{x_{1}-x_{S}}{m T}=\frac{x_{L}-x_{S}}{T} \tag{5.87}
\end{equation*}
$$

However, the Forward-Euler-based envelope-following method suffers from stability problems, when the period of the envelope $T_{\text {env }}$ is too large. To overcome this, the Backward Euler method is preferred. This will be described in Section 5.6.2.


Figure 5.30: Backward-Euler-based envelope-following method.

### 5.6.2 Multiple-Phase-Condition Envelope Following method

Here, a novel transient envelope following method, MPCENV, is proposed to determine the output frequencies corresponding to different input frequencies [55]. Consider Fig. 5.30. Let $x_{0}$ and $x_{1}$ be the state at $t_{0}=t_{\mathrm{S}}+T_{\text {env }}$ and $t_{1}=t_{\mathrm{S}}+T_{\text {env }}+T$, respectively, where $t_{\mathrm{S}}$ is the ending time of the last envelope step. Using the implicit Backward-Euler method for stability purposes, the envelope following process is described by:

$$
\begin{equation*}
\frac{x_{1}-x_{0}}{T}=\frac{x_{0}-x_{\mathrm{S}}}{T_{\mathrm{env}}} \tag{5.88}
\end{equation*}
$$

where $x_{S}=x\left(t_{\mathrm{S}}\right)$ is known from a previous step, and $x_{1}$ is determined using the trapezoidal integration method from $t_{0}$ to $t_{1}$. This means that $x_{1}$ depends on $x_{0}$ and $T$. Note that apart from the circuit variables, there are two extra unknowns, $T$ and $T_{\text {env }}$, since the period of the oscillator is always changing, and $T_{\text {env }}$ has to vary in order to remain equal to an integer number of periods $T$. To solve for the extra unknowns, two further equations are required [56]:

$$
\left\{\begin{align*}
\frac{d x_{0 l}}{d t} & =0  \tag{5.89}\\
\frac{d x_{1 l}}{d t} & =0
\end{align*}\right.
$$

### 5.6 A Novel Method Utilising the Multiple-Phase-Condition Envelope Following Method (MPCENV) to Determine the Locking Range

where $l$ denotes the $l_{t h}$ state variable. The two derivative-based phase conditions (5.89) ensure that $x_{0 l}$ and $x_{1 l}$ are the peaks or troughs of a fast cycle. In practice, value-based constraints are better for numerical handling of certain circuits such as the ILFD:

$$
\left\{\begin{array}{l}
x_{0 l}=c  \tag{5.90}\\
x_{1 l}=d
\end{array}\right.
$$

where $c$ and $d$ are constants.
Equations (5.88) and (5.90) are reorganized as a matrix and solved using the NewtonRaphson method [57]:

$$
\begin{align*}
F & =\left[\begin{array}{l}
f_{1}\left(x_{0}, T, T_{\text {env }}\right) \\
f_{2}\left(x_{0}, T, T_{\text {env }}\right) \\
f_{3}\left(x_{0}, T, T_{\text {env }}\right)
\end{array}\right] \\
& =\left[\begin{array}{c}
\left(x_{1}-x_{0}\right) T_{\text {env }}-\left(x_{0}-x_{s}\right) T \\
x_{0 l}-c \\
x_{1 l}-d
\end{array}\right]  \tag{5.91}\\
& =0 .
\end{align*}
$$

If the circuit has $n$ state variables, this system consists of $n+2$ equations with $n+2$ unknowns.

The Jacobian matrix corresponding to (5.91) is given by:

$$
\begin{align*}
J & =\left[\begin{array}{lll}
\frac{d f_{1}}{d x_{0}}, & \frac{d f_{1}}{d T}, & \frac{d f_{1}}{d T_{\text {env }}} \\
\frac{d f_{2}}{d x_{0}}, & \frac{d f_{2}}{d T}, & \frac{d f_{2}}{d T_{\text {env }}} \\
\frac{d f_{3}}{d x_{0}}, & \frac{d f_{3}}{d T}, & \frac{d f_{3}}{d T_{\text {env }}}
\end{array}\right] \\
& =\left[\begin{array}{ccc}
\frac{\partial x_{1}}{\partial x_{0}} T_{\text {env }}-\left(T_{\text {env }}+T\right) I_{n}, & \frac{\partial x_{1}}{\partial T} T_{\text {env }}-\left(x_{0}-x_{s}\right), & x_{1}-x_{0} \\
& \left.I_{n}\right|_{l}, & 0, \\
& \left.\frac{\partial x_{1}}{\partial x_{0}}\right|_{l}, & \left.\frac{\partial x_{1}}{\partial T}\right|_{l},
\end{array}\right] \tag{5.92}
\end{align*}
$$

where $I_{n}$ is an identity matrix of size $n \times n,\left.I_{n}\right|_{l},\left.\left(\partial x_{1} / \partial x_{0}\right)\right|_{l}$ and $\left.\left(\partial x_{1} / \partial T\right)\right|_{l}$ are the $l$ th row of $I_{n}, \partial x_{1} / \partial x_{0}$ and $\partial x_{1} / \partial T$, respectively.

### 5.6 A Novel Method Utilising the Multiple-Phase-Condition Envelope Following Method (MPCENV) to Determine the Locking Range

In this implementation, both $\partial x_{1} / \partial x_{0}$ and $\partial x_{1} / \partial T$ are derived using the trapezoidal integration method, as introduced in [58]. Set $x_{r}$ to be the state at $t_{r}$, where $t_{0} \leq t_{r-1}<t_{r} \leq t_{1}$. Then

$$
\begin{equation*}
\frac{d x_{r}}{d x_{0}}=\left(1-\frac{h}{2} \frac{\partial f\left(x_{r}\right)}{\partial x_{r}}\right)^{-1}\left(1+\frac{h}{2} \frac{\partial f\left(x_{r-1}\right)}{\partial x_{r-1}}\right) \frac{d x_{r-1}}{d x_{0}} \tag{5.93}
\end{equation*}
$$

where $f(x)$ is the expression to represent the derivative of the circuit variables: $\dot{x}=$ $f(x)$. The term $d x_{1} / d x_{0}$ can be found by repeatedly evaluating (5.93) from $t_{0}$ to $t_{1}$ with $d x_{0} / d x_{0}=I$, where $I$ is an $n \times 1$ matrix with all ones.

In a similar manner, $d x_{1} / d T$ can be found by solving

$$
\begin{align*}
\frac{d x_{r}}{d T}= & \left(1-\frac{h}{2} \frac{\partial f\left(x_{r}\right)}{\partial x_{r}}\right)^{-1} \\
& \times\left[\left(1+\frac{h}{2} \frac{\partial f\left(x_{r-1}\right)}{\partial x_{r-1}}\right) \frac{d x_{r-1}}{d T}+\frac{x_{r}-x_{r-1}}{T}\right] \tag{5.94}
\end{align*}
$$

starting from $d x_{0} / d T=0$.
Then the system in (5.91) can be solved using the Newton-Raphson method [57]:

$$
\begin{equation*}
Z_{\text {new }}=Z-J^{-1} F \tag{5.95}
\end{equation*}
$$

where $Z_{\text {new }}$ and $Z$ represent the current and previous states of all the variables, i.e., $Z=\left[\begin{array}{lll}x_{0} & T & T_{\text {env }}\end{array}\right]^{T}$. In the case of the ILFD, $x_{0}$ represents the capacitance voltage and the inductance current, i.e., $x_{0}=\left[\begin{array}{ll}V_{C} & I_{L}\end{array}\right]^{T}$.

As described in [9], the Devil's Staircase is a plot of $\omega_{\mathrm{inj}} / \omega_{0}$ against $\omega_{\mathrm{inj}}$. For simulation purposes, the injected frequency, $\omega_{\mathrm{inj}}$ is increased from the minimum $\omega_{\mathrm{inj}}$ with a fixed frequency step-size. $\omega_{0}$ is then determined from the MPCENV solution as:

$$
\begin{equation*}
\omega_{0}=\frac{2 \pi}{T} . \tag{5.96}
\end{equation*}
$$

### 5.7 Numerical Results and Experimental Validation

### 5.7.1 Experimental Equipments Setup

Here we describe our experimental technique [9] to measure the locking range of ILFD. In order to measure a Devil's Staircase, we must vary the frequencies of the oscillator and the injected signal relative to one another. In practice, it is easier and more accurate to vary the injected frequency automatically. Hence we fix the natural frequency of the ILFD, $f_{n}$, and adjust the frequency, $f_{i n j}$, of the injected signal.


Figure 5.31: Experiment setup

An automated measurement system has been developed which includes a personal computer, a function generator, a frequency counter and the ILFD circuit [9]. The experimental setup is shown in Fig. 5.31. The driving voltage $V_{\mathrm{inj}}$ and the driving frequency $\omega_{\text {inj }}$ are produced by a precision frequency generator (Agilent 33220A) connected to an IEEE-488 bus. The frequency of the driven oscillator $\left(\omega_{\mathrm{o}}\right)$ is obtained through the use of a precision counter (Agilent 53131A) which is also under IEEE-488 control.

### 5.7.2 Numerical Results

The LC oscillator-based ILFD (LC-ILFD) is selected as an example. It was introduced in Section 5.3.4. The schematic is shown again in Fig. 5.32.


Figure 5.32: The schematic of the LC-ILFD.

The governing equations are:

$$
\begin{align*}
C \frac{d V_{C}}{d t} & =I_{L}-\left(A+d a V_{\mathrm{inj}}\right) V_{C}+\frac{A+d a V_{\mathrm{inj}}}{V_{\mathrm{DD}}^{2}} V_{C}^{3}  \tag{5.97}\\
L \frac{d V_{C}}{d t} & =-I_{L}-V_{C}
\end{align*}
$$

where $A$ and $d a$ are the coefficients obtained from the negative resistance characteristic [7] shown in Fig. 5.11 and Fig. 5.13. The value of the selected components ${ }^{1}$ is shown in Table 5.1.

The Devil's Staircases obtained by simulation and experiment are shown in Fig. 5.33. The widths of the locking ranges agree when $\omega_{\mathrm{inj}} / \omega_{\mathrm{o}}$ is an even number [7] i.e., 2 and 4, as shown in Fig. 5.28 and Fig. 5.33. The staircase from MPCENV is almost the

[^2]Table 5.1: The value of the selected components used in both experiment and simulation.

| $C$ | $L$ | $R_{S}$ | $V_{D D}$ | $A$ | $d a$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $100 p F$ | $220 \mu H$ | $3.06 \Omega$ | $9 V$ | $-2.5 m A / V$ | $-0.1 m A / V^{2}$ |



Figure 5.33: The staircase obtained from simulation and experiment.
same as that from experiment, while the one from use of the WaMPDE described in Section 5.5.3 has an appreciably bigger difference. However, it should be accepted that this is a very basic method. It could be used to obtain an initial estimate.

Table 5.2: Locking range, where M, E and $\mathbf{W}$ represent MPCENV, Experiment and WaMPDE

|  | $\omega_{\text {inj }} / \omega_{\mathrm{o}}=2$ |  |  |
| :---: | :---: | :---: | :---: |
| $V_{\text {inj }}$ | $\mathrm{M}(\mathrm{Mrad} / \mathrm{s})$ | $\mathrm{E}(\mathrm{Mrad} / \mathrm{s})$ | $\mathrm{W}(\mathrm{Mrad} / \mathrm{s})$ |
| 1 V | 0.63 | 0.62 | 0.63 |
| 1.5 V | 0.88 | 0.92 | 0.87 |
|  | $\omega_{\text {inj }} / \omega_{\mathrm{o}}=4$ |  |  |
|  | $\mathrm{M}(\mathrm{Mrad} / \mathrm{s})$ | $\mathrm{E}(\mathrm{Mrad} / \mathrm{s})$ | $\mathrm{W}(\mathrm{Mrad} / \mathrm{s})$ |
| 1 V | 1.1 | 1.04 | 1.12 |
| 1.5 V | 1.48 | 1.51 | 1.49 |

Table 5.2 shows the locking ranges captured from the staircases. The difference between the MPCENV method and experimental results is less than $6 \%$, while that between the full WaMPDE method described in Section 5.5.2 and experimental results is almost $8 \%$. Therefore, it is sufficiently accurate to predict the locking range when designing ILFDs of this type. The locking range captured from experiments is illustrated in Fig. 5.34. As seen from it, if the amplitude of the injected signal is larger, then the locking range is larger.


Figure 5.34: The locking range of the LC-ILFD captured from experiments.

### 5.8 Conclusions

Injection-locked frequency dividers (ILFDs) are normally used as the prescaler in the feedback of a Phase-Locked-Loop(PLL). In comparison with the conventional static frequency divider, they consumes less power, but their locking range is smaller. Hence, its accurate determination is essential in design work. The WaMPDE method is used to determine the acquisition range for locking. However, it requires quite a long simulation time when implemented in full. An improved method is proposed to reduce the simulation time. However, it is less accurate. The MPCENV method is another suggested method. It follows the envelope following technique. It is used to enable a simple plotting of the Devil's Staircase from which the locking range can be determined. With this method, the simulation speed is increased and the error is reduced to $6 \%$.

## Chapter 6

## Analysis and Design of Digital Delta-Sigma Modulators (DDSM)

### 6.1 Background to Delta-Sigma Modulation

The delta-sigma modulation (DSM) technique has become popular in very-large-scale integration (VLSI) circuits in recent years, although it was first introduced in 1962 [59]. Its main use is in data converters and in frequency synthesis. Examples include analog-to-digital converters (ADC), digital-to-analog converters (DAC), direct digital frequency synthesizers (DDFS) and fractional- $N$ frequency synthesizers.

Delta-sigma modulation is based on delta modulation (DM), which is the simplest form of the differential pulse-code modulation (DPCM). A simple example of pulse-code modulation ( PCM ) is shown in Fig. 6.1: As seen from Fig. 6.1, a signal, represented by the sine wave, is sampled and quantized at fixed intervals. In other words, the signal is discrete and can be expressed by digital words. For example, the quantized values of it are $7,9,11,12,13,14,14,15,15,15,14, \ldots$ These may then be expressed in binary format as $0111,1001,1011,1100,1101,1110,1110,1111$, $1111,1111,1110, \ldots$, which can be used for digital signal processing. DPCM quantizes the difference between successional samples of a continuous-time signal into an n-bit binary word. Delta-modulation is 1-bit DPCM.


Figure 6.1: An example of 4-bit PCM.

1-bit quantizer


Figure 6.2: Block diagram of the delta-modulation process.

The block diagram of the delta-modulation process is illustrated in Fig. 6.2. The demodulation part is the inverse procedure of the delta-modulation. It converts the digital data stream to a smooth analog signal.

Integration is a linear operation and hence has the following property

$$
\begin{equation*}
\int a+\int b=\int(a+b) \tag{6.1}
\end{equation*}
$$

Consequently, the integrator in the demodulator can be moved to the position between the input analog signal and the modulator, as shown in Fig. 6.3. Furthermore, the two integrators in Fig. 6.3 (b) can be merged into one integrator as shown in Fig. 6.4. This resultant system is the so-called delta-sigma modulator [59].


Figure 6.3: Derivation of Delta-Sigma Modulation from Delta Modulation.

1-bit quantizer


Figure 6.4: Block diagram of delta-sigma modulation.


Figure 6.5: Spectrum of a signal.


Figure 6.6: Spectrum of an oversampled signal.

Oversampling is one of the most popular techniques used in delta-sigma modulation. The most attractive benefit of oversampling is that it reduces noise. Figure 6.5 illustrates the concept. $f_{m}$ and $f_{s}$ represent the bandwidth and the sampling frequency, respectively. In signal processing, if the sampling frequency is higher than twice of the bandwidth, i.e., $f_{s} \geq 2 f_{m}$, this process is termed oversampling. If the sampling frequency is $N$ times of the bandwidth, the noise power in the useful band is reduced to $1 / N$. That means the signal-noise ratio (SNR) is increased by a factor of $N$. The technique of moving the noise power to higher frequencies is termed noise shaping. The quantization noise is pushed to higher frequencies out of the useful band in a delta-sigma modulator. Consequently, it can be simply filtered by a low-pass filter.

### 6.2 Introduction of Digital Delta-Sigma Modulators

Digital delta-sigma modulators (DDSM) perform delta-sigma modulation on a digital input and produce a digital output. They take an $n_{0}$-bit input signal and produce an $m$-bit output signal. This results in a quantization error that is often modelled as an addition of white noise. The modulator filters the spectrum such that its noise power is concentrated at higher frequencies away from the signal band. They are widely used in consumer entertainment and communication equipment such as Fractional- $N$ Frequency Synthesizers, cellular telephones and MP3 players. Fractional- $N$ Frequency Synthesizers [12] [13] in general, are employed as Local Oscillators (LO) to generate exactly the required frequencies. A Fractional- $N$ Frequency Synthesizer is shown in Fig. 6.7. It consists of a phase/frequency detector, a charge pump, a loop filter and a voltage controlled oscillator in the forward path. The feedback path includes an injection-locked frequency divider (ILFD) [7] [8] [53] [54] [55] which is chosen to be the pre-scaler due to its low power consumption. The static frequency divider is controlled by a fully-digital delta-sigma modulator (DDSM). It ensures that the division ratio is the required fractional value. The input to the DDSM is a constant digital word

X . This is a high resolution value and it sets the fractional division ratio. The output is a low resolution sequence which controls the static divider. The spectrum of the output contains the desired input DC tone and the undesired quantization noise.


Figure 6.7: The Fractional-N Frequency Synthesizer.

Delta-sigma modulators (DSM) are attractive as they push most of the quantisation noise to higher frequencies and hence out of the useful frequency band, which is socalled noise shaping as introduced in Section 6.1. Thus the quantisation noise power, obtained from the DDSM appears at frequencies out of the pass-band of the fractional$N$ frequency synthesizer and can simply be filtered without affecting the signals in the frequency synthesizer.

DDSMs fall into two categories [60]: single-loop delta-sigma modulators and Multi-stAge noise SHaping (MASH) delta-sigma modulators. The architecture of an $l$ th order single-loop delta-sigma modulator is shown in Fig. 6.8. It consists of $l$ stages of 1st order modulators (DSM1) introduced in Section 6.1. $\mathrm{x}[\mathrm{n}]$ and $\mathrm{y}[\mathrm{n}]$ represent the $n_{0}$-bit input and $m$-bit output binary word, respectively. The input of the $i$ th stage of DSM1 is the output of the $(i-1)$ th DSM1 where $i \in\{2,3, \ldots, l\}$. , The input of the 1st stage is the difference between $\mathrm{x}[\mathrm{n}]$ and $\mathrm{y}[\mathrm{n}]$. The sum of the output of the $l$ DSM1s is quantized by a quantizer to produce the output of the complete modulator. The ad-


Figure 6.8: The architecture of an lth order single-loop delta-sigma modulator.
vantage of this arrangement is that it has the choice of either a single-bit or a multi-bit output, i.e., $m=1$ or $m \geq 2$. The components receiving only a single-bit data stream require less hardware than those receiving multi-bit data. However, the disadvantage of the single-loop DSM is that it is prone to instability because of the feedback loop inherent within this architecture [60]. In contrast, the MASH DDSM has has no $n$th order feedback and 1st order stages have been proven to be stable so that it is unconditionally stable [60]. However, now there is one quantizer in the DSM1 of each stage instead of one quantizer after the last DSM1. This structure results in a multi-bit output $y[n]$, i.e., $m \geq 2$. The architecture and characteristics of MASH DDSMs will be described in detail later in this Chapter.

The DDSM is a finite state machine. Hence, when the input is a constant, the output is always a repeating pattern (a cycle) [14] [61]. In particular, the quantization noise is a periodic sequence. The period of the cycle is termed the sequence length. Its length depends on the input, the initial conditions of the DDSM registers and the DDSM architecture. Short sequence lengths result in unwanted frequency components or spurs from quantisation noise in the output frequency spectrum [62]. Hence, much research has been done into detecting and removing the occurrence of a short sequence length [63].


Figure 6.9: Model of a dithered MASH DDSM.

There are two classes of technique for whitening quantization noise, stochastic and deterministic. The dithering [64] [65] method is one of the most common stochastic approaches employed. It uses a random dither sequence to disrupt the periodic cycle and thereby effectively increase the sequence length. However, it requires extra hardware and inherently introduces additional noise in the the useful frequency band. The structure of a dithered MASH DDSM is illustrated in Fig. 6.9. $d[n]$ is an 1-bit pseudorandom dither sequence. The probability of $d[n]=0$ and $d[n]=1$ is set manually to be equal.

$$
\begin{equation*}
P(d[n]=0)=P(d[n]=1)=50 \% . \tag{6.2}
\end{equation*}
$$

The block $F(z)$ represents a filter after the dither input. The power spectral density of the output sequence of a 9-bit MASH 1-1-1 ${ }^{1}$ DDSM is shown in Fig. 6.10 for three cases:

1. $F(z)=0$, no dither
2. $F(z)=1$, normal dither
3. $F(z)=\left(1-z^{-1}\right)$, shaped dither

[^3]

Figure 6.10: The power spectral density of a 9-bit MASH 1-1-1 DDSM [3].

As seen in Fig. 6.10, the spectrum from the MASH DDSM without dither has strong tonal behaviour, while that from the dithered MASH DDSM is much smoother, because the dither lengthens the sequence length. The filter $F(z)$ is used to improve the noise performance at low frequencies.

Recently, some deterministic design methodologies have been proposed to maximise the sequence length. Borkowski [14] gives a guaranteed minimum and maximum period obtained by setting the initial condition of the registers. Hosseini [3] introduced a digital delta-sigma modulator structure termed the HK-MASH with a very long sequence and the period of such a sequence is derived by mathematical analysis [66] [67] [68]. They will be described in Section 6.3. A novel design methodology for the MASH DDSM employing multi-moduli [15] will be proposed in this Chapter. The new structure is termed the MM-MASH. It can produce a longer sequence length than that of HK-MASH.

### 6.3 State-of-the-art Architectures of DDSMs

### 6.3.1 The Conventional DDSMs

There are two classes of structure for digital delta-sigma modulators: Multi-stAge noise SHaping (MASH) and error feedback modulator (EFM). Normally, a MASH modulator is comprised of several EFMs. The conventional structures of both of them are introduced in this Section.

Noise Cancellation Networks


Figure 6.11: MASH DDSM architecture.

### 6.3.1.1 The Conventional MASH

The architecture of an $l$ th order MASH digital delta-sigma modulator (DDSM) is illustrated in Fig. 6.11. It contains $l$ first-order error-feedback modulators (EFM1). $x[n]$ and $y[n]$ are an $n_{0}$-bit input digital word and an $m$-bit output, respectively. The relationship between them is

$$
\begin{equation*}
\operatorname{mean}(y)=\frac{X}{M} \tag{6.3}
\end{equation*}
$$

where $X$ is the decimal number corresponding to the digital sequence $x[n]$ [69], i.e., $x[n]=X \in\{1,2, \ldots, M-1\}$, and $M$ is the quantizer modulus which is set as $2^{n_{0}}$ in
the conventional DDSM. mean $(y)$ is the desired fraction in a fractional- $N$ frequency synthesizer.

The model of the EFM1 is shown in Fig. 6.12. This is a core component in the make-up of the MASH digital delta-sigma modulator (DDSM). The rectangle $Z^{-1}$ represents the register which stores the error $e[n]$ and delays it for one time sample. $Q(\cdot)$ is the quantization function:

$$
y[n]=Q(u[n])= \begin{cases}1, & u[n] \geq M  \tag{6.4}\\ 0, & u[n]<M\end{cases}
$$

where

$$
\begin{equation*}
u[n]=x[n]+e[n-1] . \tag{6.5}
\end{equation*}
$$

In order to simplify the analysis, a linear model of the EFM1 is introduced here. As seen in Fig. 6.13, an additive quantization noise source $e_{a s}[n]$ is used to take the place of the quantizer. The expression for the linear model can be written in $z$-domain as below:

$$
\begin{align*}
& Y(z)=U(z)+E_{\text {as }}(z)  \tag{6.6}\\
& U(z)=X(z)+E(z) z^{-1} . \tag{6.7}
\end{align*}
$$

Substituting (6.7) into (6.6),

$$
\begin{equation*}
Y(z)=X(z)+E(z) z^{-1}+E_{a s}(z) . \tag{6.8}
\end{equation*}
$$

Consider the expression for the error $e[n]$ in the $z$-domain:

$$
\begin{equation*}
E(z)=U(z)-Y(z) . \tag{6.9}
\end{equation*}
$$

By merging (6.6) and (6.9), it follows that

$$
\begin{equation*}
E(z)=-E_{a s}(z) \tag{6.10}
\end{equation*}
$$



Figure 6.12: EFM1: First-order error-feedback modulator.


Figure 6.13: The linear model of an EFM1.

Combining (6.10) and (6.8),

$$
\begin{equation*}
Y(z)=X(z)+\left(1-z^{-1}\right) E_{a s}(z) \tag{6.11}
\end{equation*}
$$

Thus it can be concluded that the EFM1 output consists of the input sequence and the output quantization error. The output quantization error is the quantizer error $E_{a s}(z)$ which is shaped by a filter $\left(1-z^{-1}\right)$. From (6.11), it follows that for the conventional EFM1, the signal transfer function (STF) and the noise transfer function (NTF) are:

$$
\begin{align*}
& S T F=1  \tag{6.12}\\
& N T F=1-z^{-1} \tag{6.13}
\end{align*}
$$

They represent a measure of the output signal versus the input signal and versus the noise in a system, respectively.

In a MASH DDSM, there is a cascade of several EFM1, as shown in Fig. 6.11. The error of each stage $\left(e_{i}[n]\right)$ is fed to the next stage and the carry-out of each stage $\left(y_{i}[n]\right)$ is fed to the noise cancellation network to eliminate the intermediate quantization noises [3]. Thus the spectrum of the final output depends only on the input signal and the shaped quantization noise power of the last stage.

$$
\begin{equation*}
Y(Z)=X(z)+\left(1-z^{-1}\right)^{l} E_{\text {asl }}(z) \tag{6.14}
\end{equation*}
$$

where $E_{\text {asl }}(z)$ is the quantization error of the last stage. Then the signal transfer function and the noise transfer function for a conventional MASH are:

$$
\begin{align*}
& S T F=1  \tag{6.15}\\
& N T F=\left(1-z^{-1}\right)^{l} . \tag{6.16}
\end{align*}
$$

### 6.3.1.2 The Conventional EFM

The structure of an lth order EFM (EFMl) is illustrated in Fig. 6.14. The symbols $R_{1}, R_{2}, \ldots, R_{l}$ and $A_{1}, A_{2}, \ldots, A_{l}$ represent the initial conditions and gains of the $i$ th registers, respectively, where $i \in\{1,2, \ldots, l\}$. The expressions governing the linear model of it as shown in Fig. 6.15 are

$$
\begin{align*}
& Y(z)=U(z)+E_{a s}(z)  \tag{6.17}\\
& U(z)=X(z)+\sum_{i=1}^{l} A_{i} z^{-i} E(z)  \tag{6.18}\\
& E(z)=U(z)-Y(z) \tag{6.19}
\end{align*}
$$

From the equations above, it is obtained

$$
\begin{equation*}
Y(z)=X(z)+\left(1-\sum_{i=1}^{l} A_{i} z^{-i}\right) E_{a s}(z) \tag{6.20}
\end{equation*}
$$

where

$$
\begin{equation*}
E_{a s}(z)=-E(z) . \tag{6.21}
\end{equation*}
$$

Thus the signal transfer function and the noise transfer function for an lth order EFM are

$$
\begin{align*}
& S T F_{E F M}=1  \tag{6.22}\\
& N T F_{E F M}=1-\sum_{i=1}^{l} A_{i} z^{-i} . \tag{6.23}
\end{align*}
$$

Since the higher-order EFMl is used as the delta-sigma modulator, it should have the same noise transfer function as a MASH composed of a cascade of first-order EFMs given by (6.16). Then

$$
\begin{equation*}
1-\sum_{i=1}^{l} A_{i} z^{-i}=\left(1-z^{-1}\right)^{l} \tag{6.24}
\end{equation*}
$$

Thus the value of the gains $A_{i}$ is obtained from:

$$
\begin{equation*}
\sum_{i=1}^{l} A_{i} z^{-i}=1-\left(1-z^{-1}\right)^{l} \tag{6.25}
\end{equation*}
$$



Figure 6.14: Higher-order error-feedback modulator.


Figure 6.15: Linear model of higher-order error-feedback modulator.

### 6.3.1.3 The Sequence Length of Conventional DDSMs

The sequence length of the two types of conventional DDSMs, the MASH and the EFM, is controlled by the initial conditions of the registers, which are represented by the symbols $\left\{R_{1}, R_{2}, \ldots, R_{l}\right\}$ shown in Fig. 6.11 and Fig. 6.14. Borkowski [14] proposed the so-called preferred initial conditions, which are shown in Table 6.1. If the DDSMs work with these initial conditions, the guaranteed minimum sequence length is dependent only on the modulator bus width $n_{0}$ and is not affected by the DC input any more. The guaranteed minimum and maximum sequence lengths for these structures have been found from simulations [14], and are as shown in Table 6.2. The modulator period varies between the guaranteed minimum and the maximum sequence length dependent on the value of the input. Note that the sequence length of an $l$ th order MASH and an lth order EFM is the same. It is found that the conventional MASH modulator period will always be approximately equal to the maximum sequence length if the quantizer modulus, $M$, is set as a prime number [66] [67] [68].

Table 6.1: The preferred initial conditions for the conventional DDSMs.

| Modulator Order | MASH | EFM $l$ |
| :---: | :---: | :---: |
| 2 | $R_{1}$ odd | $R_{1}+R_{2}$ odd |
| 3 | $R_{1}$ odd | $R_{1}+R_{3}$ odd |
| 4 | $R_{1}$ odd | $R_{1}+R_{2}+R_{3}+R_{4}$ odd |
| 5 | $R_{1}$ odd | $R_{1}+R_{5}$ odd |

Table 6.2: The sequence length of the conventional DDSMs.

| Modulator Order | Guaranteed Minimum Period | Maximum Period |
| :---: | :---: | :---: |
| 2 | $2^{n_{0}-1}$ | $2^{n_{0}+1}$ |
| 3 | $2^{n_{0}+1}$ | $2^{n_{0}+1}$ |
| 4 | $2^{n_{0}+1}$ | $2^{n_{0}+2}$ |
| 5 | $2^{n_{0}+2}$ | $2^{n_{0}+2}$ |

### 6.3.2 HK-MASH DDSMs

The HK-MASH DDSM is the current MASH that produces the longest sequence length. It has two advantages. First, it produces a much longer sequence length when compared to the conventional MASH. Secondly, its maximum sequence length can be achieved independent of the initial conditions described in Section 6.3.1 and independent of all constant inputs.


Figure 6.16: The modified EFM1 used in the HK-MASH.

The architecture of the modified EFM1 used in the HK-MASH (HK-EFM1) is illustrated in Fig. 6.16. The only difference between it and the conventional EFM1 in Fig. 6.12 is the presence of the feedback block $a z^{-1}$. The effect of this block is that, unlike the conventional DDSM, the quantizer is not reset to 0 again, but to $a$, when it overflows. Equivalently, the effective quantizer mudulus is changed to $M-a$ from $M$. $a$ is a specifically-chosen small integer to make $(M-a)$ the maximum prime number below $2^{n_{0}}[3]$. The value of $a$ chosen in [3] is shown in Table 6.3.

The sequence length of an lth order HK-MASH is determined by the following formula [3]:

$$
\begin{equation*}
N=\left(2^{n_{0}}-a\right)^{l} . \tag{6.26}
\end{equation*}
$$

Table 6.3: The specifically-chosen a for various modulator word lengths.

| $n_{0}$ | $a$ |
| :---: | :---: |
| $5,7,13,17,19$ | 1 |
| $6,9,10,12,14,20,22,24$ | 3 |
| $8,18,25$ | 5 |
| 11,21 | 9 |
| 16,23 | 15 |
| 15 | 19 |

The maximum sequence length of an lth order HK-MASH is compared with that of the conventional MASH DDSM in Table 6.4. The maximum sequence length of the HK-MASH is almost $\left(2^{n_{0}}\right)^{l-1}$ times that of the conventional MASH DDSM. Note that, since the input does not affect the period of HK-MASH output, the guaranteed minimum sequence length of it is the same as its maximum sequence length.

Table 6.4: A comparison between the maximum sequence length of the conventional MASH and the HK-MASH.

| Modulator Order $l$ | Conventional MASH | HK-MASH |
| :---: | :---: | :---: |
| 2 | $2^{n_{0}+1}$ | $\left(2^{n_{0}}-a\right)^{2} \approx\left(2^{n_{0}}\right)^{2}$ |
| 3 | $2^{n_{0}+1}$ | $\left(2^{n_{0}}-a\right)^{3} \approx\left(2^{n_{0}}\right)^{3}$ |
| 4 | $2^{n_{0}+2}$ | $\left(2^{n_{0}}-a\right)^{4} \approx\left(2^{n_{0}}\right)^{4}$ |
| 5 | $2^{n_{0}+2}$ | $\left(2^{n_{0}}-a\right)^{5} \approx\left(2^{n_{0}}\right)^{5}$ |

The linear model of the HK-EFM1 is illustrated in Fig. 6.17. As with the linear model of conventional EFM1, the quantizer in the HK-EFM1 is replaced by an additive noise source $\left(e_{a s}[n]\right)$. Since all of the signals shown in Fig. 6.17 are normalized to 1 in order to do the linear analysis, the coefficient $a$ is modified to $\alpha$ :

$$
\begin{equation*}
\alpha=\frac{a}{2^{n_{0}}} . \tag{6.27}
\end{equation*}
$$

The expression for the output $y[n]$ in terms of the input $x[n]$ and the quantizer error


Figure 6.17: The linear model of the HK-EFM1.
$e_{a s}[n]$ in the $z$-domain is:

$$
\begin{align*}
& Y(z)=U(z)+E_{a s}(z)  \tag{6.28}\\
& E(z)=U(z)-Y(Z) . \tag{6.29}
\end{align*}
$$

Thus it follows:

$$
\begin{equation*}
E(z)=-E_{a s}(z) . \tag{6.30}
\end{equation*}
$$

After summing the input and the feedback,

$$
\begin{equation*}
U(z)=X(z)+\alpha z^{-1} Y(z)+z^{-1} E(z) . \tag{6.31}
\end{equation*}
$$

Substituting (6.30) into (6.31),

$$
\begin{equation*}
U(z)=X(z)+\alpha z^{-1} Y(z)-z^{-1} E_{a s}(z) . \tag{6.32}
\end{equation*}
$$

From (6.28), it follows

$$
\begin{equation*}
U(z)=Y(z)-E_{a s}(z) . \tag{6.33}
\end{equation*}
$$

Substituting it into (6.32),

$$
\begin{equation*}
Y(z)-E_{a s}(z)=X(z)+\alpha z^{-1} Y(z)-z^{-1} E_{a s}(z) . \tag{6.34}
\end{equation*}
$$

Then (6.32) can be modified to

$$
\begin{equation*}
Y(z)=\frac{X(z)}{1-\alpha z^{-1}}+\frac{1-z^{-1}}{1-\alpha z^{-1}} E_{a s}(z) \tag{6.35}
\end{equation*}
$$

Therefore, the signal transfer function (STF) and the noise transfer function (NTF) of the HK-EFM1 are

$$
\begin{align*}
S T F_{H K} & =\frac{1}{1-\alpha z^{-1}}  \tag{6.36}\\
N T F_{H K} & =\frac{1-z^{-1}}{1-\alpha z^{-1}} \tag{6.37}
\end{align*}
$$

In comparison with the STF and NTF of the conventional EFM1, (6.15) and (6.16), there is a pole at $z=\alpha$. However, the value of $\alpha$ is such that the pole is close to the origin in the $z$-plane and hence does not affect the in-band operation of the modulator to any significant level.


Figure 6.18: The structure of the $H K-M A S H$.

The structure of the HK-MASH is shown in Fig. 6.18. The difference between the HK-MASH and the conventional MASH DDSM is that the HK-MASH utilises the HK-EFM1 introduced above instead of the conventional EFM1. Using (6.35), the expression for the first stage of HK-EFM1 in a HK-MASH is obtained:

$$
\begin{equation*}
Y_{1}(z)=\frac{X(z)}{1-\alpha z^{-1}}+\frac{1-z^{-1}}{1-\alpha z^{-1}} E_{a s 1}(z) . \tag{6.38}
\end{equation*}
$$

Combining (6.30), (6.38) is modified to

$$
\begin{equation*}
Y_{1}(z)=\frac{X(z)}{1-\alpha z^{-1}}-\frac{1-z^{-1}}{1-\alpha z^{-1}} E_{1}(z) . \tag{6.39}
\end{equation*}
$$

Similarly, for the other stages of HK-EFM1 are expressed as:

$$
\begin{align*}
Y_{2}(z)= & \frac{E_{1}(z)}{1-\alpha z^{-1}}-\frac{1-z^{-1}}{1-\alpha z^{-1}} E_{2}(z)  \tag{6.40}\\
& \vdots \\
Y_{l}(z)= & \frac{E_{l-1}(z)}{1-\alpha z^{-1}}-\frac{1-z^{-1}}{1-\alpha z^{-1}} E_{l}(z) . \tag{6.41}
\end{align*}
$$

From Fig. 6.18, it is derived that

$$
\begin{equation*}
Y(z)=Y_{1}(z)+Y_{2}(z)\left(1-z^{-1}\right)+\cdots+Y_{l}(z)\left(1-z^{-1}\right)^{l-1} . \tag{6.42}
\end{equation*}
$$

Substituting (6.39)-(6.41) into (6.42),

$$
\begin{equation*}
Y(z)=\frac{X(z)}{1-\alpha z^{-1}}-\frac{\left(1-z^{-1}\right)^{l}}{1-\alpha z^{-1}} E_{l}(z) \tag{6.43}
\end{equation*}
$$

Alternatively,

$$
\begin{equation*}
Y(z)=\frac{X(z)}{1-\alpha z^{-1}}+\frac{\left(1-z^{-1}\right)^{l}}{1-\alpha z^{-1}} E_{a s l}(z) \tag{6.44}
\end{equation*}
$$

### 6.4 Mathematical Analysis of the MASH DDSMs employing Multi-Moduli

A novel structure for the MASH DDSMs which employ Multi-Moduli (MM-MASH) is proposed in this section. The advantage of this structure will be that it increases the modulator period. It is proposed that the modulus of each quantizer is set as a different value from each other. Note that each quantizer has only one modulus. Furthermore, all of the moduli are co-prime numbers. The difference between a a prime number and co-prime numbers [70] is stated as follows:

1. A prime number is a natural number which has exactly two divisors: 1 and itself.
2. If the greatest common divisor of any two numbers is 1 , they are co-prime numbers. They do NOT have to be prime numbers, for example 8 and 9 are co-prime but 8 or 9 are not prime numbers.


Figure 6.19: MM-EFM1: The modified first-order error-feedback modulator used in the MM-MASH.

The structure of the modified EFM1 used in the MM-MASH (MM-EFM1) is shown in Fig. 6.19. The key difference between the proposed MM-MASH and the existing structures described in Section 6.3 is that the quantizer modulus of each EFM1 is different from each other. The quantizer modulus of the $i$ th stage EFM1 in an lth order MASH-DDSM is denoted by $M_{i}$, where $i \in\{1,2, \ldots, l\}$. It shall first be shown that the MM-MASH is an accurate modulator and that the use of a different modulus for each stage has not affected its output. Then the effect of the multi-moduli on the sequence length shall be investigated mathematically. In Section 6.4.1 and 6.4.2, a MM-MASH consisting of only first-order EFMs (MM-EFM1s) is mathematically analysed. The use of higher-order EFMs in the MM-MASH (MM-EFM) will be investigated in Section 6.4.3.

### 6.4.1 The Suitability of the Multi-Modulus MASH-DDSM

In a fractional- $N$ frequency synthesizer, the static frequency divider, shown in Fig. 6.7, is controlled by the average value of the delta-sigma modulator output, mean (y). For example, if the static frequency divider divides the output frequency of the frequency synthesizer by the factor 7.89, the average delta-sigma modulator output should be 0.89 , i.e., $\operatorname{mean}(y)=0.89$. The goal of this section is to show that in an MM-MASH, $\operatorname{mean}(y)$ is affected only by the quantizer modulus of the first stage EFM1, $M_{1}$, and is independent of the moduli in other stages. With this being true, having a multimodulus architecture does not affect the accuracy of the digital delta-sigma modulator. Hence, it is a suitable digital delta-sigma modulator. So it is required to prove:

$$
\begin{equation*}
\operatorname{mean}(y)=\frac{X}{M_{1}} . \tag{6.45}
\end{equation*}
$$

Proof. The structure of MM-MASH consists of only MM-EFM1 is shown in Fig. 6.20. As seen in Fig. 6.20, at the output of the last adder,

$$
\begin{align*}
v_{l-1}[1]= & y_{l-1}[1]+y_{l}[1]-y_{l}[0]  \tag{6.46}\\
v_{l-1}[2]= & y_{l-1}[2]+y_{l}[2]-y_{l}[1]  \tag{6.47}\\
& \vdots \\
v_{l-1}[N] & =y_{l-1}[N]+y_{l}[N]-y_{l}[N-1] \tag{6.48}
\end{align*}
$$

where $N$ is assumed as the sequence length of the complete MASH delta-sigma modulator. Adding all of the above equations yields:

$$
\begin{equation*}
\sum_{k=1}^{N} v_{l-1}[k]=\sum_{k=1}^{N} y_{l-1}[k]+\sum_{k=1}^{N} y_{l}[k]-\sum_{k=0}^{N-1} y_{l}[k] \tag{6.49}
\end{equation*}
$$

where the period of $y_{l}$ is assumed as $N_{l}$. As seen from Fig. 6.20, the output of the MASH modulator is obtained by simply summing and/or subtracting the output of each EFM1. Hence, the period of the MASH DDSM is the least common multiple of


Figure 6.20: The MM-MASH consists of only MM-EFM1.
the sequence length of each stage. In other words, $N$ is a multiple of $N_{i}$, where $N_{i}$ is the period of the $i$ th stage EFM1 and $i \in\{1,2, \ldots, l\}$. It follows that

$$
\begin{equation*}
\sum_{k=1}^{N} y_{l}[k]=\sum_{k=0}^{N-1} y_{l}[k] . \tag{6.50}
\end{equation*}
$$

Thus (6.49) becomes:

$$
\begin{equation*}
\sum_{k=1}^{N} v_{l-1}[k]=\sum_{k=1}^{N} y_{l-1}[k] . \tag{6.51}
\end{equation*}
$$

Similarly, each output of the other adders is obtained as:

$$
\begin{align*}
\sum_{k=1}^{N} v_{l-2}[k]= & \sum_{k=1}^{N} y_{l-2}[k]  \tag{6.52}\\
& \vdots  \tag{6.53}\\
\sum_{k=1}^{N} v_{2}[k]= & \sum_{k=1}^{N} y_{2}[k]  \tag{6.54}\\
\sum_{k=1}^{N} y[k]= & \sum_{k=1}^{N} y_{1}[k] .
\end{align*}
$$

Each side of (6.54) may be expressed as:

$$
\begin{align*}
& \sum_{k=1}^{N} y[k]=N \cdot \operatorname{mean}(y)  \tag{6.55}\\
& \sum_{k=1}^{N} y_{1}[k]=K \sum_{k=1}^{N_{1}} y_{1}[k] \tag{6.56}
\end{align*}
$$

where $N_{1}$ is the sequence length of $y_{1}, K$ is an integer and $N=K \cdot N_{1}$. Since $y_{1}$ is the output of a first-order delta-sigma modulator EFM1,

$$
\begin{align*}
\sum_{k=1}^{N_{1}} y_{1}[k] & =N_{1} \cdot \operatorname{mean}\left(y_{1}\right) \\
& =N_{1} \cdot \frac{X}{M_{1}} \tag{6.57}
\end{align*}
$$

On substitution of (6.57) into (6.56), the right-hand side of (6.54) becomes

$$
\begin{align*}
\sum_{k=1}^{N} y_{1}[k] & =K \cdot N_{1} \cdot \frac{X}{M_{1}} \\
& =N \cdot \frac{X}{M_{1}} \tag{6.58}
\end{align*}
$$

By substituting (6.55) and (6.58) into (6.54), the average value of the MASH DDSM output $y$ is determined as:

$$
\begin{equation*}
\operatorname{mean}(y)=\frac{X}{M_{1}} . \tag{6.59}
\end{equation*}
$$

### 6.4.2 The Effect of the Multi-Moduli on the Modulator Sequence Length

It is required to prove that the sequence length of the MASH modulator depends on the product of all the quantizer moduli. The expression for the lth order MASH DDSM sequence length is:

$$
\begin{equation*}
N=\frac{M_{1} \cdot M_{2} \cdot \ldots \cdot M_{l}}{\lambda} \tag{6.60}
\end{equation*}
$$

where $\lambda$ is a parameter to make $N$ the least common multiple of the sequence length of each stage $N_{i}$.

In addition, if the following two conditions are satisfied:
Cond1: $X$ and $M_{1}$ are co-prime numbers
Cond2: $\left\{M_{1}, M_{2}, \ldots, M_{l}\right\}$ are different co-prime numbers
then the sequence length of the MASH DDSM attains the maximum value:

$$
\begin{equation*}
N_{\max }=M_{1} \cdot M_{2} \cdot \ldots \cdot M_{l} . \tag{6.61}
\end{equation*}
$$

Proof. In the first-stage EFM1 shown in Fig. 6.12,

$$
\begin{align*}
e_{1}[1]= & u[1]-y_{1}[1] M_{1} \\
= & X+e_{1}[0]-y_{1}[1] M_{1}  \tag{6.62}\\
e_{1}[2]= & X+e_{1}[1]-y_{1}[2] M_{1}  \tag{6.63}\\
& \vdots  \tag{6.64}\\
e_{1}\left[N_{1}\right]= & X+e_{1}\left[N_{1}-1\right]-y_{1}\left[N_{1}\right] M_{1}
\end{align*}
$$

where $e_{1}[0]$ is the initial condition of the register. The sum of all of the above equations is:

$$
\begin{equation*}
\sum_{k=1}^{N_{1}} e_{1}[k]=N_{1} X+\sum_{k=0}^{N_{1}-1} e_{1}[k]-\sum_{k=1}^{N_{1}} y_{1}[k] M_{1} . \tag{6.65}
\end{equation*}
$$

Since in the steady state, the first EFM1 is periodic with a period $N_{1}$ [67],

$$
\begin{equation*}
\sum_{k=1}^{N_{1}} e_{1}[k]=\sum_{k=0}^{N_{1}-1} e_{1}[k] . \tag{6.66}
\end{equation*}
$$

Hence, (6.65) may be modified to

$$
\begin{equation*}
\sum_{k=1}^{N_{1}} y_{1}[k]=\frac{N_{1} X}{M_{1}} \tag{6.67}
\end{equation*}
$$

In practice, the input $\mathrm{DC} X$ is set as $0<X<M_{1}$. So in order to make the right side of (6.67) an integer, the minimum nonzero solution of $N_{1}$ has to be:

$$
\begin{equation*}
N_{1}=\frac{M_{1}}{\lambda_{1}} \tag{6.68}
\end{equation*}
$$

where $\lambda_{1}$ is the greatest common divisor of $M_{1}$ and $X$. If $M_{1}$ and $X$ are co-prime numbers, $\lambda_{1}$ equals to 1 .

If the process of (6.62)-(6.67) is repeated with the second EFM1, the sum of its output, which has a period $N_{2}$, is obtained as:

$$
\begin{equation*}
\sum_{k=1}^{N_{2}} y_{2}[k]=\frac{\sum_{k=1}^{N_{2}} e_{1}[k]}{M_{2}} \tag{6.69}
\end{equation*}
$$

If the relationship between the sequence length of the first and second stages is

$$
\begin{equation*}
N_{2}=K_{1} N_{1} \tag{6.70}
\end{equation*}
$$

(6.69) becomes

$$
\begin{equation*}
\sum_{k=1}^{N_{2}} y_{2}[k]=\frac{\sum_{k=1}^{K_{1} N_{1}} e_{1}[k]}{M_{2}} . \tag{6.71}
\end{equation*}
$$

Since $e_{1}$ is periodic with a sequence length $N_{1}$ [3],

$$
\begin{equation*}
\sum_{k=1}^{N_{2}} y_{2}[k]=\frac{K_{1} \sum_{k=1}^{N_{1}} e_{1}[k]}{M_{2}} \tag{6.72}
\end{equation*}
$$

where

$$
\begin{equation*}
\sum_{k=1}^{N_{1}} e_{1}[k]=N_{1} \cdot \operatorname{mean}\left(e_{1}\right) . \tag{6.73}
\end{equation*}
$$

Recalling (6.68),

$$
\begin{equation*}
\sum_{k=1}^{N_{1}} e_{1}[k]=\frac{M_{1} \cdot \operatorname{mean}\left(e_{1}\right)}{\lambda_{1}} . \tag{6.74}
\end{equation*}
$$

On substitution of (6.74) into (6.72), the following expression is obtained:

$$
\begin{equation*}
\sum_{k=1}^{N_{2}} y_{2}[k]=\frac{K_{1} \cdot M_{1} \cdot \operatorname{mean}\left(e_{1}\right)}{\lambda_{1} \cdot M_{2}} \tag{6.75}
\end{equation*}
$$

Normally, mean $\left(e_{1}\right)$ is a decimal fraction. However, if both sides of (6.74) are multiplied by $\lambda_{1}$, the result is:

$$
\begin{equation*}
\lambda_{1} \sum_{k=1}^{N_{1}} e_{1}[k]=M_{1} \cdot \operatorname{mean}\left(e_{1}\right) . \tag{6.76}
\end{equation*}
$$

Since $\lambda_{1} \sum_{k=1}^{N_{1}} e_{1}[k]$ has to be an integer, $M_{1} \cdot \operatorname{mean}\left(e_{1}\right)$ is always an integer.
Then the minimum solution of $K_{1}$ so that the right-hand-side of (6.75) is an integer is obtained as:

$$
\begin{equation*}
K_{1}=\frac{\lambda_{1} M_{2}}{\lambda_{2}} \tag{6.77}
\end{equation*}
$$

where $\lambda_{2}$ is the greatest common divisor of $\lambda_{1} M_{2}$ and $M_{1}$ mean $\left(e_{1}\right)$. Substituting (6.68) and (6.77) into (6.70), the sequence length of the second stage is:

$$
\begin{equation*}
N_{2}=\frac{M_{1} M_{2}}{\lambda_{2}} . \tag{6.78}
\end{equation*}
$$

If $M_{1}$ and $M_{2}$ are co-prime numbers, the greatest common divisor of $\lambda_{1} M_{2}$ and $M_{1}$ mean $\left(e_{1}\right)$ is $\lambda_{1}$, i.e., $\lambda_{2}=\lambda_{1}$. Hence,

$$
\begin{equation*}
N_{2}=\frac{M_{1} M_{2}}{\lambda_{1}} \tag{6.79}
\end{equation*}
$$

When $X$ and $M_{1}$ are also co-prime numbers, $\lambda_{1}$ equals to 1 . Thus the maximum sequence length for $y_{2}$ is obtained as:

$$
\begin{equation*}
N_{2 \_\max }=M_{1} M_{2} \text {. } \tag{6.80}
\end{equation*}
$$

Continuing in this manner, the sequence length of the $i$ th effective stage EFM1 in an $l$ th order MASH modulator is:

$$
\begin{equation*}
N_{i}=\frac{M_{1} M_{2} \ldots M_{i}}{\lambda_{i}} \tag{6.81}
\end{equation*}
$$

where $i \in\{1,2,3, \ldots, l\}$ and $\lambda_{i}$ is the maximum common divisor of $\lambda_{i-1} M_{i}$ and $M_{1} M_{2} \ldots M_{i-1} \operatorname{mean}\left(e_{i-1}\right)$. Note that when $i=1, \operatorname{mean}\left(e_{0}\right)=X$ and $\lambda_{0}=M_{0}=1$.

If $\left\{M_{1}, M_{2}, \ldots, M_{i}\right\}$ are co-prime numbers, $M_{i}$ and $\left(M_{1} \cdot M_{2} \cdot \ldots \cdot M_{i-1}\right)$ have to be co-prime numbers as well. Thus $\lambda_{i}=\lambda_{i-1}$. Since $M_{i-1}$ and $\left(M_{1} \cdot M_{2} \cdot \ldots \cdot M_{i-2}\right)$ are also co-prime numbers, $\lambda_{i-1}=\lambda_{i-2}$. Repeating this manner, it is follows that

$$
\begin{equation*}
\lambda_{i}=\lambda_{i-1}=\ldots=\lambda_{1} . \tag{6.82}
\end{equation*}
$$

Then the sequence length of the $i$ th EFM1 becomes

$$
\begin{equation*}
N_{i}=\frac{M_{1} M_{2} \ldots M_{i}}{\lambda_{1}} \tag{6.83}
\end{equation*}
$$

where $\lambda_{1}$ is the greatest common divisor of $X$ and $M_{1}$. In practice, if the input $X$ and $M_{1}$ are set as co-prime numbers, the maximum sequence length of the $i$ th stage EFM1 is:

$$
\begin{equation*}
N_{i-\max }=M_{1} M_{2} \ldots M_{i} . \tag{6.84}
\end{equation*}
$$

Since $N$ is the least common multiple of $\left\{N_{1}, N_{2}, \ldots, N_{l}\right\}$, as was stated in Section 6.4.1, the sequence length of the MASH DDSM is obtained as

$$
\begin{equation*}
N=\frac{M_{1} \cdot M_{2} \cdot \ldots \cdot M_{l}}{\lambda} \tag{6.85}
\end{equation*}
$$

where $\lambda$ is the least common multiple of $\left\{\lambda_{1}, \lambda_{2}, \ldots, \lambda_{l}\right\}$.
When $\left\{M_{1}, M_{2}, \ldots, M_{l}\right\}$ are co-prime numbers, (6.82) is true. Then

$$
\begin{equation*}
N=\frac{M_{1} \cdot M_{2} \cdot \ldots \cdot M_{l}}{\lambda_{1}} \tag{6.86}
\end{equation*}
$$

In addition, if $X$ and $M_{1}$ are co-prime numbers as well, $\lambda_{1}$ becomes 1 . Thus the maximum sequence length is:

$$
\begin{equation*}
N_{\max }=M_{1} \cdot M_{2} \cdot \ldots \cdot M_{l} . \tag{6.87}
\end{equation*}
$$

### 6.4.3 A Mathematical Analysis of Higher-Order MM-EFMs

The structure of an $m$ th order MM-EFM (MM-EFM $m$ ) is illustrated in Fig. 6.21. It is similar to the conventional EFM except that the quantizer modulus is $M_{i}$ instead of $M$, since the modulus in each EFM is assumed to be different from each other. The symbols $R_{1}, R_{2}, \ldots, R_{m}$ and $A_{1}, A_{2}, \ldots, A_{m}$ represent the initial conditions and gains of the $i$ th registers, respectively, where $i \in\{1,2, \ldots, m\}$. The noise transfer function is

$$
\begin{equation*}
N T F=\frac{(Z-1)^{m}}{Z^{m}} \tag{6.88}
\end{equation*}
$$

The value of the gains $A_{i}$ are:

$$
\begin{equation*}
\sum_{i=1}^{m} A_{i} Z^{-i}=1-N T F \tag{6.89}
\end{equation*}
$$

Since the EFM2 is the most popular higher-order EFM in practice [14], the calculation of its sequence length shall be selected as an example. The structure of EFM2 is shown in Fig. 6.22. The coefficients $\left[A_{1}, A_{2}\right]$ are obtained from (6.89) as $[2,-1]$. The $p$ th stage of the MASH DDSM is considered. Thus the state variable $u$ is dependent on the output of previous stage:

$$
\begin{equation*}
u_{p}[1]=e_{p-1}[1]+2 R_{1}-R_{2} . \tag{6.90}
\end{equation*}
$$

Using (6.90), the error is expressed as

$$
\begin{align*}
e_{p}[1] & =u_{p}[1]-M \cdot y_{p}[1] \\
& =e_{p-1}[1]+2 R_{1}-R_{2}-M_{p} \cdot y_{p}[1] \tag{6.91}
\end{align*}
$$

After several time steps, the values of the registers are changed from their initial conditions to $e_{p}$, since steady-state has been established. Then

$$
\begin{align*}
e_{p}[2]= & e_{p-1}[2]+2 e_{p}[1]-R_{2}-M_{p} \cdot y_{p}[2]  \tag{6.92}\\
e_{p}[3]= & e_{p-1}[3]+2 e_{p}[2]-e_{p}[1]-M_{p} \cdot y_{p}[3]  \tag{6.93}\\
& \vdots \\
e_{p}\left[N_{p}\right]= & e_{p-1}\left[N_{p}\right]+2 e_{p}\left[N_{p}-1\right]-e_{p}\left[N_{p}-2\right]-M_{p} \cdot y_{p}[k] . \tag{6.94}
\end{align*}
$$



Figure 6.21: Higher order error-feedback modulator.


Figure 6.22: Structure of the 2 nd order error-feedback modulator.

The sum of all of the above equations (6.91)-(6.94) is

$$
\begin{align*}
\sum_{k=1}^{N_{p}} e_{p}[k]= & \sum_{k=1}^{N_{p}} e_{p-1}[k]+2 \sum_{k=0}^{N_{p-1}} e_{p}[k] \\
& -\sum_{k=-1}^{N_{p-2}} e_{p}[k]-M_{p} \cdot \sum_{k=1}^{N_{p}} y_{p}[k] \tag{6.95}
\end{align*}
$$

Since $e_{p}$ is periodic with the period $N_{p}[66]$ in the steady state,

$$
\begin{equation*}
\sum_{k=1}^{N_{p}} e_{p}[k]=\sum_{k=0}^{N_{p-1}} e_{p}[k]=\sum_{k=-1}^{N_{p-2}} e_{p}[k] \tag{6.96}
\end{equation*}
$$

Then (6.95) becomes

$$
\begin{equation*}
\sum_{k=1}^{N_{p}} y_{p}=\frac{1}{M_{p}} \sum_{k=1}^{N_{p}} e_{p-1} \tag{6.97}
\end{equation*}
$$

If the relationship between the sequence length of the $p$ th and $(p-1)$ th stage EFM is

$$
\begin{equation*}
N_{p}=K \cdot N_{p-1} \tag{6.98}
\end{equation*}
$$

(6.97) is modified to

$$
\begin{equation*}
\sum_{k=1}^{N_{p}} y_{p}=\frac{K}{M_{p}} \sum_{k=1}^{N_{p-1}} e_{p-1} \tag{6.99}
\end{equation*}
$$

In comparison with (6.72), it is obtained that the sequence length of the EFM2 is same as that of the EFM1, if they both constitute the $p$ th stage in a lth stage DDSM [16]. In other words and the crucial point, the sequence length of a MASH DDSM does not depend on the order level, but on the number of EFMs. For example, the period of MM-MASH 1-2 is

$$
\begin{equation*}
N=\frac{M_{1} \cdot M_{2}}{\lambda} \tag{6.100}
\end{equation*}
$$

where $\lambda$ is a parameter to make $N$ the least common multiple of $N_{1}$ and $N_{2}$. If the two conditions stated in Section 6.4.2 are satisfied, the maximum sequence length is

$$
\begin{equation*}
N=M_{1} \cdot M_{2} . \tag{6.101}
\end{equation*}
$$

This is same as the period of the MM-MASH 1-1.

### 6.5 The Proposed Structure of the MASH DDSM

The novel structure for the MASH digital delta-sigma modulator (DDSM) employing multi-moduli (MM-MASH) shall now be confirmed in this section. The structure of a MM-MASH consisting of $l$ stages of MM-EFM is illustrated in Fig. 6.23. Note that the block MM-EFM represents either a first-order or a higher order MM-EFM. The crucial points to note about this structure are:

1. Every MM-EFM has a different modulus
2. Each MM-EFM still only has ONE quantizer modulus.


Figure 6.23: The structure of the MM-MASH.
$M_{1}$ is set as a prime number around $2^{n_{0}}$. This is to make $X$ and $M_{1}$ always coprime numbers and therefore satisfy the first condition Cond1 stated at the begin of Section 6.4.2. This condition must be satisfied to maximise the sequence length of the MASH DDSM and to make the sequence length independent of the value of input. In order to maintain the modulator output accuracy, the value of the input $\mathrm{DC} X$ is adjusted to

$$
\begin{equation*}
X=M_{1} \cdot \operatorname{mean}(y) \tag{6.102}
\end{equation*}
$$

where mean $(y)$ is the required output to control the static frequency divider in a fractional- $N$ frequency synthesizer.

In an $l$ th order MM-MASH, there are $l$ co-prime numbers around $2^{n_{0}}$ that need to be found in order to satisfy the second condition Cond2. The higher the modulator order, the greater the difficulty in finding suitable values for these moduli. Fortunately, the most popular MASH DDSM in modern communication systems is MASH 1-11 [14]. Note that all of the quantizer moduli should be chosen no bigger than $2^{n_{0}}$ to avoid necessitating extra hardware. Some quantizer moduli chosen by the author for MM-MASH 1-1-1 are given in Table 6.5.

Table 6.5: Some sample moduli of the 3rd order MM-MASH.

| Word length | $M_{1}$ | $M_{2}$ | $M_{3}$ |
| :---: | :---: | :---: | :---: |
| 5 bit | 31 | 32 | 29 |
| 6 bit | 61 | 64 | 63 |
| 7 bit | 127 | 128 | 125 |
| 8 bit | 251 | 256 | 255 |
| 9 bit | 509 | 512 | 511 |
| 10 bit | 1021 | 1024 | 1023 |
| 11 bit | 2039 | 2048 | 2047 |

### 6.6 The Simulation Results

All of the models of the EFM1 and MASH DDSM are built and simulated in Simulink. They are shown in Appendix E. All of the simulations in this section are performed with the input constant $X=1$. The simulations confirm that the average value of the output mean $(y)$ equals to $\frac{X}{M_{1}}$. The sequence length of the MASH DDSM is determined using the autocorrelation function [14], which is defined as

$$
\begin{equation*}
R_{x x}(m)=\frac{1}{N} \sum_{n=1}^{N-m+1} x(n) x(n+m-1) \tag{6.103}
\end{equation*}
$$

where $m=1,2, \ldots, N$ and $N$ is the number of samples. As a typical example, Fig. 6.24 shows that the sequence length of a 3rd-order 5-bit MM-MASH is 28768 and this equals $M_{1} \cdot M_{2} \cdot M_{3}$ as given in Table 6.5. The sequence length is only 64 from a 3rd-order 5-bit conventional MASH. The sequence lengths of the HK-MASH and the MM-MASH are compared in Table 6.6, since the HK-MASH has the longest sequence length of current structures of DDSMs. The MM-MASH achieves a longer sequence when the word lengths are $6,8,9,10$ and 11. Thus the choice between MM-MASH and HK-MASH depends on the word length. However, the MM-MASH produces a longer sequence length, if $M_{1} \neq 2^{n_{0}}-1$, where $M_{1}$ is the biggest prime number below $2^{n_{0}}$ as described in Section 6.5. Hence, it has a greater probability of producing a longer sequence length.


Figure 6.24: The autocorrelation result for the 5-bit MM-MASH 1-1-1.

Table 6.6: A comparison of the sequence lengths for the HK-MASH and MM-MASH.

| Word length | HK-MASH | MM-MASH | Difference |
| :---: | :---: | :---: | :---: |
| 5 bit | 32768 | 28768 | -4000 |
| 6 bit | $0.227 \times 10^{6}$ | $0.246 \times 10^{6}$ | $+0.019 \times 10^{6}$ |
| 7 bit | $2.048 \times 10^{6}$ | $2.032 \times 10^{6}$ | $-0.016 \times 10^{6}$ |
| 8 bit | $15.81 \times 10^{6}$ | $16.39 \times 10^{6}$ | $+0.58 \times 10^{6}$ |
| 9 bit | $131.87 \times 10^{6}$ | $133.17 \times 10^{6}$ | $+1.3 \times 10^{6}$ |
| 10 bit | $1.06 \times 10^{9}$ | $1.07 \times 10^{9}$ | $+10 \times 10^{6}$ |
| 11 bit | $8.48 \times 10^{9}$ | $8.55 \times 10^{9}$ | $+70 \times 10^{6}$ |

The ultimate goal in the novel design is to improve the quality of the power spectrum, i.e to reduce noise power in the useful band. Hence, the power spectrum is now examined. The power spectral density for 5-11 bit MASH DDSMs are compared and in all cases the performance of the MM-MASH is better. As an example, the power spectral density [71] of the MM-MASH 1-1-1 and dithered conventional MASH 1-1-1 are compared in Figs. 6.25-6.31. The dithering is applied to the DDSM by adding a 1-bit pseudorandom dither sequence to the LSB of the dc input [65]. It is evident from the figures that the MM-MASH is significantly more effective than the conventional MASH DDSM at lower frequencies.

Also, the power spectral density of the 11-bit MM-MASH 1-1-1 and 11-bit HKMASH 1-1-1 are compared in Fig. 6.32, since it produces the biggest sequence length difference as shown in Table 6.6. The noise power performance of MM-MASH is a little better than that of HK-MASH. Thus the noise power difference results from the MM-MASH and HK-MASH with a smaller word length can be even neglected. However, MM-MASH is a better choice in high-end products, since they desire the noise to be the lower the better.

The power spectral density [71] of the MM-MASH 1-1-1 and MM-MASH 1-2 is compared in Fig. 6.33-6.39. Both of the DDSMs are 3rd order. It is evident from the figure that the MM-MASH 1-1-1 has a better noise performance than the MM-MASH

1-2. This is expected as the sequence length of the former is longer than that of the latter as shown in Section 6.4.3. However, the MM-MASH 1-2 has the advantage of less hardware requirements and so a balance between hardware cost and noise performance is required in the selection of the most suitable structure.

### 6.7 Conclusions

Digital delta-sigma modulators have been widely used for years. The benefit of their use is that the quantisation noise is pushed to the higher frequencies out of the useful bandwidth. Then the noise can be simply eliminated by a lowpass filter. Now with DDSMs, the output sequence length is periodic and short. This is disadvantageous as it results in spurs in the useful spectrum. To avoid this, one needs to whiten the quantization noise. This can be done by dithering. However, in order to avoid the extra noise and hardware requirements resulting from the dithering method, several novel DDSM structures have been proposed to lengthen the sequence length. In this chapter, it has been proved that MM-MASH produces the maximum sequence length for most input word lengths. Expressions for the sequence length of the modulator are derived. The simulation results confirm the results. Higher-order EFMs are also investigated as previous research indicates that they yield hardware savings [72]. However, it has been proved in this thesis that they are not able to increase the sequence length. Hence, as a final note, a balance must be made between the noise performance and the hardware requirements.


Figure 6.25: The power spectral density of the dithered 5-bit conventional MASH DDSM and non-dithered 5-bit MM-MASH.


Figure 6.26: The power spectral density of the dithered 6-bit conventional MASH DDSM and non-dithered 6-bit MM-MASH.


Figure 6.27: The power spectral density of the dithered 7-bit conventional MASH DDSM and non-dithered 7-bit MM-MASH.


Figure 6.28: The power spectral density of the dithered 8-bit conventional MASH DDSM and non-dithered 8-bit MM-MASH.


Figure 6.29: The power spectral density of the dithered 9-bit conventional MASH DDSM and non-dithered 9-bit MM-MASH.


Figure 6.30: The power spectral density of the dithered 10-bit conventional MASH DDSM and non-dithered 10-bit MM-MASH.


Figure 6.31: The power spectral density of the dithered 11-bit conventional MASH DDSM and non-dithered 11-bit MM-MASH.


Figure 6.32: The power spectral density of the 11-bit HK-MASH DDSM and 11-bit MM-MASH.


Figure 6.33: The power spectral density of different structures of 3 rd order 5-bit MMMASH.


Figure 6.34: The power spectral density of different structures of 3rd order 6-bit MMMASH.


Figure 6.35: The power spectral density of different structures of 3 rd order 7 -bit MMMASH.


Figure 6.36: The power spectral density of different structures of 3 rd order 8-bit MMMASH.


Figure 6.37: The power spectral density of different structures of 3rd order 9-bit MMMASH.


Figure 6.38: The power spectral density of different structures of 3 rd order 10-bit MM-MASH.


Figure 6.39: The power spectral density of different structures of 3rd order 11-bit MM-MASH.

## Chapter 7

## Conclusions

Several novel CAD algorithms are proposed in this dissertation. The Padé-based approximation is proposed for the numerical integration of stiff differential equations. Illustrative examples have shown that the Padé-based techniques permit a significant increase in step-size when compared to a traditional predictor-corrector such as the Adams-Moulton technique. Both single-step and multi-step methods are proposed. Multi-step methods have the advantage of obviating the need for the evaluation of highorder derivatives. All of the techniques are particularly suitable for highly nonlinear systems as there is no need for the solution of a non-linear set of algebraic equations at each time step.

The Filon-type methods are explored for systems subjected to high-frequency signals. Numerical examples confirm the significant potential of Filon-type methods in this setting. Futher research is required into issues of implementation and software design, as well as detailed comparison with existing software for realistic differential equations originating in RF and communications engineering.

Two simulation strategies have been presented for the determination of the locking range of an ILFD. The first strategy involves the use of the WaMPDE in conjunction with the bisection method to identify external frequencies to which the oscillator locks. The WaMPDE enables identification of the natural frequency of the oscillator. Hence,
by examining the Jacobian matrix used in the transient evolution to steady state, it is possible to ascertain when the system becomes independent of the local frequency and can then be deemed locked. In addition, a technique for estimating the locking ranges may be obtained using linear extrapolation if an accurate result is not necessary. Results confirm the efficacy of the approaches. The second technique is based on the Multiple-Phase-Condition Envelope Following (MPCENV) algorithm. The periods of the carrier and envelope are set as variables. In order to solve for them, two extra conditions are required. The output frequencies of the ILFD corresponding to the different injected frequencies are then obtained. Thus the Devil's staircase can be plotted from the obtained data. Results for an LC-oscillator based ILFD confirm its efficacy. Computer simulation for the determination of the locking range is advantageous in avoiding the need for time-consuming experiments. It also greatly aids in design work involving ILFDs and their use as lower power frequency dividers in PLLs for wireless systems.

A novel design methodology for a MASH DDSM aimed at increasing the sequence length is introduced as well. Increasing the sequence length improves the noise performance of DDSMs. The proposed method employs different moduli in each stage of the EFM1. It is proven that the multi-modulus architecture is suitable because the output of the MASH modulator is only dependent on the quantizer modulus of the first stage EFM1 and independent of the others. Hence, having different moduli does not affect the output of the modulator. The expressions for the sequence length of the EFM1 of each stage and for the complete MASH DDSM are derived. There are two conditions given that must be satisfied to yield the maximum modulator period. A novel structure for the MASH digital delta-sigma modulator employing the multi-moduli (MM-MASH) is proposed. In most cases, the MM-MASHs produce an increased sequence length when compared to that of the current best DDSM, HK-MASH. Both of the output accuracy and the predicted sequence length of the MM-MASH are validated by simulation. The power spectral density spectra confirm that the proposed modulator
architecture is, in most cases, more effective than the HK-MASH method at moving noise from the lower frequencies.

Future work will involve investigation of ring-oscillator based injection-locked frequency dividers. In particular, expressions for the locking range and the conditions for locking will be explored. Deficiencies in previous work such as that by Lai [73] will be addressed.

## Appendix A

## $A$-stability and $L$-stability of Padé

## approximations

In order to examine the stability of Padé approximations, consider the following standard test differential equation:

$$
\begin{equation*}
y^{\prime}=\lambda y \tag{A.1}
\end{equation*}
$$

where $\lambda<0$. Now, consider the second-order Padé approximation where $h$ is the time-step and $y_{k}$ is the estimate of the solution $y$ at time $t_{k}$ :

$$
\begin{equation*}
y_{k+1}=\frac{1+(h \lambda / 3)}{1-(2 h \lambda / 3)+\left((h \lambda)^{2} / 6\right)} y_{k} . \tag{A.2}
\end{equation*}
$$

The amplification factor is:

$$
\begin{equation*}
R(\lambda h)=\frac{1+(h \lambda / 3)}{1-(2 h \lambda / 3)+\left((h \lambda)^{2} / 6\right)} . \tag{A.3}
\end{equation*}
$$

As $\forall h \lambda<0$, it is obtained

$$
\begin{equation*}
|R(\lambda h)|<1 \tag{A.4}
\end{equation*}
$$

Thus the method in equation (A.2) is $A$-stable. Similarly, higher-order Padé methods are A-stable $\forall h \lambda<0$.

In addition, the Pade method in (3.1) is $L$-stable if $m<n$. A numerical method is $L$-stable if, in addition to $A$-stability, when applied to the scalar test equation (A.1) with $\Re(\lambda)<0,|R(\lambda h)| \rightarrow 0$ as $\Re(\lambda h) \rightarrow-\infty$ [23]. Examining equation (3.3), it is evident that $\lim _{\Re(\lambda h) \rightarrow-\infty}|R(\lambda h)|=0$. This is true for higher-order methods if $m<n$ and $\forall h \lambda<0$.

## Appendix B

## Richardson extrapolation and

## improvement of accuracy

Consider the test equation:

$$
\begin{equation*}
y^{\prime}=\lambda y \tag{B.1}
\end{equation*}
$$

where $\lambda<0$. Also, consider the first-order Padé approximation:

$$
\begin{equation*}
y_{k+1}=\frac{1}{1-h \lambda} y_{k} . \tag{B.2}
\end{equation*}
$$

Let $y_{h}$ be the estimate of $y_{k+1}$ obtained with a time-step of $h$. Let $y_{h} / 2$ be the estimate obtained using a time-step of $h / 2$. The second-order error in $y_{h}$ is $E_{h}=\lambda^{2} / 2$. The second-order error in $y_{h / 2}$ is $E_{h / 2}=\lambda^{2} / 8$. Now consider:

$$
\begin{align*}
& y_{e x}=y_{h}+E_{h} h^{2}+O\left(h^{3}\right)  \tag{B.3}\\
& y_{e x}=y_{h / 2}+E_{h / 2} h^{2}+O\left(h^{3}\right) \tag{B.4}
\end{align*}
$$

From (B.3) and (B.4), the second-order errors may be eliminated by evaluating:

$$
\begin{equation*}
y_{e x}=\frac{4 y_{h / 2}-y_{h}}{3}+O\left(h^{3}\right) . \tag{B.5}
\end{equation*}
$$

In general, for an $(L-1)$ th-order method, Richardson extrapolation eliminates $L$ th errors as follows:

$$
\begin{equation*}
y_{e x}=\frac{2^{L} y_{h / 2}-y_{h}}{2^{L}-1}+O\left(h^{L+2}\right) . \tag{B.6}
\end{equation*}
$$

## Appendix C

## MESFET amplifier details



Figure C.1: Schematic of the MESFET amplifier.

$$
\begin{align*}
& \frac{d v_{g s}}{d t}=\frac{1}{C_{g s}}\left[\frac{v_{g d}-v_{g s}+v_{d s}+R_{g d i}\left(i_{s}-i_{d}-i_{g s}-i_{g d}\right)}{R_{g s i}+R_{g d i}}\right]  \tag{C.1}\\
& \frac{d v_{g d}}{d t}=\frac{1}{C_{g d}}\left[\frac{v_{g s}-v_{g d}-v_{d s}+R_{g s i}\left(i_{s}-i_{d}-i_{g s}-i_{g d}\right)}{R_{g s i}+R_{g d i}}\right] \tag{C.2}
\end{align*}
$$

$$
\begin{gather*}
\frac{d v_{d s}}{d t}=\frac{1}{C_{d s}}\left[i_{s}-\frac{v_{g d}-v_{g s}+v_{d s}+R_{g d i}\left(i_{s}-i_{d}-i_{g s}-i_{g d}\right)}{R_{g s i}+R_{g d i}}-i_{g s}-i_{d s}-\frac{v_{d s}-v_{x x}}{R_{x}}\right]  \tag{C.3}\\
\frac{d v_{x x}}{d t}=\frac{1}{C_{x}}\left[\frac{v_{d s}-v_{x x}}{R_{x}}\right] \tag{C.4}
\end{gather*}
$$

Let:

$$
\begin{gather*}
v_{d x}=\frac{L_{g} L_{s}\left(v_{d}-R_{d} i_{d}\right)+L_{d} L_{s}\left(v_{g}-v_{g d}-R_{g} i_{g}\right)+L_{d} L_{g}\left(R_{s} i_{s}+v_{d s}\right)}{L_{g} L_{s}+L_{d} L_{s}+L_{d} L_{g}}  \tag{C.5}\\
\frac{d i_{s}}{d t}=\frac{1}{L_{s}}\left(v_{d x}-v_{d s}-R_{s} i_{s}\right)  \tag{C.6}\\
\frac{d i_{d}}{d t}=\frac{1}{L_{d}}\left(v_{d}-v_{d x}-R_{d} i_{d}\right) \tag{C.7}
\end{gather*}
$$

$$
\begin{equation*}
\frac{d v_{g}}{d t}=\frac{\left(C_{p g d}+C_{p d}\right)\left[\left(e_{g}-v_{g}\right) G_{s}-i_{\lg }-i_{g}\right]+C_{p g d}\left(-G_{l} v_{o}-i_{l d}-i_{d}\right)}{C_{p g} C_{p d}+C_{p g d}\left(C_{p g}+C_{p d}\right)} \tag{C.8}
\end{equation*}
$$

$$
\begin{equation*}
\frac{d v_{d}}{d t}=\frac{C_{p g d}\left[\left(e_{g}-v_{g}\right) G_{s}-i_{\lg }-i_{g}\right]+\left(C_{p g d}+C_{p d}\right)\left(-G_{l} v_{o}-i_{l d}-i_{d}\right)}{C_{p g} C_{p d}+C_{p g d}\left(C_{p g}+C_{p d}\right)} \tag{C.9}
\end{equation*}
$$

$$
\begin{equation*}
\frac{d i_{\mathrm{lg}}}{d t}=\frac{v_{g}-R_{g g} i_{\mathrm{lg}}-V_{G G}}{L_{b g}} \tag{C.10}
\end{equation*}
$$

$$
\begin{equation*}
\frac{d i_{1 g}}{d t}=\frac{v_{d}-R_{d d} i_{1 g}-V_{D D}}{L_{b d}} \tag{C.11}
\end{equation*}
$$

$$
\begin{equation*}
i_{g s}=\text { IS_GS }\left[\mathrm{e}^{\frac{\mathrm{qvgs}_{\mathrm{g}}}{\mathrm{NkT}}}-1\right] \tag{C.12}
\end{equation*}
$$

$$
\begin{equation*}
i_{g d}=\text { IS_GS }\left[\mathrm{e}^{\left.\frac{\left.\mathrm{q(vgss-vg}_{\mathrm{gd}}\right)}{\mathrm{NkT}}-1\right]}\right. \tag{C.13}
\end{equation*}
$$

The Curtice-Ettenberg model is employed for the MESFET shown in Fig. C.1. Its details are as follows:

$$
\begin{gather*}
v_{1}= \begin{cases}v_{g s}\left(1+\beta\left(v_{\text {out } 0}-v_{d s}\right)\right), & v_{d s} \geq 0 \\
v_{g d}\left(1+\beta\left(v_{\text {out } 0}-v_{d s}\right)\right), & v_{d s}<0\end{cases}  \tag{C.14}\\
i_{d s 0}=\left\{\begin{array}{cl}
A_{0}+v_{1}\left(A_{1}+v_{1}\left(A_{2}+v_{1} A_{3}\right)\right), & v_{1} \geq v_{P \mathrm{MAX}} \\
A_{0}+v_{P \operatorname{MAX}}\left(A_{1}+v_{P \operatorname{MAX}}\left(A_{2}+v_{P \operatorname{MAX}} A_{3}\right)\right), & v_{1}<v_{P \mathrm{MAX}}
\end{array}\right. \tag{C.15}
\end{gather*}
$$

where

$$
\begin{equation*}
v_{P \mathrm{MAX}}=\frac{-2 A_{2}+\sqrt{4 A_{2}^{2}-12 A_{1} A_{3}}}{6 A_{3}} \tag{C.16}
\end{equation*}
$$

If $v_{d s}>0$ and $i_{d s 0}<0$

$$
\begin{equation*}
i_{d s}=0 \tag{C.17}
\end{equation*}
$$

else

$$
\begin{align*}
\chi_{2} & =-2 \gamma v_{d s} \\
z_{1} & =e^{\chi_{2}}  \tag{C.18}\\
i_{d s} & =i_{d s 0} \frac{1-z_{1}}{1+z_{1}}
\end{align*}
$$

Let $\phi=F_{C} \cdot V_{B I}$. Then the capacitances $C_{g s}$ and $C_{g d}$ are given as:

$$
\begin{align*}
& C_{g s}=\left\{\begin{array}{cc}
\frac{C_{g s 0}}{\sqrt{1-\frac{v_{g s}}{V_{B I}}}} & v_{g s}<\phi \\
\frac{C_{g s 0}\left(1-1.5 F_{C}+0.5 \frac{v_{g s}}{V_{B I}}\right)}{\left(1-F_{C}\right)^{\frac{3}{2}}} & v_{g s} \geq \phi
\end{array}\right.  \tag{C.19}\\
& C_{g d}=\left\{\begin{array}{cl}
\frac{C_{g d 0}}{\sqrt{1-\frac{v_{g d}}{V_{B I}}}} & v_{g d}<\phi \\
\frac{C_{g d 0}\left(1-1.5 F_{C}+0.5 \frac{v_{g d}}{V_{B I}}\right)}{\left(1-F_{C}\right)^{\frac{3}{2}}} & v_{g d} \geq \phi
\end{array}\right. \tag{C.20}
\end{align*}
$$

The following parameters are used:

$$
\begin{align*}
& G_{S}=0.02 \mathrm{~s}  \tag{C.21}\\
& G_{l}=0.02 \mathrm{~s}  \tag{C.22}\\
& L_{b g}=100 \mathrm{nH}  \tag{C.23}\\
& L_{b d}=100 \mathrm{nH}  \tag{C.24}\\
& R_{g g}=R_{d d}=0  \tag{C.25}\\
& C_{p g}=0.15 p F  \tag{C.26}\\
& C_{p d}=0.15 p F  \tag{C.27}\\
& L_{g}=0.35 \mathrm{nH}  \tag{C.28}\\
& L_{d}=0.35 \mathrm{nH}  \tag{C.29}\\
& L_{s}=0.35 \mathrm{nH}  \tag{C.30}\\
& R_{g}=7.21 \mathrm{~W}  \tag{C.31}\\
& R_{d}=5.0686 \mathrm{~W}  \tag{C.32}\\
& R_{s}=3.6953 \mathrm{~W}  \tag{C.33}\\
& C_{p g d}=0.01 p F  \tag{C.34}\\
& R_{g d i}=5.5 \mathrm{~W}  \tag{C.35}\\
& R_{g s i}=3.18 \mathrm{~W}  \tag{C.36}\\
& T=300.15 \mathrm{~K}  \tag{C.37}\\
& I S \_G S=1.0 \times 10^{-14} \mathrm{~A}  \tag{C.38}\\
& I S \_G D=0.0  \tag{C.39}\\
& N=1.0  \tag{C.40}\\
& V_{B I}=1.5 \mathrm{~V}  \tag{C.41}\\
&
\end{align*}
$$

$$
\begin{align*}
C_{g s 0} & =0.7243 p F  \tag{C.42}\\
C_{g d 0} & =0.0197 p F  \tag{C.43}\\
F_{C} & =0.5  \tag{C.44}\\
C_{x} & =0.19 p F  \tag{C.45}\\
R_{x} & =800 W  \tag{C.46}\\
C_{d s} & =0.1037 p F  \tag{C.47}\\
b & =0.05396 / V  \tag{C.48}\\
g & =7.680735 / V  \tag{C.49}\\
V_{\text {out } 0} & =19.996182 V  \tag{C.50}\\
A_{0} & =0.17229 A  \tag{C.51}\\
A_{1} & =0.093461 A / V  \tag{C.52}\\
A_{2} & =-0.053499 A / V^{2}  \tag{C.53}\\
A_{3} & =-0.028237 A / V^{3} \tag{C.54}
\end{align*}
$$

## Appendix D

## Schematics of the diode rectifier

## circuit used in ADS



Figure D.1: The diode rectifier circuit with amplitude modulated input signal.


Figure D.2: The diode rectifier circuit with digitally modulated input signal.

## Appendix E

## Models of DDSMs used in Simulink



Figure E.1: The model of EFM1 used in Simulink.


Figure E.2: The model of HK-EFM1 used in Simulink.


Figure E.3: The model of EFM2 used in Simulink.


Figure E.4: The model of MASH 1-1-1 used in Simulink.


Figure E.5: The model of MASH 1-2 used in Simulink.

## Appendix F

## List of relevant publications

## Journal Papers:

[27] M. Condon, E. Dautbegovic and T. Xu, "Novel Padé-based algorithms for numerical integration of ODEs," COMPEL: The International Journal for Computations and Mathematics in Electrical and Electronic Engineering, vol. 27, issue 6, 2008, pp. 1402-1417.
[74] M. Condon, A. Deaño, A.Iserles, K. Maczynski and T. Xu, "On highly oscillatory problems arising in electronic engineering," Accepted by COMPEL: The International Journal for Computations and Mathematics in Electrical and Electronic Engineering.

## Conference Papers:

[16] T. Xu and M. Condon, "Comparative Study of the MASH Digital Delta-Sigma Modulators," in Proc. IEEE Ph.D. Research in Microelectronics and Electronics, Cork, Ireland, July 2009, pp. 196-199.
[15] T. Xu and M. Condon, "Design methodology for a maximum sequence length MASH digital delta-sigma modulator," in Proc. World Congress on Engineering, London, U.K., July 2009.
[55] T. Xu and M. Condon, "Accurate simulation of the Devil's staircase of an injectionlocked frequency divider," in Proc. International Conference on Scientific Computing in Electrical Engineering, Espoo, Finland, Sept. 2008, pp. 105-106.
[54] T. Xu and M. Condon, "An effective method for the determination of the locking range of an injection-locked frequency divider," in Proc. Emerging Trends in Wireless Communications, Dublin, Ireland, April 2008, pp. 47-50.
[53] C. E. Christoffersen, M. Condon and Tao Xu, "A new method for the determination of the locking range of oscillators," in Proc. European Conference on Circuit Theory and Design, Seville, Spain, Aug. 2007, pp. 575-578.
[7] T. Xu, Z. Ye and M. P. Kennedy, "Mathematical analysis of injection-locked frequency dividers," in Proc. International Symposium on Nonlinear Theory and its Applications, Bologna, Italy, Sept. 2006, pp. 639-642.
[8] Z. Ye, T. Xu and M. P. Kennedy, "Locking range analysis for injection-locked frequency dividers," in Proc. International Symposium on Circuits and Systems, Island of Kos, Greece, May 2006, pp. 4070-4073.

## Bibliography

[1] Circuit Envelope Simulation, Agilent Technologies, Santa Clara, California, Aug. 2005.
[2] R. Adler, "A study of locking phenomena in oscillators," Proc. of the IEEE, vol. 61, no. 10, pp. 1380-1385, Oct. 1973.
[3] K. Hosseini and M. P. Kennedy, "Maximum sequence length mash digital deltasigma modulators," IEEE Trans. on Circuits and Systems-I: Regular Papers, vol. 54, no. 12, pp. 2628-2638, Dec. 2007.
[4] Electronic design automation. [Online]. Available: http://en.wikipedia.org/wiki/ Electronic_design_automation
[5] M. Condon, E. Dautbegovic, and T. J. Brazil, "An efficient numerical algorithm for the transient analysis of high-frequency non-linear circuits," in European Microwave Conference, Milan, Italy, Oct. 2002, pp. 1-4.
[6] S. Olver, "Numerical approximation of highly oscillatory integrals," Ph.D. dissertation, University of Cambridge, Nov. 2007.
[7] T. Xu, Z. Ye, and M. Kennedy, "Mathematical analysis of injection-locked frequency dividers," in Proc. International Symposium on Nonlinear Theory and its Applications, Bologna, Italy, Sep. 2006, pp. 639-642.
[8] Z. Ye, T. Xu, and M. P. Kennedy, "Locking range analysis for injection locking frequency dividers," in Proc. International Symposium on Circuits and Systems, Island of Kos, Greece, May 2006, pp. 4070-4073.
[9] D. O'Neill, D. Bourke, and M. P. Kennedy, "The devil's staircase as a method of comparing injection-locked frequency divider topologies," in Proc. European Conference on Circuit Theory and Design, vol. 3, Cork, Ireland, Sep. 2005, pp. 317-320.
[10] B. Razavi, "A study of injection locking and pulling in oscillators," IEEE Journal of Solid-Stagte Circuits, vol. 39, no. 9, pp. 1415-1424, Sep. 2004.
[11] O. Narayan and J. Roychowdhury, "Analysing oscillators using multitime PDEs," IEEE Trans. on Circuits and Systems, vol. 50, no. 7, pp. 894-903, July 2003.
[12] R. Gu and S. Ramaswamy, "Fractional- $N$ phase locked loop design and applications," in Proc. The 7th International Conference on ASIC, Guilin, China, Oct. 2007, pp. 327-332.
[13] Z. Ye, W. Chen, and M. P. Kennedy, "A novel dual-loop multi-phase frequency synthesizer," in Proc. European Conference on Circuit Theory and Design, Seville, Spain, Aug. 2007, pp. 567-570.
[14] M. J. Borkowski, T. A. D. Riley, J. Häkkinen, and J. Kostamovaara, "A practical $\Delta-\Sigma$ modulator design method based on periodical behavior analysis," IEEE Trans. on Circuits and Systems-II: Express Briefs, vol. 52, no. 10, pp. 626-630, Oct. 2005.
[15] T. Xu and M. Condon, "Design methodology for a maximum sequence length mash digital delta-sigma modulator," in Proc. World Congress on Engineering, London, U.K., July 2009.
[16] ——, "Comparative study of the mash digital delta-sigma modulators," in Proc. IEEE Ph.D. Research in Microelectronics and Electronics, Cork, Ireland, July 2009.
[17] M. K. Jain, Numerical Solution of Differential Equations, 2nd ed. New Delhi: Wiley Eastern Limited, 1984.
[18] E. Hairer, S. P. Nørsett, and G. Wanner, Solving ordinary differential equations I. Springer-Verlag, 1987.
[19] R. Bronson, Differential Equations, E. J. Bredensteiner, Ed. McGraw-Hill, 2003.
[20] J. Butcher, Numerical Methods for Ordinary Differential Equations, 2nd ed. John Wiley and Sons, 2008.
[21] F. Bashforth and J. C. Adams, Theories of Capillary Action. London: Cambridge University Press, 1883.
[22] K. A. Atkinson, An Introduction to Numerical Analysis, 2nd ed. John Wiley and Sons, 1988.
[23] G. Dahlquist, "A special stability problem for linear multistep methods," BIT, vol. 3, pp. 27-43, 1963.
[24] B. L. Ehle, "On padé approximations to the exponential function and $a$-stable methods for the numerical solution of initial value problems," in Report 2010. University of Waterloo, 1969.
[25] L. W. Nagel, "Spice2, a computer program to simulate semiconductor circuits," University of California, Berkeley, CA., Tech. Rep. ERL- M520, 1975.
[26] E. Dautbegovic, "Transient simulation of complex electronic circuits and systems operating at ultra high frequencies," Ph.D. dissertation, Dublin City University, Dublin, Ireland, Jan. 2005.
[27] M. Condon, E. Dautbegovic, and T. Xu, "Novel padé-based algorithms for numerical integration of odes," COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering, vol. 27, no. 6, pp. 1402-1417, 2008.
[28] C. F. Gerald and P. O. Wheatley, Applied Numerical Analysis, 6th ed. Harlow, England: Addison Wesley Longman, 1999.
[29] M. Stabrowski, "An efficient algorithm for solving stiff ordinary differential equations," Simulation Practice and Theory, vol. 5, no. 4, pp. 333-344, May 1997.
[30] M. C. Jeruchim, P. Balaban, and K. S. Shanmugan, Simulation of Communication Systems, Modeling, Methodology and Techniques, 2nd ed. New York: Kluwer Academic/Plenum Publishers, 2000.
[31] M. Khanamirian, "Quadrature methods for systems of highly oscillatory odes," DAMTP, University of Cambridge, Cambridge, UK, Tech. Rep. NA2008/01, 2008.
[32] A. Iserles and S. P. Nørsett, "Efficient quadrature of highly oscillatory integrals using derivatives," Proc. Royal Soc. A, vol. 461, pp. 1383-1399, 2005.
[33] M. Abramowitz and I. A. Stegun, Eds., Handbook of Mathematical Functions. New York: Dover Publications, 1965.
[34] U. Miekkala and O. Nevanlinna, "Iterative solution of systems of linear differential equations," Acta Numerica, vol. 5, p. 259307, 1996.
[35] O. Nevanlinna, "Remarks on picardlindeloff iteration," BIT, vol. 29, p. 328346, 1989.
[36] S. Vandewalle, Parallel multigrid waveform relaxation for parabolic problems. Stuttgart, Germany: B.G. Teubner GmbH, Dec. 1993.
[37] E. Ngoya and R. Larcheveque, "Envelope transient analysis: A new method for the transient and steady-state analysis of microwave communications circuits and systems," in Digest IEEE MTT-S International Microwave Symposium, vol. 3, San Francisco, CA, USA,, June 1996, pp. 1365-1368.
[38] D. Sharrit, "New method of analysis of communication systems," in Proc. IEEE MTT-S WMFA: Nonlinear CAD Workshop, San Francisco, CA, USA,, June 1996.
[39] A Users Guide to Envelope Following Analysis Using Spectre RF, Cadence Design Systems, San Jose, CA , USA, April 2007.
[40] U. Singh and M. Green, "Dynamics of high-frequency cmos dividers," in Proc. IEEE International Symposium on Circuits and Systems, vol. 5, Scottsdale, Arizona, USA, May 2002, pp. 421-424.
[41] T. H. Lee, The Design of CMOS Radio-Frequency Integrated Circuits. Cambridge University Press, 1998.
[42] A. Buonomo and A. L. Schiavo, "Modelling and analysis of differential VCOs," International Journal of Circuit Theory and Applications, vol. 32, pp. 117-131, 2004.
[43] H. R. Rategh and T. H. Lee, "Superharmonic injection-locked frequency dividers," IEEE Journal of Solid-State Circuits, vol. 34, pp. 813-821, June 1999.
[44] R. L. Miller, "Fractional-frequency generators utilizing regenerative modulation," Proc. of IRE, vol. 27, pp. 446-457, July 1939.
[45] R. J. Betancourt-Zamora, S. Verma, and T. H. Lee, "1-GHz and 2.8GHz injectionlocked ring oscillator prescalers," in Symp. VLSI Circuits Dig. Tech. Papers, Kyoto, Japan, June 2001, pp. 47-50.
[46] S. Verma, H. Rategh, and T. H. Lee, "A unified model for injection-locked frequency dividers," IEEE Journal of Solid-State Circuits, vol. 38, no. 6, pp. 10151027, June 2003.
[47] S.-M. Wu and W.-L. Chen, "A 5.8-GHz CMOS VCO with injection-locked frequency divider for IEEE 802.11a application," in Proc. European Conference on Circuit Theory and Design, Cracow, Poland, Sep. 2003.
[48] D. O’Neill, D. Bourke, Z. Ye, and M. P. Kennedy, "Accurate modeling and experimental validation of an injection-locked frequency divider," in Proc. European Conference on Circuit Theory and Design, vol. 3, Cork, Ireland, Sep. 2005, pp. 409-412.
[49] H. P. Moyer and A. S. Daryoush, "A unified analytical model and experimental validations of injection-locking processes," IEEE Trans. on Microwave Theory and Techniques, vol. 38, no. 4, pp. 493-499, Apr. 2000.
[50] B. Razavi, "Analysis, modeling and simulation of phase noise in monolithic voltage-controlled oscillators," in Proc. IEEE Custom Integrated Circuits Conference, May 1995, pp. 323-326.
[51] R. L. Burden and J. D. Faires, Numerical Analysis, 5th ed. Boston: PWS Publishing Company, 1993.
[52] E. L. Allgower and K. Georg, Numerical Continuation Methods. New York: Springer-Verlag, 1990.
[53] C. E. Christoffersen, M. Condon, and T. Xu, "A new method for the determination of the locking range of oscillators," in Proc. European Conference on Circuit Theory and Design, Seville, Spain, Aug. 2007, pp. 575-578.
[54] T. Xu and M. Condon, "An effective method for the determination of the locking range of an injection-locked frequency divider," in Proc. Emerging Trends in Wireless Communications, Dublin, Ireland, Apr. 2008, pp. 47-50.
[55] -_, "Accurate simulation of the devil's staircase of an injection-locked frequency divider," in Proc. International Conference on Scientific Computing in Electrical Engineering, Espoo, Finland, Sep. 2008, pp. 105-106.
[56] T. Mei and J. Roychowdhury, "A time-domain oscillator envelope tracking algorithm employing dual phase conditions," IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, vol. 27, no. 1, pp. 59-69, Jan. 2008.
[57] H. P. Langtangen, Computational Partial Differential Equations: Numerical Methods and Diffpack Programming. Springer, 1999.
[58] T. J. Aprille and T. N. Trick, "Steady-state analysis of nonlinear circuits with period inputs," Proc. of IEEE, vol. 60, no. 1, pp. 108-114, Jan. 1972.
[59] H. Inose, Y. Yasuda, and J. Marakami, "A telemetering system by code modulation, delta-sigma modulation," IRE Trans. on Space, Electronics and Telemetry, vol. 8, pp. 204-209, Sept. 1962.
[60] A. Reddy, "Noise shaping with sigma-delta modulators in fractional- $N$ synthesizers," in Proc. IEEE Intl. Workshop on Radio-Frequency Integration Technology, Singapore, Dec. 2007, pp. 329-332.
[61] M. Kozak and I. Kale, "Rigorous analysis of delta-sigma modulators for fractional-n pll frequency synthesis," IEEE Trans. on Circuits and Systems-I: Regular Papers, vol. 51, no. 6, pp. 1148-1162, June 2004.
[62] D. Reefman, J. Reiss, E. Janssen, and M. Sandler, "Description of limit cycles in sigma-delta modulators," IEEE Trans. on Circuits and Systems-I: Regular Papers, vol. 52, no. 6, pp. 1211-1223, June 2005.
[63] J. Reiss and M. Sandler, "Detection and removal of limit cycles in sigma delta modulators," IEEE Trans. on Circuits and Systems-I: Regular Papers, vol. 55, no. 10, pp. 3119-3130, Nov. 2008.
[64] S. Pamarti, J. Welz, and I. Galton, "Statistics of the quantization noise in 1-bit dithered single-quantizer digital delta-sigma modulators," IEEE Trans. on Circuits and Systems-I: Regular Papers, vol. 54, no. 3, pp. 492-503, Mar. 2007.
[65] S. Pamarti and I. Galton, "Lsb dithering in mash delta-sigma d/a converters," IEEE Trans. on Circuits and Systems-I: Regular Papers, vol. 54, no. 4, pp. 779790, Apr. 2007.
[66] K. Hosseini and M. P. Kennedy, "Mathematical analysis of a prime modulus quantizer mash digital delta-sigma modulator," IEEE Trans. on Circuits and Systems-II: Express Brief, vol. 54, no. 12, pp. 1105-1109, Dec. 2007.
[67] -_, "Mathematical analysis of digital mash delta-sigma modulator for fractional- $N$ frequency synthesizers," in Proc. IEEE Ph. D. Research in Microelectronics and Electronics, Otranto (Lecce), Italy, June 2006, pp. 309-312.
[68] K. Hosseini, M. P. Kennedy, and C. McCarthy, "Calculation of sequence lengths in mash 1-1-1 digital delta sigma modulators with a constant input," in Proc.

IEEE Ph. D. Research in Microelectronics and Electronics, Bordeaux, France, July 2007, pp. 13-16.
[69] M. J. Borkowski and J. Kostamovaara, "Spurious tone free digital delta-sigma modulator design for dc inputs," in Proc. IEEE International Symposium on Circuits and Systems, Kobe, Japan, May 2005, pp. 5601-5604.
[70] R. E. Crandall and C. Pomerance, Prime numbers: a computational perspective. Springer, 2001.
[71] A. V. Oppenheim and R. W. Schafer, Discrete-Time Signal Processing. Upper Saddle River: NJ: Prentice-Hall, 1999.
[72] L. Sun, T. Lepley, F. Nozahic, A. Bellissant, T. Kwasniewski, and B. Heim, "Reduced complexity, high performance digital delta-sigma modulator for fractionaln frequency synthesis," in Proc. of the 1999 IEEE International Symposium on Circuits and Systems, vol. 2, Orlando, FL, USA, July 1999, pp. 152-155.
[73] X. Lai and J. Roychowdhury, "Capturing oscillator injection locking via nonlinear phase-domain macromodels," IEEE Trans. Microwave Theory Tech., vol. 52, no. 9, p. 22512261, Sep. 2004.
[74] M. Condon, A. Deano, and A. Iserles, "On highly oscillatory problems arising in electronic engineering," To appear in Math. Modelling \& Num. Anal., 2009.


[^0]:    ${ }^{1}$ In a realistic model $x(t)$ is random, rather than alternating. However, introducing stochastic component would have made the comparison of different methods considerably more difficult.

[^1]:    ${ }^{1}$ Note that in fundamental locking, $e^{j \omega_{\text {out }} t}=e^{j \omega_{\text {in } \mathrm{j}} t}$

[^2]:    ${ }^{1} V_{D D}$ is supplied by a dry battery.

[^3]:    ${ }^{1}$ MASH $m-n-l$ means that it consists of an $m$ th order error feedback modulator (EFM $m$ ), an $n$th order EFM (EFM $n$ ) and an $l$ th order EFM (EFM $l$ ), cascaded.

