An Integrated Heterodyne Interferometer with on-chip Detectors and Modulators

by

David B. Cole

B.S., New York University (1996)
B.E., Stevens Institute of Technology (1996)
M.S., University of Pennsylvania (1998)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Science in Electrical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2015

© Massachusetts Institute of Technology 2015. All rights reserved.
An Integrated Heterodyne Interferometer with on-chip Detectors and Modulators

by

David B. Cole

Submitted to the Department of Electrical Engineering and Computer Science on August 3, 2015, in partial fulfillment of the requirements for the degree of Doctor of Science in Electrical Engineering

Abstract

The application of CMOS processing techniques developed in the microelectronics world to that of silicon photonics has been the catalyst for the rapid proliferation of smaller, higher performance, and more densely integrated photonic devices that are rapidly advancing the field with large-scale implementations of photonic systems. It is only logical that as the silicon photonics library continues to grow, that these devices will be synthesized into complete systems for applications including optical networking and communications, imaging, and sensors. Among complex optical systems, the interferometer represents perhaps the most important class of optical sensors and scientific instruments ever developed. Today, interferometric techniques are key to applications such as displacement measurement, photolithography, vibrometry, optical coherence tomography (OCT), and LIDAR.

In these applications, the preferred mode is the heterodyne interferometer. However, modern heterodyne interferometers are complex systems requiring bulk optical devices to be implemented. As a result, they are large and expensive precision instruments limited to industrial and scientific applications. The development of a chip-scale integrated interferometer, with its significantly smaller form factor, increased stability, and lower cost, could greatly expand the application of interferometric techniques. Leveraging silicon photonics, the required components can be realized on-chip, allowing for a low-cost, high-precision interferometer to be implemented. In this thesis, the design and experimental results of the first silicon chip-scale heterodyne interferometer is presented. The device is constructed of a series of on-chip beam-splitters, modulators, and germanium detectors in a Michelson-like configuration. The interferometer achieves a noise-limited position resolution of approximately 2 nm in a 1 mm by 6 mm footprint. The vibrometer and LIDAR modes; two modalities with important scientific, industrial, and consumer applications, are also demonstrated.

Thesis Supervisor: Michael R. Watts
Title: Associate Professor of Electrical Engineering
Acknowledgments

The pursuit of a doctorate degree has been described as one of the most isolating and solitary endeavors into which humans willingly enter. In seeking to create ‘original work’ which makes ‘a significant contribution to knowledge’, doctoral students are expected to become experts in their particular domain, no matter how small it may be. Therefore, there comes a time when we feel as if there is no one to whom we can turn when we don’t have the answer, an experiment isn’t working, or we’re lacking confidence in which path to take. Despite this seemingly self-imposed exile into an unknown intellectual wilderness, few students, no matter how intelligent, driven, or tenacious, can complete this journey on their own. And to this I am no exception, as I owe my success to colleagues, friends, and family who have provided steadfast support over not only the past four years, but throughout my entire life.

I would like to thank my adviser Professor Michael Watts for welcoming me into the Photonic Microsystems Group (PMG), and for allowing me to work with such a talented group of colleagues in one of the world’s foremost photonics groups. Looking to change my research direction, Professor Watts was willing to take a chance on me, and I have grown tremendously as a researcher under his supervision. Even if we didn’t always agree on a particular direction, he always provided the freedom to follow a thread to my satisfaction. I would also like to thank my colleagues, both within PMG and affiliated groups, past and present. From the beginning they willingly gave of their time and expertise in teaching me the ins and outs of our research, the tools of the trade, and the intricacies of working in a photonics lab. They allowed me to assist with their own experiments on many occasions, often trusting their own success to me. And because of this it wasn’t long before I was able to make my own contributions. Thank you to Cheri Sorace-Agaskar, Michele Moresco, Zhan Su, Jie Sun, Erman Timurdogan, Ami Yaacobi, Ehsan Shah Hoseini, Johnathan Bradley, Purnawirman, Patrick Callahan, Katia Shteyrkova, Sasha Biberman, Salih Magden, Anna Baldycheva, Brad Cordova, Christos Samolis, Manan Raval, Chris Poulton, Ben Moss, Nanxi Li, Gurpreet Singh, and Jerry Zhou for your contributions to this
work. And thank you to Stacy McGrath for ensuring that the administrative issues were never a barrier to the research.

I would like to sincerely thank my thesis committee, Professors Leslie Kolodziejski and Erich Ippen. Professor Ippen did something that in my experience very few professors of his stature would; he responded to my emails as an aspiring MIT applicant not once, but throughout the application process. And since my arrival he has always made himself available to offer advice and to serve as an unofficial mentor. As my academic "grandfather" and someone who is widely respected throughout the department, I feel honored to have him as a committee member. Professor Kolodziejski has consistently been one of my strongest supporters throughout my time at MIT. From easing my anxieties following the TQE exams after my first semester to offering guidance on the obstacles graduate students often face, she was always there when I needed her. And I am proud that when my daughter is old enough to appreciate what I have accomplished, that she will know that a female professor was so instrumental in her father's success.

I would like to thank Professor James Fujimoto for serving as my academic adviser and for allowing me to listen to his classes, despite his no listener rule. His courses helped me remember that my true love is electromagnetics, and his work in Optical Coherence Tomography was a key foundation upon which I built the heterodyne interferometer. I would like to thank the ladies of the EECS Graduate Office: Alicia Duarte, Joycelyn Thompson, and Claire Benoit. They helped me to navigate the bureaucracy that is so often part of graduate school, and always had encouraging words whether they saw me in the graduate office or in the hallways. In particular, Janet Fischer was there for me on so many occasions that it’s incalculable, and I understand why she is such an indispensable part of this department. I am also appreciative of my "MIT Mother", Dean Blanche Staton. Having served in that role for so many students of color, she was a confidant and a fount of wisdom when I needed it the most. Dr. Andrew Puryear was a valuable resource both before and after arriving at MIT, and provided the perspective of a fellow graduate student that helped make my transition less daunting. I would like to thank my friend Dr. Jerry
Selvaggi, who raised the bar as a teacher of physics, math, and engineering during my years at NYU. It was his example that helped mold me into a well-rounded engineer, and he consistently demonstrated the patience and skill that exemplify teaching at its best. Thanks to Nivedita Chandrasekaran for our many enjoyable conversations that were often a respite from the day-to-day routine. And I would be remiss if I didn’t thank Dr. Shan-Yuan Ho. I could always depend on Ho-Ho for her friendship, or if I needed help on some complicated math issue. But most importantly, she provided valuable firsthand experience on being a MIT doctoral student that helped keep life in perspective during the hard times.

But long before I arrived at MIT, the foundation was laid by family who provided the love and guidance necessary to remain on the right path, and who stood by me when times were good and bad. My cousin Eula Morris was my first grade teacher. Because I was a very quiet and shy child who didn’t talk much, the administrators at my elementary school wanted to hold me back in kindergarten instead of moving me on to the first grade. But she refused to let that happen to her little cousin, whom she knew better than almost anyone. She took me into her class and placed me on a trajectory that would have been radically different without her intervention. My mother Mary Chatman raised me as a single parent to value education. She sacrificed to give me opportunities she didn’t have, and she kept me away from the destructive influences that claimed so many of my peers. She allowed me to try my hand at whatever I was interested, and set the stage for my lifelong love of learning. During those most difficult days during my freshman and sophomore years at NYU, she cried with me when I wasn’t sure how I would make it through. She made me the man I am today. I would also like to thank my father and mother in-law, Vervice and Jacquelyn Mitchell, for giving me the gift of their daughter and for welcoming me into their incredible family. To my sisters and brothers in-laws Joni, Stephanie, Damon, and Ramon, thank you for treating me like your brother and for all of your support. Thank you to Cousin Desi for coming to help us during these past few years, and for the love you have shown your little cousin. And thank you to friends so numerous I can’t name them individually.
But none of this would have been possible without the love, sacrifice, and support of my wife Renee and daughter Gabrielle. You persevered in my absence during my first term at MIT. You bore the many late nights and weekends when I had to work, the times when I was physically present but mentally absent, and the frequent changing of plans that became a part of our lives. This would not have been possible without you, and I want you to know how much I love and appreciate you. I dedicate this work to both of you.

Last but not least, I would like to thank my father, Curtis Cole, who we lost during my time at MIT. I wish you were here to share this with us.
# Contents

1. **Silicon Photonics and the Interferometer**
   - 1.1 Background ........................................... 19
   - 1.2 Previous Work ........................................ 23
   - 1.3 Thesis Overview and Outline ..................... 25

2. **Background and Theory** .................................. 27
   - 2.1 Transfer Function for a Free Space Heterodyne Interferometer ........ 27
   - 2.2 Balanced Detection .................................... 33
   - 2.3 Displacement Resolution and Measurement Limitations ............. 37
     - 2.3.1 Two Wavelength Measurements for Increased Unambiguous Range ... 40
   - 2.4 Discrete vs. Continuous Phase Measurements .................... 43

3. **Noise and Performance Analysis** ....................... 49
   - 3.1 Background ............................................ 49
     - 3.1.1 Additive Noise Sources .......................... 52
     - 3.1.2 Phase Noise Sources .............................. 64
     - 3.1.3 Balanced Detection Signal-to-Noise Ratio ............. 66
   - 3.2 The Effect of Noise on Phase Measurement ..................... 71
   - 3.3 The Coherent Demodulator .................................. 77
     - 3.3.1 Resolution of Doppler and Phase Sign Ambiguity ............ 79
   - 3.4 Conclusion .............................................. 80

4. **The Integrated Heterodyne Interferometer** .......... 81
Appendix C  A Jones Matrix Analysis of a Forward and Reverse Pass through a Quarter Waveplate

Appendix D  Utilization of Amplitude Modulators in the Heterodyne Interferometer

D.1  Synthetic Wavelength Interferometry utilizing an Amplitude Modulator

Appendix E  Input Waveguides, Adiabatic Couplers, and the Adiabatic Theorem

Appendix F  The Effect of an Under-suppressed Carrier and Sidelobes in the Single-Sideband Modulator
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Patent application submitted by Wanlass for the first CMOS circuit.</td>
</tr>
<tr>
<td>1-2</td>
<td>Scanning electron microscope image of a 64 by 64 element nano-photonic phased array with 9 μm by 9 μm unit cells.</td>
</tr>
<tr>
<td>1-3</td>
<td>An original diagram of Michelson’s 1881 Interferometer.</td>
</tr>
<tr>
<td>1-4</td>
<td>A Heterodyne Interferometer on Lithium Niobate.</td>
</tr>
<tr>
<td>1-5</td>
<td>A Laser Doppler Vibrometer for the measurement of arterial blood flow.</td>
</tr>
<tr>
<td>2-1</td>
<td>Conceptual model for the Integrated Heterodyne Interferometer.</td>
</tr>
<tr>
<td>2-2</td>
<td>Conceptual diagram of the balanced detection scheme.</td>
</tr>
<tr>
<td>2-3</td>
<td>True versus aliased phase over multiple half-wavelength displacements.</td>
</tr>
<tr>
<td>3-1</td>
<td>A conceptual diagram of the fabricated TIA circuit.</td>
</tr>
<tr>
<td>3-2</td>
<td>A simple sinusoid for examining amplitude noise contributions to phase noise.</td>
</tr>
<tr>
<td>3-3</td>
<td>A conceptual drawing of the IQ demodulator for the coherent measurement of phase.</td>
</tr>
<tr>
<td>4-1</td>
<td>A conceptual diagram of the Integrated Heterodyne Interferometer.</td>
</tr>
<tr>
<td>4-2</td>
<td>The Integrated Heterodyne Interferometer</td>
</tr>
<tr>
<td>4-3</td>
<td>Conceptual diagram of the Single-Sideband Modulator (SSBM).</td>
</tr>
<tr>
<td>4-4</td>
<td>Single sideband operation of the SSBM at 1570 nm.</td>
</tr>
<tr>
<td>4-5</td>
<td>Images of the (a) SSBM and (b) detector sections of the fabricated device.</td>
</tr>
<tr>
<td>4-6</td>
<td>3-D images of the Ge Detectors.</td>
</tr>
<tr>
<td>Page</td>
<td>Image/Text Description</td>
</tr>
<tr>
<td>------</td>
<td>------------------------</td>
</tr>
<tr>
<td>4-7</td>
<td>Experimental setup for the measurement of nanometer-scale displacements.</td>
</tr>
<tr>
<td>4-8</td>
<td>Functional diagram of the TIA receiver circuit.</td>
</tr>
<tr>
<td>4-9</td>
<td>Schematic for the TIA receiver board.</td>
</tr>
<tr>
<td>4-10</td>
<td>TIA receiver circuit as fabricated on PCB.</td>
</tr>
<tr>
<td>4-11</td>
<td>Measurement Signal versus Displacement.</td>
</tr>
<tr>
<td>4-12</td>
<td>Static noise and PDF for a one-second measurement period.</td>
</tr>
<tr>
<td>4-13</td>
<td>Phase noise spectrum for a one-second measurement period.</td>
</tr>
<tr>
<td>4-14</td>
<td>Balanced and unbalanced phase noise from the measurement detectors taken over one second.</td>
</tr>
<tr>
<td>4-15</td>
<td>Wrapped phase versus displacement over a distance of 1.55 μm.</td>
</tr>
<tr>
<td>4-16</td>
<td>Lissajous curves and phase for uncorrected and corrected 50 Hz triangular-wave displacement.</td>
</tr>
<tr>
<td>4-17</td>
<td>Unwrapped corrected and uncorrected phase over a distance of 1.55 μm.</td>
</tr>
<tr>
<td>5-1</td>
<td>Vibrometer measurements using a speaker-mounted retroreflector.</td>
</tr>
<tr>
<td>5-2</td>
<td>The vibrational spectrum of a sinusoidally-driven speaker at 100 Hz and 200 Hz.</td>
</tr>
<tr>
<td>5-3</td>
<td>Phase plots for 1 Hz and 50 Hz vibrational frequencies.</td>
</tr>
<tr>
<td>5-4</td>
<td>Measurement spectrum of a target driven at 5 Hz and 500 nm pk-pk amplitude.</td>
</tr>
<tr>
<td>5-5</td>
<td>Phase versus time for 25 Hz oscillation at 10 Vpk-pk displacement.</td>
</tr>
<tr>
<td>5-6</td>
<td>Derived Doppler frequency for 1 Hz and 25 Hz oscillations at 10 Vpk-pk.</td>
</tr>
<tr>
<td>5-7</td>
<td>Derived velocity for 1 Hz and 25 Hz oscillations at 10 Vpk-pk.</td>
</tr>
<tr>
<td>5-8</td>
<td>Noisy velocity obtained without a Savitzky-Golay filter.</td>
</tr>
<tr>
<td>6-1</td>
<td>A linearly-chirped CW waveform in the time domain.</td>
</tr>
<tr>
<td>6-2</td>
<td>The Integrated Heterodyne LIDAR.</td>
</tr>
<tr>
<td>6-3</td>
<td>Sawtooth FMCW waveform for range measurement.</td>
</tr>
<tr>
<td>6-4</td>
<td>Illustration of the Rayleigh criterion.</td>
</tr>
<tr>
<td>6-5</td>
<td>An Agilent 81600B Tunable Laser.</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Synthetic wavelengths for several wavelength pairs. .......................... 41
2.2 N-ratio for several synthetic wavelength pairs. ............................... 42

4.1 Static phase noise statistics for a one-second measurement period. ...... 93
4.2 Summary of measurements for 5 nm, 10 nm, and 25 nm displacement steps. ................................................................. 98

5.1 Vibrational amplitude computed using the relative amplitudes of the first five harmonic sidebands. ................................................. 113
5.2 $\beta$s computed using the relative sideband amplitudes. ...................... 113

6.1 Summary of measurement results for a 630 GHz/s sweep rate. ........... 131
6.2 Summary of measurement results for a 1.26 THz/s sweep rate. ........... 131
6.3 Results of displacement measurement using phase shift for a 630 GHz/s sweep rate. ................................................................. 134

F.1 Spectral components for equal responsivities when analyzing the effect of an unsuppressed carrier and sidebands in the SSBM. ................. 200
F.2 Spectral components for non-equal responsivities when analyzing the effect of an unsuppressed carrier and sidebands in the SSBM. ................. 200
Chapter 1

Silicon Photonics and the Interferometer

1.1 Background

In 1833, Michael Faraday described what he called the "extraordinary case" of electrical conduction increasing with temperature in silver sulfide crystals, which was the opposite of what normally occurred in metals. His discovery was the first known observation of what would eventually be known as the semiconductor effect [1]. The first transistor, made of germanium, was invented over 100 hundred years later in 1947 by Bell Lab physicists John Bardeen and Walter Brattain. But while germanium has many desirable properties, such as its ease of fabrication and high frequency response, its narrow band gap led to high transistor leakage currents even in the “off” state, and germanium transistors shut down completely above 75° C. Morris Tanenbaum, a Bell Labs chemist, demonstrated the first silicon transistor in January 1954. But it was not pursued by Bell because it was deemed commercially unattractive. However, a team at Texas Instruments (TI) independently developed a silicon transistor in April of that same year. TI quickly commercialized their product and would go on to dominate the silicon transistor market for years; and by the end of the 1950s silicon had replaced germanium as the preferred semiconductor material [2].

In 1963, Frank Wanlass and C.T. Sah of the Fairchild R&D Laboratory pub-
lished a conference paper describing logic circuits combining p-channel and n-channel MOS transistors in a complementary symmetric configuration [3]. This configuration, patented by Wanlass in 1963 [4], is what is known today as Complimentary-Metal-Oxide-Semiconductor (CMOS), and has been the catalyst for many of the incredible technological advances that have fundamentally changed the world over the past fifty years.

Figure 1-1: Patent application submitted by Wanlass for the first CMOS circuit [5].

Since their introduction, CMOS devices have closely followed Moore’s law, which predicts a doubling of the transistor count per unit area every eighteen to twenty-four months. Today, minimum feature sizes on the order of 20 nm are possible, compared to the tens of microns when Moore first stated his now famous law. But it is the field of silicon photonics that would become one of the greatest beneficiaries of these advances in microelectronics. Perhaps it is serendipitous that silicon possesses not only desirable electrical properties, but also numerous optical properties that has made it the medium of choice for photonic devices. Silicon and its dioxide (SiO$_2$) are transparent throughout most of the infrared spectrum, and their large refractive index contrast provides strong mode confinement in silicon waveguides, permitting sharp
bends and close-packing of devices. Silicon has a strong thermo-optic effect that allows the refractive index to be changed with temperature, and a strong electro-optic effect that allows the refractive index to be changed electrically via the plasma-dispersion effect. The thermo-optic effect is relatively slow, with a frequency response on the order of kilohertz. It is ideal for tuning parameters that change slowly over time, such as adjusting the resonances of ring filters. The electro-optic effect is much faster and has a frequency response on the order of tens of gigahertz. This makes silicon ideal for fabricating high-speed modulators for optical communications and data networks.

The application of CMOS processing techniques developed in the microelectronics world to that of silicon photonics has been the catalyst for the rapid proliferation of smaller, higher performance, and more densely integrated photonic devices that are rapidly advancing the field of photonics with large-scale implementations of photonic microsystems. The demonstration by Sun et al. [6] of a 64 by 64 optical phased array, developed on a 300 mm Silicon-on-Insulator (SOI) process and shown in Figure 1.2, highlights the integration possible with silicon photonics. Moreover, this advanced 300 mm silicon photonics platform has led to a number of demonstrations that will be discussed later in this thesis. But it is only logical that as the silicon photonics library continues to grow that these devices will be synthesized into complete photonic systems for applications such as optical networking and communications, imaging, and sensors.

Among complex optical systems, the interferometer represents perhaps the most important class of optical sensors and scientific instruments ever developed. Since the famous work of Michelson [7,8], the interferometer has become one of the most important of all scientific instruments. Today, interferometric techniques are key to applications such as displacement measurement, Doppler vibrometry and velocimetry, optical surface characterization, Fourier transform spectroscopy, optical coherence tomography (OCT), holography, gravitational wave detection, and light detection and ranging (LIDAR). Importantly, interferometers are crucial for photolithography and high-precision semiconductor manufacturing, where they are used to control wafer stepper position, e-beam location in mask writing applications, and for determining
layer thickness and uniformity [9]. In these critical applications, the preferred operational mode is the heterodyne interferometer. In this configuration, displacement information is encoded as a timing difference, or equivalently, a phase difference between a pair of equal frequency sinusoids. This is in contrast to the less sensitive and error-prone amplitude measurement common to traditional homodyne interferometers. However, this increased accuracy and performance comes at a cost. Bulk heterodyne interferometers are complex optical systems requiring multiple beam-splitters, waveplates, and frequency modulators to be implemented. As a result, heterodyne interferometers used in modern applications are large and expensive precision instruments that are primarily limited to industrial and scientific applications [10]. The development of a chip-scale integrated interferometer, with a significantly smaller form factor, increased stability, and lower cost, could greatly expand the application of interferometry to automobiles and handheld medical and consumer devices. Lever-
aging silicon photonics, all the required components can indeed be realized on-chip, allowing for a chip-scale, high-precision interferometer to be implemented.

Figure 1-3: An original diagram of Michelson’s 1881 Interferometer [7].

1.2 Previous Work

Dahlquist, Peterson, and Culshaw first applied heterodyning techniques to interferometry in 1966 by using the Zeeman split of a He-Ne laser to generate two frequencies with orthogonal polarizations [11]. This work eventually lead to the commercialization of the Hewlett-Packard Interferometer [10,12], for years the industrial standard and now marketed and sold by Keysight Technologies. The push to miniaturize the interferometer has followed closely with progress in photonic systems. And while silicon photonics has allowed an unprecedented level of integration, on-chip interferometers on LiNbO$_3$ [13,14] were demonstrated as far back as 1991. The device developed by Toda et al., measuring 47 mm by 5 mm and shown in Figure 1-4, achieved a resolution of 3 nm at a wavelength of 633 nm. The heterodyne frequencies were generated using an on-chip mode converter and phase modulator. However, the photodetectors were external to the device. Because both frequencies were generated concurrently by a single modulator, they were separated utilizing orthogonal TE and TM modes. Therefore, on-chip mode converters, mode splitters, and polarizers were required, similar to bulk interferometers. An external quarter waveplate was also required to rotate the outgoing TM wave to TE upon reflection to allow interference between the
In recent years, researchers have sought to implement chip-scale interferometric devices using the CMOS-compatible SOI process. One such example is a dual Laser Doppler Vibrometer for measuring the flow rate of arterial blood, illustrated in Figure 1-5 [15]. This device, measuring less than 5 mm$^2$, consists of a series of on-chip splitters, 90° hybrid couplers, and two pairs of grating couplers that allow each vibrometer to couple light on and off the chip. The input laser is divided into three paths: one path is coupled to each of the two transmitting grating couplers that focus light on the artery. The third path is further split in two, with each path used as a reference oscillator for each vibrometer. The displacement of the artery due to the passage of blood modulates the reflected waves, and the time delay between the two Doppler-modulated tones are measured to determine the flow rate. The vibrometers use external photodetectors to convert the optical signals into photocurrents, which are then filtered and processed off-chip. Because only a single laser frequency is employed, the device is fundamentally a homodyne system, with a beat frequency present only during Doppler modulation of the reflected waveforms.

In this thesis, the first chip-scale heterodyne interferometer fabricated on a CMOS-compatible SOI process is demonstrated. Measuring 1 mm by 6 mm and exhibiting a

Figure 1-4: A Heterodyne Interferometer on Lithium Niobate [13].
2 nm noise-limited position resolution, the interferometer is also the smallest heterodyne interferometer demonstrated to date and could impact numerous interferometric and metrology applications, including displacement measurement, laser Doppler vibrometry, imaging, and light detection and ranging (LIDAR). The interferometer is constructed of a series of on-chip beam-splitters, single-sideband modulators (SS-BMs), and germanium detectors in a Michelson-like configuration. The device is demonstrated in three operational modes: displacement measurement, laser Doppler vibrometer, and LIDAR.

1.3 Thesis Overview and Outline

- In Chapter 2 the theory of interferometry is reviewed and the transfer function of a bulk interferometer is derived before describing balanced detection and limitations on measurement accuracy.

- In Chapter 3 the primary sources of noise in an interferometer are described, and the relevant statistics characterizing these sources are reviewed before developing expressions for the Signal-to-Noise Ratio (SNR). The relationship between
SNR and performance limitations on system phase and position noises are then demonstrated.

- Chapter 4 provides background on the 300 mm SOI process used in the device fabrication, before proceeding with a description of the integrated heterodyne interferometer and the experimental setup for nanometer-scale displacement measurements. The chapter closes by presenting results of the displacement mode measurements.

- Chapters 5 and 6 build on the results of the displacement measurements in Chapter 4, and lay the theoretical foundation for operating the interferometer as a Laser Doppler Vibrometer (LDV) and LIDAR; achieving good results in both modalities.

- Chapter 7 concludes by summarizing the results of all three operational modes and discussing design improvements that should not only increase device performance, but also increase the interferometer’s utility across a broad range of scientific and consumer applications.
Chapter 2

Background and Theory

2.1 Transfer Function for a Free Space Heterodyne Interferometer

In this section the transfer function is derived for a bulk heterodyne interferometer operating in free space. This will serve as the foundation for the performance analysis that occurs throughout the remainder of this thesis. A conceptual model of a bulk heterodyne interferometer in a Michelson configuration is shown in Figure 2-1. For an interferometer operating in free space, the reference and measurements signals on the outbound path must not only be separated in space, but also collinearly recombined prior to photodetection. To provide this spatial separation in a free space system, polarizing beam splitters (PBS), polarization analyzers, and the appropriate waveplates are required to manipulate the polarization. For this analysis several simplifying assumptions can be made.

1. The laser source is linearly polarized along the x-axis and given by the time harmonic field $E_0 e^{-j\omega t}$, where $E_0$ is real with units of V/m. In reality, laser sources often have a slightly elliptical polarization that introduces polarization contamination into the two paths and leads to phase nonlinearities. Such phase nonlinearities will be discussed in a subsequent chapter.

2. Devices are assumed lossless. This is not too hard to justify, as the field am-
plitudes are not critical for a heterodyne interferometer, and maintaining equal amplitudes in the various paths is not a limiting factor in device performance. The frequency separation of the measurement and reference paths are assumed close enough such that the response of frequency-dependent devices can be ignored.

3. There is no unwanted polarization mixing due to imperfections in the PBS or other polarization components, so polarization purity is assumed. However, this is not the case in reality, and cross-talk between polarization components must be taken into account when analyzing actual system performance, as mentioned in the first assumption.

4. No reflection at the boundary between the measurement arm aperture and free space is assumed.

The fields at 2a and 2b are the result of splitting the source field at an ideal 3-dB coupler, in this case a non-polarizing beam splitter (NPBS). Recalling that the transfer matrix relating the fields at the input ports to the fields at the output ports
for a 3-dB hybrid coupler is given by

\[
T = \begin{bmatrix}
\frac{1}{\sqrt{2}} & -j \frac{1}{\sqrt{2}} \\
-j \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}
\end{bmatrix} \tag{2.1}
\]

Then the fields at 2a and 2b are given respectively as

\[
-j \frac{E_0}{\sqrt{2}} e^{-j\omega_0 t} \hat{x} \tag{2.2}
\]
\[
\frac{E_0}{\sqrt{2}} e^{-j\omega_0 t} \hat{x} \tag{2.3}
\]

The fields at 2a and 2b pass through modulators with angular frequency shifts \(\Delta \omega_1\) and \(\Delta \omega_2\), respectively, such that the fields at 3a and 3b are modulated to \(\omega_1\) and \(\omega_2\). Furthermore, the polarization of one path is rotated 90° such that the heterodyne carriers are differentiated by polarization as well as frequency. In this model the phase rotator is placed in the lower path, indicated here by a half waveplate. To account for any un-characterized phase shifts that may occur between the two paths, arbitrary phase shifts \(\phi_1\) and \(\phi_2\) are ascribed to the lower and upper paths. The fields can therefore be written as

\[
-j \frac{E_0}{\sqrt{2}} e^{-j(\omega_1 t + \phi_1)} \hat{y} \tag{2.4}
\]
\[
\frac{E_0}{\sqrt{2}} e^{-j(\omega_2 t + \phi_2)} \hat{x} \tag{2.5}
\]

The fields at 3a and 3b become the inputs to a second 3-dB coupler, and hence half the power at 3a and 3b will appear superimposed on the coupler outputs represented by the points 4a and 4b, which have total field amplitudes given by

\[
-j \frac{E_0}{2} \left[ e^{-j(\omega_2 t + \phi_2)} \hat{x} + e^{-j(\omega_1 t + \phi_1)} \hat{y} \right] \tag{2.6}
\]
\[
\frac{E_0}{2} \left[ e^{-j(\omega_2 t + \phi_2)} \hat{x} - e^{-j(\omega_1 t + \phi_1)} \hat{y} \right] \tag{2.7}
\]
The field at 4a is used to establish the reference beat frequency for the heterodyne system. This beating is a form of optical mixing and is the result of interference between fields of different frequency. This “mixing”, or down-conversion, is necessary as the photodetectors are unable to respond to optical frequencies. Before analyzing the response of the photodetectors, recall first that a photodetector is a square-law device responding to a field intensity \( I \) with units of \( \text{W/m}^2 \) and is defined as the squared magnitude of the total field, or \( I \propto |E_t|^2 = \vec{E}_t \cdot \vec{E}_t^* \). This expression takes into account both the complex and vectorial nature of the fields. For fields such as those at the output of the second coupler, the total field consists of orthogonal components. The intensity for such a field is given by

\[
I = |E_x e^{-j(\omega_1 t + \phi_1)} \hat{x} + E_y e^{-j(\omega_2 t + \phi_2)} \hat{y}|^2
= (E_x e^{-j(\omega_1 t + \phi_1)} \hat{x} + E_y e^{-j(\omega_2 t + \phi_2)} \hat{y}) \cdot (E_x e^{j(\omega_1 t + \phi_1)} \hat{x} + E_y e^{j(\omega_2 t + \phi_2)} \hat{y})
= |E_x|^2 + |E_y|^2
\]

This result simply acknowledges what is already known about interference; that for fields with orthogonal polarizations, the total intensity is the sum of the individual intensities and is a constant when time-averaged over a period greater than the signal period. Mathematically, the dot product in the second equality only allows parallel fields to “interact”, so no mixing terms are possible in this configuration. For the sake of completeness, consider the intensity of two collinear components of different frequency. If the fields in the above example both lie along the x axis, then the total intensity is given as

\[
I = |E_{x1} e^{-j(\omega_1 t + \phi_1)} \hat{x} + E_{x2} e^{-j(\omega_2 t + \phi_2)} \hat{x}|^2
= (E_{x1} e^{-j(\omega_1 t + \phi_1)} \hat{x} + E_{x2} e^{-j(\omega_2 t + \phi_2)} \hat{x}) \cdot (E_{x1} e^{j(\omega_1 t + \phi_1)} \hat{x} + E_{x2} e^{j(\omega_2 t + \phi_2)} \hat{x})
= |E_{x1}|^2 + |E_{x2}|^2 + 2E_{x1} E_{x2} \cos ((\omega_2 - \omega_1) t + (\phi_2 - \phi_1))
\]

The total intensity is not only the sum of the individual intensities, but also contains a third term that captures the interference between the two monochromatic waves. But unlike the first case with orthogonal field polarizations, the intensity for the collinear
fields is not constant, but rather has an additional beat term that varies sinusoidally as a function of the difference between the signal frequencies and their initial phases. Because the fields at 4b are orthogonal, it is necessary to mix the components by rotating the fields along a common axis. By inserting a polarization analyzer with its axis at 45° with respect to the x and y axes, half the power is passed in each component. The total intensity at detector $D_R$, the reference detector, after passage through the polarization analyzer, is thus

$$I_{D_R} = -j \frac{E_0}{\sqrt{8}} \left( e^{-j(\omega_2 t + \phi_2)} + e^{-j(\omega_1 t + \phi_1)} \right) \cdot j \frac{E_0}{\sqrt{8}} \left( e^{j(\omega_2 t + \phi_2)} + e^{j(\omega_1 t + \phi_1)} \right)$$

$$= \frac{E_0^2}{4} \left[ 1 + \cos \left( (\omega_2 - \omega_1) t + (\phi_2 - \phi_1) \right) \right]$$  \hspace{1cm} (2.10)

Finally, by analyzing the reference and measurement (probe) arms, it can be observed that there are few differences between the total intensity derived for the reference detector path and that of the measurement detector path. The signal at 4b passes through a polarizing beam splitter (PBS), where the orthogonal polarizations are split into what are referred to as the reference arm and the measurement arm. Because the two heterodyne frequencies were also polarization differentiated, one frequency is directed into the measurement arm by the PBS and the other into the reference arm. The signal from the measurement arm is directed at an object of interest through a quarter waveplate (QWP), and a phase shift is induced on this optical field proportional to the round-trip distance to the measured object.

Because the object may be in motion relative to the interferometer, this phase change can in general be written as a function of time $\phi(t)$. The derivative of this time-varying phase shift represents an apparent shift in the carrier frequency due to the Doppler effect, and both velocity and displacement of an object can be measured. Propagation loss, scattering, and absorption of the field at the target are accounted for by multiplying the original signal amplitude by a scaling factor $\gamma$, which is assumed real. Upon reflection, the return signal passes the QWP a second time. This double-pass through the QWP rotates the polarization of the incident field by 90°, so the reflected reference arm and measurement arm fields exit the PBS port marked
5 rather than returning through the PBS port coincident with point 4b. A Jones matrix analysis of a double-pass through the QWP is given in Appendix C. The reference arm, as the name implies, provides a phase reference for the measurement arm that allows a determination of the object’s displacement based on the phase shift with respect to the reference arm. It is not necessary to ensure that the phase difference between the reference and measurements arms are equal prior to measurement (balanced interferometer), as long as the path difference is less than the coherence length of the laser. The reference arm field also traverses the QWP twice to change its polarization by $90^\circ$ such that it recombines with the measurement arm field at the PBS. The combined fields prior to the measurement detector $D_M$ can be written as

$$E_0 \frac{e^{-j(\omega_2 t + \phi_2)}}{\sqrt{8}} - \gamma E_0 \frac{e^{-j(\omega_1 t + \phi_1(t))}}{\sqrt{8}} \hat{y}$$

(2.11)

This allows the total intensity at detector $D_M$, after the orthogonal fields are aligned by the polarizer, to be written as

$$I_{D_M} = \frac{E_0^2}{8} \left[ (1 + \gamma^2) - 2\gamma \cos ((\omega_2 - \omega_1) t + (\phi_2 - \phi_1 - \delta \phi(t))) \right]$$

(2.12)

which equals $\frac{E_0^2}{4} \left[ 1 - \cos ((\omega_2 - \omega_1) t + (\phi_2 - \phi_1 - \delta \phi(t))) \right]$ for $\gamma = 1$. Allowing the initial phase difference $\phi_2 - \phi_1$ to equal zero, the total intensity at the reference and measurement detectors can be recast as

$$I_{D_R} = \frac{E_0^2}{4} \left[ 1 + \cos (\Delta \omega t) \right]$$

(2.13)

$$I_{D_M} = \frac{E_0^2}{8} \left[ (1 + \gamma^2) - 2\gamma \cos (\Delta \omega t - \delta \phi(t)) \right]$$

(2.14)

where $\Delta \omega = \omega_2 - \omega_1$. 

32
2.2 Balanced Detection

In many applications, the power from the hybrid coupler is only collected by a single detector. Intuitively this seems like a waste of valuable signal, and indeed it has an effect on the obtainable performance of such systems. Achieving shot-noise-limited performance with single-detector receivers is difficult due to high laser intensity noise and insufficient local oscillator (LO) power, such that LO shot noise fails to dominate other receiver noise sources [16]. An easily realizable solution to this inefficiency is to place detectors at both coupler outputs; a configuration known as balanced detection and shown conceptually in Figure 2-2. Unlike single-detector systems where only half the available power is collected, balanced detection allows the capture of an additional 3-dB of signal power, with the added benefit of removing most of the DC intensity components [16–19]. In fact, any common-mode noise that limits detector sensitivity can be eliminated. Also, dividing the LO power across two detectors can avoid detector saturation at high input power.

The use of balanced detection has also been shown to remove a considerable portion of amplitude modulation noise [20–22]. To achieve the photon (shot) noise limit, a single frequency laser with narrow linewidth and low amplitude noise, combined with a balanced detector configuration, should be employed [16,20]. The impact of various noise contributors is demonstrated in a subsequent analysis, which will draw from the existing literature. Intensity fluctuations due to laser Relative Intensity Noise (RIN) in the LO laser are correlated between the two photodetectors, and are therefore canceled to a degree dependent on how well the detectors are matched in incident intensity and responsivity. The shot noise in each detector is uncorrelated and does not cancel. Again, the details will be deferred until the next chapter. Achieving an accurate and stable balance between the photocurrents and equal paths from the optical coupler to each photodetector is crucial to obtaining cancellation of wideband excess noise. The 3-dB couplers used in this design are adiabatic in nature. The transfer function for such couplers is seen to be that of a sum-difference (180°-hybrid)
coupler and is given by

$$T_{\Sigma, \Delta} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$  (2.15)

These couplers have demonstrated coupling ratios within a few percent of 50:50. If the coupling ratio is not exactly 50:50, there is incomplete cancellation of the common-mode noise terms. However, as band-pass filtering will be applied to the output voltages after amplification, the DC components will be eliminated in any case. Based on the transmission matrix for the coupler, the two input signals, $E_{1,IN}$ and $E_{2,IN}$, will produce outputs given by

$$E_\Sigma = \frac{1}{\sqrt{2}} (E_{1,IN} + E_{2,IN})$$  (2.16)
$$E_\Delta = \frac{1}{\sqrt{2}} (-E_{1,IN} + E_{2,IN})$$  (2.17)

Now, assume the output of the coupler described by Eq.(2.15) is incident on a pair of photodetectors in the configuration shown schematically in Figure 2-2. Both detectors are reverse-biased and work in a push-pull configuration. If the current is sampled at the node connecting the two photodetectors, the difference between the currents is obtained. To show why this configuration is advantageous, the power delivered by the sum and difference ports, and the resulting photocurrents at each detector, are

![Figure 2-2: Conceptual diagram of the balanced detection scheme.](image-url)
computed. Results from earlier in this section will be used to move from the fields at the sum and difference ports to currents at the respective photodetectors. First, the input fields at Port 1 and Port 2 are expressed as

\begin{align}
E_{1,IN} &= E_1 e^{-j(\omega_1 t + \phi_1)} \\
E_{2,IN} &= E_2 e^{-j(\omega_2 t + \phi_2)}
\end{align}

From here the fields and powers (intensity over a unit area) at the sum and difference ports can be easily written as

\begin{align}
E_\Sigma &= \left[ \frac{E_2}{\sqrt{2}} e^{-j(\omega_2 t + \phi_2)} + \frac{E_1}{\sqrt{2}} e^{-j(\omega_1 t + \phi_1)} \right] \\
E_\Delta &= \left[ \frac{E_2}{\sqrt{2}} e^{-j(\omega_2 t + \phi_2)} - \frac{E_1}{\sqrt{2}} e^{-j(\omega_1 t + \phi_1)} \right]
\end{align}

\begin{align}
P_\Sigma &\propto \vec{E}_\Sigma \cdot \vec{E}_\Sigma^* = \frac{|E_1|^2}{2} + \frac{|E_2|^2}{2} + E_1 E_2 \cos \left( (\omega_2 - \omega_1) t + (\phi_2 - \phi_1) \right) \\
P_\Delta &\propto \vec{E}_\Delta \cdot \vec{E}_\Delta^* = \frac{|E_1|^2}{2} + \frac{|E_2|^2}{2} - E_1 E_2 \cos \left( (\omega_2 - \omega_1) t + (\phi_2 - \phi_1) \right)
\end{align}

As shown previously, the power incident on the photodetectors has DC terms related to the intensities of the incident fields, and an oscillating component with frequency equal to the frequency difference of the incident fields. The second-order nonlinearity that arises from squaring the sum and difference fields also introduces a term with a frequency equal to the sum of the individual heterodyne frequencies. However, the photodetector cannot pass signals at optical frequencies and this term is filtered. This low-pass behavior of the photodetector is due to its shunt capacitance, which forms an equivalent RC circuit with the detector series resistance, as well as transit time effects. The detector photocurrent is related to the average incident power (intensity) via the responsivity \( R \), given in units of A/W, by the equation \( I_{PD} = R \cdot P_{PD} \). Furthermore, each photodetector can have a different responsivity owing to differences
in the quantum efficiency $\eta$, with $\mathcal{R}$ defined as

$$\mathcal{R} = \eta q / h \nu \text{ (A/W)} \quad (2.24)$$

Denoting the responsivity of the sum and difference detectors as $\mathcal{R}_\Sigma$ and $\mathcal{R}_\Delta$ and applying them to the respective incident powers, expressions for the photocurrents produced by the sum and difference detectors are readily obtained.

$$I_\Sigma = \mathcal{R}_\Sigma \left( \frac{|E_1|^2}{2} + \frac{|E_2|^2}{2} + E_1 E_2 \cos(\ldots) \right)$$

$$= I_{\Sigma,DC} + \mathcal{R}_\Sigma \sqrt{P_1 P_2} \cos((\omega_2 - \omega_1) t + (\phi_2 - \phi_1)) \quad (2.25)$$

$$I_\Delta = \mathcal{R}_\Delta \left( \frac{|E_1|^2}{2} + \frac{|E_2|^2}{2} - E_1 E_2 \cos(\ldots) \right)$$

$$= I_{\Delta,DC} - \mathcal{R}_\Delta \sqrt{P_1 P_2} \cos((\omega_2 - \omega_1) t + (\phi_2 - \phi_1)) \quad (2.26)$$

Here the following definitions are used for the DC currents.

$$I_{\Sigma,DC} = \mathcal{R}_\Sigma \left( \frac{|E_1|^2}{2} + \frac{|E_2|^2}{2} \right) ; I_{\Delta,DC} = \mathcal{R}_\Delta \left( \frac{|E_1|^2}{2} + \frac{|E_2|^2}{2} \right) \quad (2.27)$$

The difference between the sum and difference photocurrents yields the following expression for the balanced current.

$$I_B = I_{\Sigma,DC} - I_{\Delta,DC} + (\mathcal{R}_\Delta + \mathcal{R}_\Sigma) \sqrt{P_1 P_2} \cos(\omega_{IF} t + \delta \phi) \quad (2.28)$$

The frequency difference $\omega_2 - \omega_1$ is rewritten as $\omega_{IF}$, the initial phase difference $\phi_2 - \phi_1$ as $\delta \phi$, and the product of the fields is written in terms of the respective powers. For responsivities at the sum and difference detectors that are close in value, the DC terms are suppressed and the AC term is amplified. In fact, in the limit of equal responsivities, the DC term is completely canceled and the AC term is doubled. The balanced detector in the presence of noise is revisited in a subsequent section, and the
benefit of this configuration in the presence of correlated noise will become apparent.

This framework is used extensively in the subsequent analysis of the integrated heterodyne interferometer. Because the integrated interferometer utilizes waveguides for both the Mach-Zehnder and Michelson sections, it is not necessary to spatially separate the tones using orthogonal polarizations. So while the integrated heterodyne interferometer is an inherently more difficult device to design, fabricate, and characterize because of its micron-scale components, and hence sensitivity to the fabrication process; the elimination of polarization elements can simplify the device complexity considerably. However, if errors in the polarization devices are ignored, as they were here, the transfer functions for both devices should be equivalent. The next section will focus on the operation of the interferometer in the real world, and examine some factors that cause it to deviate from the ideal assumptions in this section.

2.3 Displacement Resolution and Measurement Limitations

In this section the performance metrics of a heterodyne interferometer are extensively analyzed. In particular, the displacement resolution, or the smallest measurable change in displacement, is of interest. Depending on the intended application of the interferometer, it may also be desirable to measure velocity using the Doppler frequency shift induced by an object’s motion. In this case the achievable frequency resolution is also of interest. One commonly used measure of the maximum displacement is the unambiguous range. The unambiguous range is determined by the optical wavelength of the measurement signal, and is attributable to the inherent modulo $2\pi$ nature of phase measurements. This phase ambiguity, and strategies to resolve it, are discussed in due course. However, the true upper performance limit on the displacement measurement is set by the laser coherence length [23], which is inversely proportional to the laser linewidth.

There are additional performance limitations inherent to the interferometer that
do not exist in coherent communications systems. A free space heterodyne interferometer, such as the one analyzed earlier, utilizes polarization to separate the fields representing the measurement and reference arms. Because the bulk interferometer incorporates optical components that modify the polarization, errors will be introduced resulting from the imperfect nature of these components. As an example, phase rotators designed to change the polarization of a wave from TE to TM or vice-versa have a finite extinction ratio, which allows undesired polarization components to pass into each arm. A component such as a polarizing beam splitter that spatially separates orthogonal polarizations will also allow some polarization leakage. In addition to polarization mixing in the polarization-dependent devices, polarization errors can also be introduced by the laser source. Instead of a linearly polarized laser source, actual laser sources often have elliptical polarizations. This crosstalk will introduce contamination in the fields that must be taken into account. The consequences of a similar effect due to frequency crosstalk in the modulators is observed in this device, and a method is presented to compensate for these errors when the crosstalk is not excessive. All displacement measuring interferometers are inherently subject to a $2\pi$ phase ambiguity. The interferometer determines displacement by measuring the phase shift induced by a change in the measurement path length as compared to a reference path of fixed length. But this phase shift is confined to a maximum of $2\pi$ due to the periodic nature of the exponential, such that

$$e^{j(\phi + m2\pi)} = e^{j\phi}$$

(2.29)

for $m$ equal to an integer. To see why this is a limitation in interferometric measurements, assume there are two carriers with some initial phase difference between them. These carriers traverse different optical paths and are later recombined. If one carrier traverses an additional round-trip distance $\Delta L$ with respect to the other, an additional relative phase shift is introduced between them given by

$$\Delta \phi = 2\pi \frac{\Delta L}{\lambda}$$

(2.30)
where \( \Delta L \) is the additional round-trip delay experienced by a carrier with wavelength \( \lambda \). This expression can be written in terms of the one-way displacement \( \Delta d \) using \( \Delta L = 2\Delta d \) such that
\[
\Delta \phi = 4\pi \frac{\Delta d}{\lambda}
\] (2.31)

If \( \Delta L \) is the round-trip distance between the interferometer and the object to be measured, the maximum one-way distance that can be measured without ambiguity is \( \lambda/2 \), such that the round-trip displacement \( L = \lambda \) corresponds to a \( 2\pi \) phase shift. For round-trip displacements greater than \( \lambda \), the total displacement can be written as \( L = L_{amb} + \Delta L \). Here \( L_{amb} = m\lambda \) and \( m \), an integer greater than or equal to one, represents the number of whole wavelengths in the round-trip. \( \Delta L \) is the incremental distance between \( m\lambda \) and \( (m + 1)\lambda \), and can be considered the remainder when the length is measured modulo \( \lambda \). The total phase shift can then be expressed as
\[
\Delta \phi = 2\pi \frac{L_{abs} + \Delta L}{\lambda} = 2\pi \left( \frac{m\lambda + \Delta L}{\lambda} \right) = 2\pi \left( m + \frac{\Delta L}{\lambda} \right) = 2\pi m + \Delta \phi_{frac} \] (2.32)

The fractional phase shift \( \Delta \phi_{frac} \) is the phase shift remaining after wrapping the phase modulo \( 2\pi \). For displacements less than \( \lambda/2 \), \( m \) equals zero and the total phase shift equals the fractional phase shift as expected. This demonstrates that for round-trip delays larger than \( \lambda \), a phase ambiguity occurs that aliases one-way distances greater than a half-wavelength into distances less than a half-wavelength. This phase wrapping is a well-known limitation and is illustrated in Figure 2-3. For motion in one direction, the true phase shift as a function of displacement is a continuous monotonic function. But the measured displacement is limited by the aforementioned ambiguity and repeats every half-wavelength. This aliasing is not an inherent property of the motion, but rather an artifact of how the phase is computed. Discussions in the following sections will focus on methods to overcome the phase ambiguity.
2.3.1 Two Wavelength Measurements for Increased Unambiguous Range

One method for solving the range ambiguity is to use either a single source with a longer wavelength, or to use two or more sources to synthesize a longer equivalent wavelength. The latter method has been used since the 1970s and is commonly known as synthetic wavelength or multiple-wavelength interferometry. Some of the earliest applications of these techniques were in holographic optical measurements [24–27]. However, these methods were cumbersome and required exacting requirements of the testing environment that diminished their usefulness outside the laboratory. As a brief introduction, synthetic heterodyne interferometry uses two or more wavelengths that are close in value to create a synthetic wavelength considerably longer than the individual wavelengths. In particular, the synthetic wavelength can be viewed as deriving simultaneous phase measurements of the same object using wavelengths $\lambda_1$ and $\lambda_2$ such that

$$\phi_1 = 4\pi \frac{d}{\lambda_1}; \phi_2 = 4\pi \frac{d}{\lambda_2}$$  \hspace{1cm} (2.33)

Since phase is measured modulo $2\pi$, the above expressions can be written without loss of generality as

$$\phi_1 = 2\pi \left( m + 2 \frac{d}{\lambda_1} \right) \mod 2\pi; \phi_2 = 2\pi \left( n + 2 \frac{d}{\lambda_2} \right) \mod 2\pi$$  \hspace{1cm} (2.34)
where \( m \) and \( n \) are integers. The difference in phase using the two measurements is

\[
\Delta \phi = \phi_2 - \phi_1 = \left[ 2\pi (n - m) + 4\pi d \frac{\lambda_2 - \lambda_1}{\lambda_2 \lambda_1} \right] \mod 2\pi = 4\pi \frac{d}{\Lambda_{eq}} \tag{2.35}
\]

The synthetic wavelength, also known as the equivalent wavelength, is given by

\[
\Lambda_{eq} = \frac{\lambda_2 \lambda_1}{|\lambda_2 - \lambda_1|} = \frac{\lambda_2 \lambda_1}{\Delta \lambda} \tag{2.36}
\]

Table 2.1 shows examples of the relatively large synthetic wavelengths and unambiguous ranges that can be obtained for a small difference in wavelength. Utilizing a small wavelength difference, it is possible to derive a large synthetic wavelength that produces the same interferometric results as a single wavelength of equivalent length. Wavelength separations of the order listed in Table 2.1 are well within the capabilities of available lasers such as the Agilent 81600B, which can be tuned with one picometer resolution. One useful effect is that the unambiguous range of the interferometer is increased accordingly, allowing for longer absolute measurements. The corollary is that longer wavelengths produce fewer interference fringes over a given distance than shorter wavelengths, hence reducing the overall sensitivity of the device \[24\]. In some applications this reduction in sensitivity may be desired. However, this is not the case when extremely accurate measurements on the order of the individual wavelengths \( \lambda_1 \) and \( \lambda_2 \) are of interest.

As is typically the case, there exists a performance trade-off in that longer wavelengths can measure longer distances before reaching the phase ambiguity. However, for displacements much less than a wavelength, the distances becomes a smaller fraction of the total wavelength, resulting in phase changes increasingly less than \( 2\pi \).
These small phase changes may be less than the minimum phase resolution imposed by noise and other error sources within the system, which ultimately reduce the obtainable displacement resolution. Therefore the mode of operation must be chosen to satisfy the task at hand. Because the heterodyne interferometer is a dual-frequency device by nature, it should be possible to operate in a high-fidelity mode for measuring sub-wavelength scales phenomena using a single carrier frequency, and a low-fidelity mode using both frequencies to unambiguously measure longer lengths. An additional limitation becomes apparent from considering the relationship between the two phase measurements. The ratio of the two phase measurements is equal to

$$\frac{\phi_1}{\phi_2} = \frac{\lambda_2}{\lambda_1} = N \quad (2.37)$$

Equivalently, the displacement in terms of the interference orders $N_1$ and $N_2$ of the two wavelengths can be written as

$$d = N_1\lambda_1 = N_2\lambda_2 \quad (2.38)$$

The ratio of the interference orders for a two-frequency measurement of the same displacement can be written in terms of the ratio of the individual wavelengths as

$$\frac{N_1}{N_2} = \frac{\lambda_2}{\lambda_1} = N \quad (2.39)$$

For the wavelength pairs given in Table 2.1, the ratio $N$ is given in Table 2.2. This

<table>
<thead>
<tr>
<th>$\lambda_1$(nm)</th>
<th>$\lambda_2$(nm)</th>
<th>$\Lambda_{eq}$(mm)</th>
<th>$N$</th>
<th>$\lambda_1$(nm)</th>
<th>$\lambda_2$(nm)</th>
<th>$\Lambda_{eq}$(mm)</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1550</td>
<td>1548</td>
<td>1.2</td>
<td>1.0013</td>
<td>1550</td>
<td>1549.7</td>
<td>8.0</td>
<td>1.0002</td>
</tr>
<tr>
<td>1550</td>
<td>1548.5</td>
<td>1.6</td>
<td>1.001</td>
<td>1550</td>
<td>1549.8</td>
<td>12.0</td>
<td>1.00013</td>
</tr>
<tr>
<td>1550</td>
<td>1549</td>
<td>2.4</td>
<td>1.001</td>
<td>1550</td>
<td>1549.9</td>
<td>24.0</td>
<td>1.0001</td>
</tr>
<tr>
<td>1550</td>
<td>1549.1</td>
<td>2.7</td>
<td>1.001</td>
<td>1550</td>
<td>1549.95</td>
<td>48.0</td>
<td>1.00003</td>
</tr>
</tbody>
</table>

Table 2.2: N-ratio for several synthetic wavelength pairs.

shows that as the wavelengths become closer, so do their orders. The interference order $N$ is a real number consisting of an integer (number of full wavelengths) and an excess component that gives the fractional displacement. As was put forth initially,
large synthetic wavelengths require closely spaced λs with orders that are equal to within a fraction of one percent. But this ratio also describes the relationship between the measured phases φ₁ and φ₂. This places a requirement on the minimum achievable phase resolution that the system must provide to resolve the two individual phase measurements. To achieve a synthetic wavelength of 1.2 mm requires a phase resolution of about one part-per hundred, whereas a synthetic wavelength of 48 mm requires a resolution of about thirty parts-per million, which is a formidable requirement. In the next section, an estimate of the achievable phase resolution of a system with a given Signal-to-Noise Ratio (SNR) is derived. Lastly, the displacement in terms of the synthetic wavelength Λ_EQ and the interference order can be written as

\[ d = N_{EQ} \Lambda_{EQ} \]  

(2.40)

Using the equations previously presented, \( N_{EQ} \) is given by

\[ \phi_2 - \phi_1 = 4\pi N_2 - 4\pi N_1 = 4\pi \frac{d}{\Lambda_{eq}} = 4\pi N_{EQ} \]  

(2.41)

\[ |N_{EQ}| = |N_2 - N_1| = |N_2 (1 - N)| \]  

(2.42)

By definition, for displacements less than \( \Lambda_{EQ}/2 \), \( N_{EQ} \) is less than one. This is indeed born out using the results of Table 2.2.

### 2.4 Discrete vs. Continuous Phase Measurements

The theory of displacement measurement via phase and its limitations were presented in the previous section. That single wavelength measurements lead to high resolution, but only offer unambiguous lengths of a half-wavelength, was also discussed. To achieve longer unambiguous lengths, either the wavelength of a single wavelength system must be increased, or multiple, closely-spaced wavelengths must be used to make simultaneous displacement measurements. But measurements conducted with
synthetic interferometry are still limited to half the synthetic wavelength before the phase ambiguity resurfaces. One method of operating the heterodyne interferometer is to only take measurements when displacement information is required, such that phase measurements occur asynchronously at discrete times. This approach is simple and does not impose excessively onerous data storage requirements. However, a problem with this approach is that if the object has moved more than a half-wavelength between measurements, then it is impossible to unwrap the phase ambiguity. The ambiguity can be eliminated only if the maximum displacement between measurements is guaranteed to be limited to less than one-half the wavelength.

One strategy to eliminate the ambiguity is to sample the phase at regular intervals at an appropriate rate. This requires a more complex data acquisition system and higher storage requirements. But there are several advantages to synchronously measuring the phase. If an object moves between two positions $x_A$ and $x_B$ during a time interval $t_A$ to $t_B$, then it does so, assuming uniform motion, according to the kinematic equation

$$x_B = x_A + v_{B,A}(t_B - t_A)$$

where $v_{B,A}$ is the velocity in m/s. So if measurements are taken frequently enough, it should be possible to identify phase discontinuities and unwrap them appropriately. The question is how frequently is enough? The very nature of this question will lead naturally to sampling and the Nyquist-Shannon Theorem. How fast the displacement is measured is directly tied to how fast the object is moving. The Doppler effect is a shift in the frequency of a wave or periodic signal due to the line-of-sight motion between a wave source and target. In general, the frequency shift observed between the source and a target whose velocity vector makes an angle $\theta$ with the line of sight between the two is

$$f_d = \frac{\pm 2v_t \cos (\theta)}{c} f_c \text{ (Hz)}$$

where $v_t$ is the velocity of the target in m/s, $\theta$ is the angle between the target’s velocity vector and the line of sight, $f_c$ is the frequency of the source in Hz, and $c$ is the speed of light in vacuum. The factor of two accounts for round-trip propagation
to the target and back in a monostatic (single antenna) system. It should be noted that this is the non-relativistic form of the Doppler effect. The sign indicates whether the target is moving toward or away from the source. So a target moving toward the source will modulate the frequency upwards (positive) and a receding target will modulate the frequency downwards (negative). For example, a target moving at 1 m/s directly towards an optical source with a wavelength of 1.55 μm will increase the source frequency by approximately 1.29 MHz. If both receding and advancing targets with this velocity are considered, the total bandwidth imparted to the carrier is 3.58 MHz. The non-relativistic frequency shift can also be derived from the time derivative of the round-trip phase according to

$$f = \frac{1}{2\pi} \frac{d\phi}{dt} = \frac{1}{2\pi} \frac{4\pi dR/dt}{\lambda} = \frac{2v_t}{c} f_c$$

This is indeed equivalent to the non-relativistic Doppler shift when the target motion is along the line of sight. Consequently, the Doppler-modulated carrier imposes a larger bandwidth requirement on the system to capture the wider frequency content. But there is a catch. In Chapter 3, the variance of the phase noise is shown to be inversely proportional to the SNR. Because the increased bandwidth requirement leads to a reduction in SNR due to increased noise, this makes fine displacement resolution more difficult and necessitates a trade-off between the maximum velocity and the maximum displacement resolution. The maximum velocity is also limited by the reference (IF) frequency \[28\]. To avoid aliasing, it is necessary to ensure that the lowest frequency present in the signal spectrum is greater than zero. In other words, the maximum Doppler frequency and the IF must satisfy

$$f_{IF} - f_d > 0$$

(2.46)

If this relationship doesn’t hold, negative frequency components will overlap (alias) with positive frequency components near zero Hz, making accurate recovery of the signal difficult, if not impossible. The final requirement is to sample the measurement and reference signals at the Nyquist-Shannon frequency, which is at least twice the
maximum bandwidth of the spectrum $2f_d$. This may appear to violate the Nyquist theorem regarding the minimum sampling rate to accurately reconstruct the signal. For a band-pass signal centered at the intermediate frequency, this would imply a sampling rate $f_{\text{sample}} > 2(f_{IF} + f_d)$. The subtlety of the sampling theorem and the Nyquist criteria is that it applies to a baseband signal with bandwidth $B$ centered about the frequency origin $[29,30]$. To avoid overlap between the true spectra and the aliased components, the sampling frequency must be greater than twice the bandwidth $B$. That $B$ represents the highest frequency is simply a consequence of the waveform being centered about the origin ($f = 0$) such that $f_{\text{max}} = B$. However, the overlap is actually a function of the bandwidth itself. The Nyquist-Shannon theorem states that the sampling rate is not twice the highest frequency, but rather twice the bandwidth $B$. For baseband signals these two statements are equivalent, but may have different implications for band-pass signals.

For an IF signal centered at 10 MHz with a bandwidth of 2 MHz, the minimum sampling rate that can be achieved without aliasing is not 24 MHz (2x12 MHz), but rather 4 MHz (2x2 MHz). The ability to sample a band-pass signal at a rate below the Nyquist rate such that complete signal reconstruction is possible is known as undersampling, or band-pass sampling. This concept will not be explored further, but it is important to understand this technique when considering requirements for an analog-to-digital (ADC) converter. If a trade-off between the speed of the ADC and the resolution (number of bits) is required, undersampling can be utilized to handle a band-pass signal as though it were a baseband signal, allowing an increase in ADC resolution and a simultaneous reduction in quantization noise. Returning to the question of how frequently measurements must be made to capture displacements larger than a half-wavelength, please refer to Figure 2-3. If a measurement is taken at a particular point, then the object is moved by an integer multiple of half-wavelengths to a new point before another measurement is taken, the measured phases at both points will be identical. So without additional information, it is not possible to resolve the ambiguity when displacements greater than a half-wavelength are encountered. However, if the object moves at a maximum velocity $v_{\text{max}}$, it may be surmised, without
first resorting to the frequency domain, that if at least two samples are collected during
the time it takes the object to move a half-wavelength at \( v_{\text{max}} \) that the phase can be
resolved unambiguously. The time to transit a half-wavelength is given by

\[ t_{\text{half}} = \frac{\lambda}{2v_{\text{max}}} \text{ (s)} \]  \hspace{1cm} (2.47)

If two measurements are taken during this period, the required sampling rate is given
by

\[ t_{\text{sample}} = \frac{\lambda}{4v_{\text{max}}} \text{ (s)} \]  \hspace{1cm} (2.48)

Finally, writing \( v_{\text{max}} \) in terms of the Doppler frequency

\[ v_{\text{max}} = \frac{f_{\text{Doppler}}}{2} \lambda \text{ (Hz)} \]  \hspace{1cm} (2.49)

the required sampling rate for an object under uniform motion is given by

\[ t_{\text{sample}} > \frac{1}{2f_{\text{Doppler}}} \text{ (s)} \]  \hspace{1cm} (2.50)

This period is inversely proportional to twice the Doppler frequency, and so is equiva-
 lent to sampling at twice the Nyquist-Shannon frequency. This result will be revisited
when demonstrating the vibrometer mode, where the target motion may simulta-
neously introduce a Doppler frequency shift and multiple phase ambiguities. The
next chapter will examine the noise sources of the interferometer by applying well-
established principals of heterodyne system analysis. Following the noise analysis,
bounds can be placed on the achievable resolution and overall system performance.
THIS PAGE INTENTIONALLY LEFT BLANK
Chapter 3

Noise and Performance Analysis

3.1 Background

Interferometers are in theory capable of making long-range measurements with high resolution. However, in most applications of interest, both attributes are not simultaneously required. When attempting long-range measurements, say on the order of kilometers and spanning many multiples of the optical wavelength, fine position resolution is likely not a requirement. But in metrology, where the goal is often sub-wavelength position measurement, the goal is typically to achieve the theoretical resolution \[23\]. Ultimately, an interferometer’s capabilities can be viewed as being limited by noise through the Signal-to-Noise Ratio (SNR) \[31\]. With regards to SNR, the goal is to achieve shot-noise-limited performance such that other forms of noise are negligible \[32\].

This section will briefly review the primary sources of noise in a coherent system, and under what conditions the optimal shot-noise-limited operation can be achieved. Heterodyne interferometry shares many similarities to angle modulated (FM/PM) communication systems in that such systems are susceptible to amplitude noise only to the extent that the carrier frequency or phase is modified by the noise \[33\]. First, the primary sources of amplitude noise are characterized before demonstrating how amplitude noise is coupled into phase noise. Sources of phase noise, such as the laser, signal generators, environmental vibrations, and atmospheric turbulence are
A sinusoidal signal can always be written in terms of an in-phase, multiplicative amplitude modulation (AM) noise term $a(t)$, a quadrature phase noise term $\phi(t)$, and an additive noise term $n(t)$ such that

$$s(t) = [A_0 + a(t)] \sin(2\pi ft + \phi(t)) + n(t)$$  \hspace{1cm} (3.1)

The effect of quadrature phase noise on the interferometer is clear, as it directly alters the phase and introduces timing jitter onto the waveform. The severity of the phase noise terms will depend on how well-correlated they are between the reference and measurement paths. The AM noise in well-designed oscillators is typically small and can be minimized. The role of additive noise is not as obvious, but will be made clearer in this chapter. Because the heterodyne interferometer encodes displacement and velocity measurements in the phase and frequency of the photocurrent, the noise performance is intimately linked to the overall performance of the system.

A considerable benefit of heterodyne interferometry is that analysis techniques developed for coherent communications systems, such as those detailed extensively in [32, 34], can be readily applied to an interferometric system. In such a system, the goal is to transmit signals with sufficient power such that detector shot noise dominates the thermal noise of the receiver electronics, which is the dominate noise in microwave systems [20]. Heterodyne systems, and coherent systems in general, achieve this performance by optically multiplying a strong, locally-generated local oscillator (LO) signal with a relatively weak received signal prior to photodetection. The multiplication of the strong LO with the weaker received signal improves the detection of the received signal in the presence of thermal noise, providing a form of optical amplification [17, 32]. Therefore, doubling the power of the LO leads to a 3-dB increase in the carrier power. However, increasing the LO power directly increases shot noise, and the LO power is typically chosen such that shot noise exceeds thermal noise, if possible. This usually occurs for a total LO and received power into the detector on the order of 1 mW [17, 20]. In this regime the system is shot-noise-limited and yields the best possible SNR as desired for optimal performance [17].

considered only at a qualitative level. A sinusoidal signal can always been written in terms of an in-phase, multiplicative amplitude modulation (AM) noise term $a(t)$, a quadrature phase noise term $\phi(t)$, and an additive noise term $n(t)$ such that
The multiplication, or beating, of the LO and the received signal at the photodetector is also responsible for producing an information bearing signal at an RF frequency known as the intermediate frequency (IF), which was demonstrated in an earlier analysis of the interferometer transfer function.

In practice, the primary impediments to achieving shot-noise-limited performance at optical frequencies are the amplitude and phase noise of the laser source [20]. But by deriving the LO and measurement signals coherently from the same source, modulation effects due to beating can be eliminated. Because laser intensity fluctuations and other low frequency terms are centered around baseband (\( f = 0 \)), they can be spectrally separated from the information bearing signals at the IF, allowing the phase modulation to be recovered with a minimum of distortion [32][35]. To further minimize the effect of laser amplitude noise and other forms of correlated noise, the photodetectors are employed in a balanced optical configuration [36]. Balanced detection was presented in Section 2.2 and it was shown how balanced detection could improve system performance by increasing SNR.

This section will attach additional meaning to the noise terms used in the balanced detector derivation. These results will then be used to derive a shot-noise-limited SNR. The sensitivity of an interferometer is limited by a number of environmental and noise sources [37]. In bulk interferometers, mechanical vibrations, thermal expansion of components, and fluctuations in the refractive index of air have the most deleterious effect on performance in part because bulk systems consist of optical components separated by free space distances on the order of tens of centimeters to several meters. Thermal gradients may cause unequal expansion of components and/or deviations of the refractive index at different points within the interferometer’s path. Due to the compact and integrated nature of the heterodyne interferometer, environmental perturbations are assumed to affect the entire chip equally. Furthermore, because the only free space path is between the interferometer and the device under test, the optical and electronic noise sources are assumed to provide the primary limitations on sensitivity.
3.1.1 Additive Noise Sources

Optical Background Noise

Receivers in terrestrial and space-based optical communication systems collect not only the desired information-bearing signal, but also undesired optical radiation from a number of ambient and man-made sources. This optical background noise adds to the desired signal and degrades overall system performance. The optical background noise is normally characterized as either extended background or point source \[32\]. Extended sources of background noise such as the sky are assumed to occupy the entire receiver field of view and are always present, depending on receiver orientation. Point sources, on the other hand, are localized and may be far more intense than extended background sources. In a space-based system, the sun and celestial objects are examples of point objects, whereas in terrestrial systems any object along the path between the transmitter and receiver may act as a source of background noise.

In the laboratory environment, reflecting walls assume the role of the extended background, while lights and other reflecting objects act as localized sources \[32\]. A description of optical background noise is included because it can represent a non-trivial contribution to the overall noise budget in an optical communication system, depending on the operating environment. The optical background noise is ignored in subsequent analysis with the assumption that background noise is negligible compared to photodetection and electronics noise. This decision is based on several reasonable assumptions.

1. The interferometer will be operated in a laboratory environment where the system can be shielded from most sources of extraneous optical background by surrounding it with a suitable enclosure.

2. The interferometer will measure displacements over a distance of only tens of centimeters, so the measured signal at the input aperture will be much stronger than the optical background. This assumption may need to be revisited when operating the device as a long-baseline LIDAR.
Laser Excess Intensity Noise

The laser source contributes noise that in general reduces measurement sensitivity. Yet, a primary strength of the heterodyne configuration is that because the reference and measurement signals are derived from the same source, variations in the laser amplitude have little impact on interferometer performance, allowing for extremely sensitive phase measurements [38]. Laser noise generally consists of two components: an amplitude or intensity noise that can be modeled as an in-phase amplitude modulation of the signal $a(t)$, and a quadrature phase noise term $\phi(t)$ which introduces a time-dependent phase modulation such that the instantaneous laser output becomes

$$s(t) = [A + a(t)] \sin(2\pi ft + \phi(t))$$ (3.2)

The optical noise in a laser is primarily attributed to quantum mechanical fluctuations within the laser gain medium and vacuum fluctuations due to optical losses. However, variations induced by mechanical and/or thermal perturbations within the laser cavity may also generate amplitude and phase noise. Because the interferometer LO and measurement signals are derived from a single source, the intensity noises of the two signals are highly correlated and the use of balanced detection is highly effective at canceling this noise.

Thermal (Johnson-Nyquist) Noise

Of the noise sources that must be considered when analyzing performance, thermal noise is one of the two most dominant forms. Thermal noise is an additive term that is always present in dissipative resistive elements when their temperature is above absolute zero, whether a signal is present or not. This is due to the random motion of thermally-agitated charge carriers in the resistive elements [33, 34]. Thermal noise is familiar to anyone working in radio frequency (RF) and microwave engineering, as it is the dominant noise in this frequency regime [34, 39]. Thermal noise processes are assumed to possess several well-known statistical characteristics. Thermal noise is assumed to be a stationary, ergodic, zero-mean Gaussian process [33]. Specifically,
the thermal noise is assumed to be wide-sense stationary (WSS) with constant mean and an auto-correlation function $R_{xx}(\tau)$ that depends only on the interval $\tau$ between measurements, not the absolute start and finish times. The process is ergodic if the time average of a sample function is equal to the ensemble average of all the sample functions. These definitions are widely used to characterize random processes [33,40,41]. Because thermal noise is assumed to consist of random contributions from a large number of independently moving electrons within the resistive element, the thermal noise has a probability distribution that is approximately Gaussian via the Central Limit Theorem [33,34,41]. The last assumption is that the thermal noise power spectrum is that of a “white” process, or one with equal power in each 1 Hz of bandwidth. Therefore, the spectrum of a true white noise process is flat across all frequencies. Of course a true white noise spectrum is a mathematical idealization; however, the approximation usually holds well over a wide range of frequencies. The one-sided power spectral density for thermal noise, derived quantum mechanically [42,43], has the well-known form given by [32,34,39]

$$S_{tn}(f) = \left(\frac{h\nu}{e^{\frac{h\nu}{kT}} - 1}\right) \left(\frac{W}{Hz}\right)$$

(3.3)

where $h = 6.546 \times 10^{-34}$ J-sec is Planck’s constant, $k = 1.380 \times 10^{-23}$ J/K is the Boltzmann constant, $T$ is the temperature of the resistive element in kelvin (K), and $\nu$ is the signal frequency in Hz and is equivalent to $f$. From this expression it is clear that the thermal noise spectrum is not truly white, but decays exponentially towards zero as $h\nu$ becomes much greater than $kT$, which corresponds to optical frequencies. For frequencies in the radio and microwave range (MHz-GHz) and temperatures in the operating range of most electronics, the thermal energy may exceed the optical energy by several orders of magnitude since $kT \gg h\nu$ in this regime. This condition leads to the Rayleigh-Jeans approximation [39], which states that in the limit of high temperatures or long wavelengths, the argument of the exponential is much smaller
than unity. By invoking the Taylor series expansion of the exponential,

\[ e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots, \quad -\infty < x < \infty \]  

(3.4)

and realizing that higher-order terms in the series are negligible compared to the first two terms, the power spectral density can be rewritten as

\[ S_m(f) \approx \frac{h\nu}{1 + \frac{h\nu}{kT} - 1} = kT = N_0 \left( \frac{W}{\text{Hz}} \right) \]  

(3.5)

At room temperature (T \approx 300 K), the power spectral density of thermal noise in dBm is given by the well-known value of \( S_m(f) = -174 \) dBm/Hz. This form of the Johnson noise density represents the power delivered to the rest of the system by a noisy resistor. However, it is not as intuitive when analyzing the noise performance of a circuit how this noise should be treated. A more intuitive model is one that considers the noisy resistor as either a Thevenin equivalent circuit of a noisy voltage source in series with a noiseless resistor, or its Norton equivalent of a noisy current source in parallel with a noiseless resistor. Both approaches are equivalent and are well-established from basic circuit analysis. So the form chosen is the one that makes the circuit analysis easier. The Thevenin equivalent circuit is considered first, with the results for the Norton equivalent circuit stated afterwards.

The equation for the Johnson noise power spectral density given in Eq. (3.5) represents the noise power delivered to the system from a noisy resistor. According to the maximum power transfer theorem, the maximum power that can be delivered to a resistive load \( R_L \) occurs when \( R_L \) equals the source resistance \( R_S \). In this case, the source impedance is the Thevenin resistance, or equivalently, the resistance of the noisy resistor under examination. If the noisy resistor \( R_S \) is replaced with a voltage source in series with a noiseless resistor \( R_S \), the power delivered into a matched resistance such that \( R_S = R_L \) is given by

\[ P_{\text{load}} = V_{\text{load}}I_{\text{load}} = \frac{V_{\text{in}}^2 R_L}{(R_s + R_L)^2} = \frac{V_{\text{in}}^2}{4R_s} \text{ (W)} \]  

(3.6)
Given the thermal noise power spectral density in Eq. (3.5), an equivalent voltage spectral density can be derived. Dropping the subscript on the resistance, the voltage spectral density is given by

\[ V_{tn}(f) = \sqrt{4R_sS_{tn}(f)} = \sqrt{4kT R \left( \frac{V}{\sqrt{\text{Hz}}} \right)} \] (3.7)

The Norton equivalent current noise density can be written as

\[ I_{tn}(f) = V(f)_{tn} = \sqrt{\frac{4kT}{R}} \left( \frac{\text{A}}{\sqrt{\text{Hz}}} \right) \] (3.8)

These expressions will be useful when examining the noise characteristics of demodulator circuits and their effect on overall system performance. Before proceeding, it is instructive to understand why a noise process that dominates at microwave frequencies and lower is important in an optical system. This is because the detection process converts an optical signal into a radio frequency signal. So the dominate noise sources are shot noise added during the photodetection process, and thermal noise added to the RF signal in the receiver electronics. Because thermal noise and shot noise dominate so strongly in their respective regimes, operating where quantum noise is the dominate noise source maximizes receiver sensitivity.

**Quantum (Shot) Noise**

Due to the quantum theory of light, electromagnetic fields are understood to have both wave and particle properties. In an optical receiver, photodetection is utilized to convert the optical signals into an electrical signal suitable for signal processing and analysis. The details of the photodetection process will not be explored, as there are numerous treatments of this important topic. However, a review of the theory related to shot noise processes and key properties that have implications for optical system performance is warranted. The optical field is composed of discrete particles known as photons, each carrying an exact quantity of energy directly proportional to the optical frequency. In the photodetector, photons with sufficient energy interact
with either a semiconductor material in photodiodes and transistors, or the metallic cathode in a vacuum tube to excite an electron from the valence band of the material into the conduction band. Once in the conduction band the electron is free to move about under an applied voltage to produce a current.

The process of photodetection is not perfect, and hence only a fraction of the incident photons lead to an electron being promoted to the conduction band. This quantity, denoted $\eta$, is known as the quantum efficiency of the detector. The efficiency of the detector is not constant, but rather represents a statistical average. The quantum efficiency is also highly dependent upon the semiconductor material and ranges from approximately 60-90% in a silicon PIN photodiode to 70-90% in an InGaAs PIN diode. From a statistical perspective, noise is typically defined as any stochastic deviation of a signal from its average value. So it is not surprising that the statistical nature of the quantum efficiency produces noise in the detector. From a quantum mechanical perspective, an efficiency less than unity can be accounted for by assuming that photons which fail to generate electrons are in vacuum states. These vacuum states, while containing no particles on average, do have fluctuations that produce noise. However, this is not the noise source of greatest concern in the photodetection process. The photons comprising the optical field each have energy $h\nu$ J/photon, and for an incident optical field of power $P$ J/s, the rate of photon arrival in number per second is given by

$$\bar{N}_{\text{photon}} = \frac{P}{h\nu} \left( \frac{\text{photons}}{s} \right) \quad (3.9)$$

The electron arrival can be related to the photon arrival rate through the quantum efficiency as

$$\bar{N}_{\text{electron}} = \frac{\eta P}{h\nu} \left( \frac{\text{electrons}}{s} \right) \quad (3.10)$$

The incident photons at the photodetector can be described by a Poisson arrival process with an arrival rate given in Eq. (3.9) \[32,34,43\]. Therefore, the electron emissions are governed by a similar arrival process with a rate equal to the electron arrival rate in Eq. (3.10). Since the arrival process is Poisson, the mean arrival rate is

57
equal to the variance of the arrival rate.

\[ \bar{N}_{\text{electron}} = \text{var}(N_{\text{electron}}) \quad (3.11) \]

The average photocurrent produced by the detector is directly proportional to the photon arrival rate through the relationship

\[ E[i_{\text{photo}}] = q\bar{N}_{\text{electron}} = \frac{q\eta P}{h\nu} = \mathcal{R} P \text{ (A)} \quad (3.12) \]

where \( \mathcal{R} = \frac{q\eta}{h\nu} \text{ (A/W)} \) is defined as the detector responsivity and \( q \) is the electron charge. So far the arrival processes have been assumed to be homogeneous Poisson processes such that the arrival rates are constant. In reality, the photon arrival rate is a function of time, and hence the arrival process is more aptly described by an inhomogeneous Poisson process \[41\]. A time-varying arrival rate will have the effect of modulating the incident power, adding sidebands to the received signal spectrum.

The total current is often decomposed into an average DC value plus a zero-mean term describing the noise. This is somewhat intuitive, as by definition noise is any random fluctuation of the signal around its mean value. Decomposing the total photocurrent into a DC component with zero variance and a zero-mean noise term yields

\[ i_{\text{photo}} = i_{\text{signal}} + i_{\text{noise}} \quad (3.13) \]

The expected value of this current is represented entirely by the signal current and is given by

\[ E[i_{\text{photo}}] = E[i_{\text{signal}}] = \frac{q\eta P}{h\nu} \text{ (A)} \quad (3.14) \]

since \( E[i_{\text{noise}}] = 0 \) by definition. The variance of the photocurrent is given as

\[ \sigma_{\text{photo}}^2 = \sigma_{\text{signal}}^2 + \sigma_{\text{noise}}^2 + 2\text{Cov}(i_{\text{signal}}, i_{\text{noise}}) \quad (3.15) \]
For $\sigma_{signal}^2 = 0$, Eq.(3.15) reduces to

$$\sigma_{noise}^2 = \sigma_{photo}^2 = \frac{\eta q^2 P}{h\nu} = qE[i_{photo}]$$

(3.16)

As the noise current is zero mean, the mean square value is equal to the variance

$$E[i_{noise}^2] = \sigma_{noise}^2$$

(3.17)

The signal and noise powers into a resistive load $R_L$ are given by $i_{signal}^2 R_L$ and $E[i_{noise}^2] R_L$, respectively. At this point it is worthwhile to reintroduce the Signal-to-Noise Ratio (SNR). There are a number of definitions for SNR in use depending on the application. However, the SNR will be defined here as

$$\text{SNR} = \frac{i_{signal,rms}^2 R_L}{E[i_{noise}^2] R_L} = \frac{i_{signal,rms}^2}{\sigma_{noise}^2}$$

(3.18)

Just as SNR is an important performance metric in communication systems, subsequent analysis will show that SNR is equally useful in determining the performance limits of the interferometer. How the discrete nature of photons and electrons manifest themselves as shot noise within the photodetector has already been described at a high level. But it would also be instructive to provide a description of the shot noise spectrum so that it can be compared to the Johnson noise in the frequency domain. The Wiener-Khinchin Theorem is a well-known theorem relating the autocorrelation of a process to its power spectral density by the Fourier transform pair

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} S_{xx}(f) e^{-j2\pi f\tau} df$$

(3.19)

$$S_{xx}(f) = \int_{-\infty}^{\infty} R_{xx}(\tau) e^{j2\pi f\tau} d\tau$$

(3.20)

In terms of the photocurrent, the autocorrelation $R_{xx}(\tau)$ can be written as

$$R_{xx}(\tau) = E(i(t) i^* (t - \tau)) = Cov(i(t), i^* (t - \tau)) + E(i(t)) E(i^* (t - \tau))$$

(3.21)
Assuming a stationary process such that the expected value of the current is time invariant, the autocorrelation can be rewritten as

\[ R_{xx}(\tau) = \text{Cov}(i(t), i(t-\tau)) + \bar{i}(t)^2 \] (3.22)

The assumptions regarding the stochastic nature of shot noise are similar to those made in the discussion of thermal noise. The covariance can be evaluated simply if the photocurrent is described in terms of a Poisson arrival process. It can be shown that Poisson processes obey the independent increment property, which states that two counting processes \( N(t_1) \) and \( N(t_2 = t_1 - \tau) \) are statistically independent for \( t_1 \neq t_2 \) [41]. Therefore, the covariance term of the autocorrelation can be written as

\[ \text{Cov}(i(t), i(t-\tau)) = 0; \text{ for } \tau \neq 0 \]
\[ q^2 \bar{N}; \text{ for } \tau = 0 \] (3.23)

Using this result for the covariance, the definition for the electron arrival rate in Eq.(3.10), and rewriting the expected photocurrent as \( \bar{I} \), the autocorrelation describing the shot noise is given by

\[ R_{xx}(\tau) = \frac{\eta q^2 P}{h\nu} \delta(\tau) + \left( \frac{\eta q P}{h\nu} \right)^2 = q\bar{I}\delta(\tau) + \bar{I}^2 \] (3.24)

The last step is to apply the Wiener-Khinchin Theorem to the autocorrelation, yielding the following expression for the shot noise power spectral density.

\[ S_{xx}(f) = \int_{-\infty}^{\infty} (q\bar{I}\delta(\tau) + \bar{I}^2)e^{j2\pi f\tau} d\tau = q\bar{I} + \bar{I}^2\delta(f) \] (3.25)

The resulting power spectral density is comprised of two components: a term that is constant across all frequencies, and a discrete term containing a Dirac delta function that exists only at zero Hz. The second term represents the spectral component due to the average current. But the most interesting term is the first, which represents
the double-sided power spectral density of the shot noise.

\[
S(f)_{\text{shot, double-sided}} = q\bar{I} \left(\frac{\text{W}}{\text{Hz}}\right)
\]  

(3.26)

As there is no frequency dependence in the power spectral density, the shot noise is white with the same power in any 1 Hz bandwidth. The power contained in a bandwidth of B Hz is given by

\[
P_{\text{shot noise}} = \int_{-B}^{B} q\bar{I} df = 2q\bar{I}B \left(\text{W}\right)
\]  

(3.27)

The shot noise power spectral density is a two-sided function with components at both negative and positive frequencies. But in accordance with convention, the single-sided shot noise power spectral density is defined as twice the double-sided density such that

\[
S(f)_{\text{shot, single-sided}} = 2q\bar{I} \left(\frac{\text{W}}{\text{Hz}}\right)
\]  

(3.28)

Using this definition, the shot noise power can be found by integrating only over positive frequencies. One important caveat is that shot noise is contributed not only by the incident and background optical signals, but also by the photodetector dark current. Dark current is present in photosensitive devices even with no incident optical power. Dark current is due to the generation of electrons and holes within the device’s depletion region, often via thermionic emission, that are then swept across the depletion region by electric fields. In general, dark current is present in any diode. Until this point the shot noise has been attributed solely to the discrete nature of photons striking the detector, which leads to the generation of charge carriers within the depletion region. But this is only part of the story. The charge carriers, both electrons and holes, are also quantized in discrete units. In this case the units are charge rather than photons, and in the presence of a potential barrier these charge carriers generate a non-continuous current that obeys the same Poisson statistics as the photon-generated charge carriers. Therefore, it is difficult to discern photon-
generated shot noise from that generated by discrete charge carriers in a potential barrier [34]. As a historical aside, shot noise was discovered by Walter Schottky during his work on vacuum tubes. So it was the noise generated by electrons crossing high potential barriers that led to the discovery of shot noise, which would later be applied to photons. The total shot noise spectral density due to the average DC component of the desired signal, background radiation, and dark current can be written compactly as

\[
S(f)_{\text{shot, single-sided}} = 2q \left( \bar{I} + \bar{I}_{BG} + \bar{I}_{Dark} \right) \left( \frac{\text{W}}{\text{Hz}} \right)
\]

In practice, shot noise due to dark current and background radiation are negligible compared with that of the signal and LO, and will be ignored going forward. Hence, the dominant noise sources are signal shot noise, receiver Johnson noise, and laser intensity noise. There is one last point that should be made before continuing. When discussing thermal noise, several statistical properties were attributed to it. One of those assumptions was that the noise probability distribution could be described by a zero-mean Gaussian function. This has been shown to hold not only for the thermal noise, but for shot noise as well. The position was also put forth that the shot noise process, under most circumstances, is well-described by a Poisson process. By employing a key result from statistics, the Central Limit Theorem, the probability distribution is shown to converge to a Gaussian in the limit of a large number of individual observations. The probability density function (PDF) for a zero-mean Gaussian distribution is given by the equation

\[
f_X (X = x) = \frac{1}{\sqrt{2\pi\sigma^2_x}} e^{-\frac{x^2}{2\sigma^2_x}}
\]

This expression provides a means to determine the probability that the noise, denoted by the random variable (RV) \( X \), will assume a particular value \( x \). Technically, the probability that a continuous RV will assume a specific value \( X = x \) is zero, as there are an uncountably infinite number of possible values of the RV. The correct way to quantify the probability of a continuous RV at a specific value is to state the probability that the RV will lie within an infinitesimally small interval \( \delta x \) about the
value $x$. Assuming the PDF is constant in the infinitesimally small region about $\delta x$, the probability $X = x$ can be described as

$$P(x - \delta x \leq x \leq x + \delta x) = \lim_{\delta x \to 0} \int_{x-\delta x}^{x+\delta x} f(x') \, dx' \approx 2f(x) \, \delta x \quad (3.31)$$

The constant two can be dropped as it becomes superfluous in the limit $\delta x$ approaches zero. Given that $f(x)$ is defined as the probability density over a continuous random variable, $f(x)$ can be interpreted as the probability $P(X = x)$. To reiterate, a true white noise process is a mathematical idealization. Applying the Wiener-Khinchin Theorem, a true white noise process is seen to possess infinite variance, and hence infinite signal energy.

$$\sigma^2 = R_{xx}(0) = \int_{-\infty}^{\infty} S_{xx}(f) \, df = \infty \quad (3.32)$$

A requirement of Gaussian processes is that the covariance matrix consist of finite values [40,41]. This is an artifact of the independent increment property of Poisson counting processes, which led to an infinite delta function in the autocorrelation function of Eq. (3.24). Of course neither assumption holds in a physical system. This discrepancy can be reconciled by understanding that the autocorrelation is not a true delta function, but looks instead like a tall, narrow rectangular function. This form of the autocorrelation represents a process that is perfectly correlated for a narrow duration about the origin, and then quickly falls to zero. The frequency response of such a correlation function would be a \text{sinc}(\sin(x)/x) function with a width inversely proportional to the width of the correlation function. However, in any physical system, the frequency response is limited to a finite bandwidth imposed at various points throughout the system, such as the photodetector response, external filters, amplifiers, etc. Because Gaussian processes filtered through Linear Time-Invariant (LTI) systems retain their Gaussian nature [41], both the filtered shot noise and thermal noise contributions remain Gaussian, but band-limited by the system.
frequency response.

### 3.1.2 Phase Noise Sources

This section briefly describes several sources of phase noise that alter the phase directly and lead to jitter in the time domain. The effect of the laser source, which is typically the primary source of phase noise in an interferometric system, were mentioned previously. But some of the more important points will be repeated in this section. Laser noise generally consists of two components: an amplitude or intensity noise that can be modeled as an in-phase amplitude modulation, and a phase noise term that introduces a time-dependent phase modulation such that the instantaneous phase becomes

\[
\Phi(t) = \omega_0 t + \phi(t)
\]

where \(\omega_0\) is the angular frequency of the laser and \(\phi(t)\) is the stochastic, time-varying phase noise. The optical noise in a laser is primarily attributed to quantum mechanical fluctuations within the laser gain medium and vacuum fluctuations due to optical losses. However, variations induced by mechanical and/or thermal perturbation within the laser cavity may also generate amplitude and phase noise. Taking the derivative of the instantaneous phase, the instantaneous frequency is recovered and given as

\[
\Omega(t) = \frac{d\Phi(t)}{dt} = \omega_0 + \frac{d\phi(t)}{dt}
\]

The instantaneous frequency of a single frequency laser consists of \(\omega_0\), representing the desired angular frequency, and a term owing to the time-varying phase noise. It is this term that causes the spectrum of a single frequency laser to deviate from the pure spectral component at \(\omega_0\) and to acquire a finite linewidth. This spectral broadening is the mechanism responsible for the temporal de-coherence that occurs between the laser and a replica of itself delayed by a distance known as the coherence length, which is inversely proportional to the linewidth and dependent on the specific line shape. For example, the Agilent 81600B tunable laser used in these experiments has a linewidth of approximately 100 kHz, which corresponds to a coherence length on
the order of 3 km at 1.55 μm. Because the interferometer reference and measurement signals are derived from a single source, the phase noises of the two signals are highly correlated. So the use of balanced detection should be highly effective at canceling this noise.

Laser phase noise is perhaps the greatest contributor to the phase noise budget, but any phenomenon that causes a relative phase shift between the reference and measurement arms will manifest itself as phase noise. Two particularly important sources of phase noise are atmospheric turbulence and structural vibration. Atmospheric turbulence is caused by inhomogeneities in the refractive index of air, and is a separate field of study with an extensive literature. Refractive index fluctuations are caused by non-uniformities in atmospheric temperature, pressure, and humidity. These localized atmospheric gradients are generated primarily by wind shear and convective heating, with wind shear being the most dominant. Wind shear creates localized circulating air currents, or eddies, that have their own values of velocity, vorticity, pressure, and temperature. The resulting refractive index fluctuations across the optical path adversely affect interferometer performance because of statistical variations in the optical path length as the signal travels to and from the target.

Therefore, it is important to reduce the measurement path length to the minimum necessary and to shield the free space path to the extent possible. In addition to phase noise due to random fluctuations in the optical path length, turbulence also causes very deep intensity fading. A useful analysis of turbulence-induced errors in interferometers can be found in [9]. Noise caused by structural and equipment vibrations are also a key component of the phase noise budget. These vibrations can be caused by building sway, air conditioning and heating units, and automobile traffic, to name a few. These vibrations can have frequencies ranging from a few tens of hertz to several kilohertz, so they can introduce close-in phase noise that is difficult to filter.
3.1.3 Balanced Detection Signal-to-Noise Ratio

The principal of balanced detection was first introduced in Chapter 2, where the ability of this scheme to suppress certain forms of common-mode noise and to recapture the 3-dB of power normally lost in the coupler were put forth as beneficial to interferometer performance. Balanced detection will be revisited here, but now with the machinery of noise statistics available. The detector shot noise and thermal noise originate from independent processes, and hence are statistically independent. Furthermore, the shot noise of one detector is known to be statistically independent of the shot noise in a different detector. Similarly, the thermal noise of a resistive element with temperature $T$ is independent of the thermal noise of a separate resistive element at the same temperature. However, laser RIN noise is correlated between detectors and can be suppressed. The sum and difference detector currents can be written to include their respective noise terms.

\[
I_{\Sigma} = I_{\Sigma,DC} + I_{\Sigma,IF} + I_{\Sigma,Shot} + I_{\Sigma,Thermal} + I_{\Sigma,Corr} \quad (3.35)
\]

\[
I_{\Delta} = I_{\Delta,DC} - I_{\Delta,IF} + I_{\Delta,Shot} + I_{\Delta,Thermal} + I_{\Delta,Corr} \quad (3.36)
\]

Recall that the IF currents were defined as

\[
I_{\Sigma,IF} = R_{\Sigma} \sqrt{P_1 P_2} \cos (\omega_{IF} t + \delta \phi) ; \quad I_{\Delta,IF} = R_{\Delta} \sqrt{P_1 P_2} \cos (\omega_{IF} t + \delta \phi) \quad (3.37)
\]

The balanced current, taken as the difference between the sum and difference current, can be written as

\[
I_B = (I_{\Sigma,DC} - I_{\Delta,DC}) + (R_{\Delta} + R_{\Sigma}) \sqrt{P_1 P_2} \cos (\omega_{IF} t + \delta \phi)
\]

\[
+ (I_{\Sigma,Shot} - I_{\Delta,Shot}) + (I_{\Sigma,Thermal} - I_{\Delta,Thermal}) + (I_{\Sigma,Corr} - I_{\Delta,Corr}) \quad (3.38)
\]

While the expression for the balanced current is correct, in its current instantiation it offers little insight into the performance of a heterodyne interferometer. A better metric would be to examine the SNR. An expression for the SNR provided in the
previous section was

\[ \text{SNR} = \frac{i_{\text{signal, rms}}^2 R_L}{E[i_{\text{noise}}^2]} = \frac{i_{\text{signal, rms}}^2}{\sigma_{\text{noise}}^2} \] (3.39)

An equally valid expression can be written in terms of the RMS voltage and the voltage noise, but since current is the quantity supplied by the photodetection process, the analysis will use the form given in Eq.(3.39). Because noise power is equivalent to the variance, a well-known result from statistics can be applied which states that for independent random variables, the total variance is the sum of the individual variances, as all covariance terms equal zero. So the total noise for a particular photodetector, if dark current and background noise are ignored, is the sum of the signal shot noise, thermal noise, and correlated noise sources.

\[ \sigma_{\text{noise, total}}^2 = \sigma_{\text{shot}}^2 + \sigma_{\text{thermal}}^2 + \sigma_{\text{correlated}}^2 \] (3.40)

In a balanced detector, the three noise terms in Eq.(3.40) are comprised of contributions from both the sum and difference detector. This is expected from the noise terms of the balanced detector output current in Eq.(3.38). It is useful to apply another result from statistics relating the variance of the difference of two random variables A and B such that

\[ \text{Var} (A - B) = \text{Var} (A) + \text{Var} (B) - 2 \text{Cov} (A, B) \] (3.41)

where Cov(A, B) is the covariance of the random variables and is zero for independent random variables. This relationship enables the individual noise variances to be written in a straightforward manner, assuming independence of the shot noise and thermal noise processes of the sum and difference detectors.

\[ \sigma_{\text{shot}}^2 = \text{Var} (I_{\Sigma, \text{Shot}} - I_{\Delta, \text{Shot}}) = \sigma_{\Sigma, \text{shot}}^2 + \sigma_{\Delta, \text{shot}}^2 \] (3.42)

\[ \sigma_{\text{thermal}}^2 = \text{Var} (I_{\Sigma, \text{thermal}} - I_{\Delta, \text{thermal}}) = \sigma_{\Sigma, \text{thermal}}^2 + \sigma_{\Delta, \text{thermal}}^2 \] (3.43)
The noise powers will add for independent shot noise and thermal noise processes in the detectors. Assuming the shot noise and thermal noise in both balanced detectors is equal, the process of balanced detection results in twice the noise of a single detector configuration. But because signal power is also doubled, the SNR remains the same with respect to the single detector configuration. A unity gain in SNR is certainly not worth the added complexity of a balanced detector configuration versus a single detector configuration. As discussed previously, the primary benefit of balanced detection is in its rejection of common-mode terms such as correlated noise and DC offsets. The correlated noise powers between the two detectors can be written as

\[ \sigma_{corr}^2 = \sigma_{\Sigma,corr}^2 + \sigma_{\Delta,corr}^2 - 2Cov (I_{\Sigma,corr}, I_{\Delta,corr}) \] (3.44)

Depending on the degree of correlation between the two detectors, some cancellation of the correlated noise terms can be achieved, hence providing a SNR gain with respect to a single detector. In fact, if the variance of each detector is equal and the noise in both detectors is completely correlated, the term in Eq. (3.44) completely vanishes. For the sake of simplicity, assume this is the case and consider only shot and thermal noise contributions. With these results, an expression for the SNR is within reach. The benefits of operating in the shot-noise-limited regime such that \( \sigma_{\text{shot}}^2 \gg \sigma_{\text{thermal}}^2 \) were explained earlier. The shot-noise-limited regime represents the limit of performance and is typically achievable for relatively small levels of incident power at the detector. However, given the large losses possible while coupling into and out of an integrated chip, the system may be power limited to below the level necessary to ensure shot-noise-limited operation. In any case it is necessary to minimize the thermal noise added by the electronics. Employing the expressions for the RMS signal and noise currents for balanced detection, the SNR is given by

\[ \text{SNR} = \frac{i_{\text{signal, rms}}^2}{\sigma_{\text{noise}}^2} = \frac{(R_{\Delta} + R_{\Sigma})^2 P_1 P_2}{(2qI_{\Delta} + 2qI_{\Sigma} + 4kT/R_{\Delta,\text{EQ}} + 4kT/R_{\Sigma,\text{EQ}})2B} \] (3.45)

where the first two terms in the denominator represent the shot noise of the sum and difference detector, the second two terms represent the thermal noise each detector
sees due to an equivalent load resistance, and B is the system bandwidth that filters
the white spectral noise. Assuming equal responsivities $\mathcal{R}$ at each detector, equivalent
noise resistances $R_{EQ}$, and equal average currents in both detectors, the SNR reduces to

$$\text{SNR} = \frac{2\mathcal{R}^2 P_1 P_2}{2q\mathcal{R}(P_1 + P_2) + 4kT/(R_{EQ}/2)} B \quad (3.46)$$

The expression in Eq. (3.46) is the general form of the SNR when both shot noise and
thermal noise dominate over other noise sources. In the case where the shot noise
dominate the thermal noise, the shot-noise-limited SNR may be written as

$$\text{SNR}_{\text{Shot, Noise}} = \frac{\mathcal{R}P_1 P_2}{q(P_1 + P_2) B} \quad (3.47)$$

This is indeed the desired operational regime, but achieving it depends on the ability
to couple sufficient power into the device without exciting nonlinearities in the silicon
waveguides or saturating the photodetectors. The next section will demonstrate how
the SNR influences the phase measurement. The powers $P_1$ and $P_2$ into the reference
detectors come directly from the single-sideband modulators and are fairly strong with
approximately equal power, assuming 50:50 couplers and equivalent losses through
the modulators. In this case, the shot-noise-limited SNR at the reference detector
can be written, assuming $P_1 = P_2 = P_0$, as

$$\text{SNR}_{\text{Ref, Shot, Noise}} = \frac{\mathcal{R}P_0}{2qB} \quad (3.48)$$

However, the SNR at the measurement detectors more closely matches the results
seen in the traditional analysis of optical heterodyne systems. In this case there
is a strong local oscillator with power $P_{LO}$ and a weaker signal $P_S$ that carries the
information to be demodulated. In the limit $P_{LO} \gg P_S$, which is usually the case,
the shot-noise-limited SNR at the measurement detectors can be approximated by

$$\text{SNR}_{\text{Meas, Shot, Noise}} = \frac{\mathcal{R}P_{LO}P_S}{q(P_{LO} + P_S)B} \approx \frac{\mathcal{R}P_{LO}P_S}{qP_{LO}B} = \frac{\mathcal{R}P_S}{qB} \quad (3.49)$$
Applying the definition of detector responsivity, the measurement arm SNR can be simplified to

\[
\text{SNR}_{\text{Meas,ShotNoise}} = \frac{\eta P_s}{h\nu B}
\]  

(3.50)

It can be seen that the shot-noise-limited SNR in the measurement arm with a weak signal and strong LO depends on the detector efficiency, system bandwidth, optical frequency, and the signal power. Lastly, since expressions are available for the detector shot noise current density and an equivalent thermal current noise density, the crossover point between electronic and shot-noise-limited operation can be found. Because the detector shot noise represents a current density, using the Norton equivalent model of a current noise source in parallel with a noiseless resistor is appropriate to describe the thermal noise, and hence provides an equal basis of comparison. The transimpedance amplifier (TIA) is a circuit that converts photocurrents into voltages. While TIAs exist in many forms, one particularly simple configuration that is frequently used is the operational amplifier with shunt feedback. A simplified TIA using this configuration is shown in Figure 3-1. Here, \(i_{sn}\) is the shot noise current

\[
\begin{align*}
\text{Figure 3-1: A conceptual diagram of the fabricated TIA circuit.}
\end{align*}
\]

generated by the photodetector and \(i_{th}\) is the equivalent thermal noise current generated by the noiseless feedback resistor \(R_{fb}\). The output noise voltage is equal to

\[
V_{o,N} = -(i_{sn} + i_{th})R_{fb}
\]

when the op-amp is operated in the inverting configuration shown in Figure 3-1. Equating the shot noise current density and the resistor thermal
noise current density yields

\[
\sqrt{2q \mathcal{R} P_S} = \sqrt{\frac{4kT}{R_{fb}} \left( \frac{A}{\sqrt{\text{Hz}}} \right)} \tag{3.51}
\]

The minimum incident power \( P_S \) on the photodetector such that the shot noise equals the thermal noise is therefore

\[
P_{S,\text{Min}} = \frac{2kT}{RqR_{fb}} \tag{3.52}
\]

That only shot noise and thermal noise are included is a convenience to simplify the analysis. In reality, other current noise sources exist at the op-amp input that will generate voltage noise at the output. A practical op-amp has an input noise current density generated by its bias currents. These currents, which are necessary to bias the internal transistors, generate their own shot noise that manifests itself as an input noise source. Depending on the internal structure of the op-amp (BJT, JFET, etc.), the value of the feedback resistor, and the operating frequency, the op-amp’s input noise current density can far exceed the thermal current of the feedback resistor. If this is the case, the op-amp input noise must be considered the dominate source of electrical noise. But this is rarely ever the case. This concludes the discussion of noise sources and the balanced detection SNR. In the next section these results are applied to the analysis of phase performance.

### 3.2 The Effect of Noise on Phase Measurement

In interferometry absolute accuracy is not necessary, but stability, linearity, and resolution are of the highest importance. The angular resolution obtainable with a phase detector depends on the signal frequency and the timing resolution \[44\], as many detectors directly determine phase by measuring the timing interval between successive zero crossings of the reference and measurement waveforms. The motion represented by this phase shift can either be displacement or velocity. This follows from the fact
that the measurement signal in a translational interferometer is phase modulated by the displacement, and at the same time, frequency modulated by the velocity if the object is in motion \[45\]. The instantaneous values of these quantities are obtained by counting the signal periods (preferably the positive zero crossings) and measuring the time intervals between them.

In phase modulated systems, multiplicative noise will not affect the zero crossings, and hence should not have an impact on the phase measurement. However, additive noise sources are of concern. To see how additive noise translates into phase errors, it is instructive to consider the timing jitter introduced into a periodic signal. Phase noise produced by signal generators and laser sources is usually the primary contributor to timing jitter. But several assumptions are put forth, which if they hold true, should mitigate the impact of the laser and signal generators on timing jitter. First, assume the laser source has low phase noise, which is reasonable given the narrow linewidth of the lasers being employed. But perhaps most importantly, both the reference and measurement signals are derived from the same laser source, causing their phase noises to be highly correlated. From the discussion of noise in balanced detection systems, assume that phase noise introduced by the source can be ignored if the delay between the reference and measurement arms is less than the coherence length. Phase noise from the RF signal generators used to modulate the optical carriers can also be correlated to a high degree by deriving them from the same generator. But even if separate signal generators are used to drive each SSBM, signals from both SSBMs are present in both detector pairs.

While it may seem counter-intuitive, amplitude noise imposed on the signal can have an impact on the timing jitter performance of a heterodyne interferometer. To see why this is the case, consider the sinusoid shown in Figure 3-2. The subsequent analysis follows those found in \[31\],\[46\]. If amplitude noise is superimposed on this sinusoid, noise near the zero crossing points may cause a shift in the zero-crossing position. Assume the signal of interest is a sinusoid with amplitude \(A_0\) much greater than the RMS noise amplitude \(\sigma_N\), such that the additive noise causes a timing shift only in the vicinity of the zero-crossing points. The timing shifts are related to the
Figure 3-2: A simple sinusoid for examining amplitude noise contributions to phase noise.

amplitude shift through the signal slope. The voltage signal is written as

\[ s(t) = A_0 \sin(\omega t) \quad (3.53) \]

with a time derivative given by

\[ \frac{ds(t)}{dt} = \omega A_0 \cos(\omega t) \quad (3.54) \]

If a further assumption is made that the derivative is approximately constant in a neighborhood about the zeroes of \( s(t) \), then the derivative can be approximated as a difference equation such that

\[ \left| \frac{ds(t)}{dt} \right| = \frac{\Delta s}{\Delta t} \approx \omega A_0 \quad (3.55) \]

Denoting the statistical change of the signal \( \Delta s \) by the RMS noise \( \sigma_N \), the timing shift due to noise can be approximately written as

\[ \Delta t = \frac{\sigma_N}{\omega A_0} \quad (3.56) \]

From this it follows that the noise induced phase shift relative to the signal period
\[ T = \frac{2\pi}{\omega} \] can be written as

\[ \Delta \phi_{\text{Noise}} = \frac{\Delta t}{T} = \frac{\sigma_N}{2\pi A_0} \quad (3.57) \]

Recognizing that the SNR can be represented in terms of the noise variance and peak amplitude by

\[ \text{SNR} = \frac{A_0^2}{2\sigma_N^2} \quad (3.58) \]

the phase noise induced by amplitude fluctuations can also be written in terms of the SNR as

\[ \Delta \phi_{\text{Noise}} = \frac{1}{2\pi \sqrt{2\text{SNR}}} \quad (3.59) \]

This relationship between the SNR and the minimum displacement resolution is quite pleasing. Because the phase noise is directly proportional to the amplitude standard deviation \( \sigma_N \), it is possible to relate the phase noise statistics to those of the amplitude fluctuations. The expression derived for the phase noise is not a mean value, since the mean value is zero, but rather the standard deviation. This is because the total noise, which was assumed to consist of shot and thermal noises with Gaussian distributions, is also zero mean. Since the sum of a number of Gaussian random variables is also Gaussian, with a variance equal to the sum of the individual variances, the phase noise also has a Gaussian distribution with variance given by

\[ \sigma_\phi^2 = \frac{\sigma_N^2}{(2\pi A_0)^2} \quad (3.60) \]

The required SNR to achieve a given phase resolution is therefore equal to

\[ \text{SNR}_{\text{Min}} = \frac{1}{\sqrt{8\pi \Delta \phi_{\text{Noise}}} \quad (3.61) \]

And while not critical, the phase resolution can be used to determine the required wavelength discretization. The fractional wavelength resolution \( N \), first discussed in
Section 2.3.1 can be written in terms of the phase noise as

\[
\frac{\Delta t}{T} = \frac{\lambda/N}{\lambda} = \frac{1}{N} = \Delta \phi_{\text{Noise}}
\]  

(3.62)

This allows the fractional wavelength resolution \( N \) to be expressed in term of SNR as

\[
N = \frac{1}{\Delta \phi_{\text{Noise}}} = 2\pi \sqrt{2\text{SNR}}
\]  

(3.63)

Thus the achievable displacement resolution for a two-pass interferometer per unit of achievable phase resolution is

\[
\Delta d = \frac{\lambda/N}{4\pi \left( \frac{m}{\text{radian}} \right)}
\]  

(3.64)

But while the timing uncertainty due to amplitude noise has already been examined, one could ask whether the choice of intermediate frequency has an impact on the achievable phase resolution. By re-examining Eq.(3.57), the fractional phase error is seen to not only depend on the timing uncertainty, but also on the electrical period of the reference and measurement signals.

\[
\frac{\Delta t}{T} = \Delta \phi
\]  

(3.65)

This equation implies that the phase accuracy is also inherently tied to the IF frequency through its period. Rewriting the above relationship in terms of the intermediate frequency, the minimum timing accuracy required to resolve a given phase shift is

\[
\Delta t = \frac{\Delta \phi}{f_{IF}}
\]  

(3.66)

This states that for a desired phase resolution \( \Delta \phi \), the required timing accuracy actually decreases (improves) as frequency decreases. Intuitively this makes sense in that for longer wavelengths (lower frequencies), the \( 2\pi \) phase of the sinusoid is spread out over a longer period. Therefore, the required timing resolution of the measurement apparatus required to achieve a particular phase resolution can be reduced by using a
lower intermediate frequency. Or better yet, for a given temporal resolution, a higher phase resolution can be achieved at lower frequencies. But it was previously demonstrated that if Doppler frequency shifts are present, the intermediate frequency must be large enough to ensure that aliasing does not occur when operating at the maximum expected Doppler shift. So there is an inherent tension between the attainable displacement resolution and the maximum Doppler frequency of either sign that can be accommodated with the same demodulator.

A final point is made with respect to the choice of the optical wavelength. When discussing the phase ambiguity and synthetic wavelengths, it was stated that single wavelength measurements allow increased phase resolution versus longer synthetic wavelengths. In general, this behavior is expected. From radar to microscopy, the minimum resolvable feature size is inversely proportional to frequency. In the case of diffraction limited optics, optical frequencies on the order of the minimum dimension size are required. This illustrates that more accurate measurements are possible utilizing a finely-ruled standard (shorter wavelength) than with a coarser standard (longer wavelength). So while using a higher optical frequency reduces the maximum displacement that can be measured without ambiguity, it also provides a higher phase resolution.

But Eq. (3.66) showed that lower intermediate frequencies are better for measuring phase with higher resolution. And while there appears to be a contradiction, there is not. The phase imparted to the optical signal is determined by both the object displacement and the optical frequency \( \nu \). In converting the optical signal to a RF signal via the process of heterodyning, the optical phase shift is preserved in the intermediate frequency. Once in the electrical domain, a lower IF is actually preferable for a phase meter with fixed timing resolution. So for optimal results, the highest optical frequency should be chosen to provide the minimum acceptable ambiguity, and the lowest IF should be chosen to accommodate the expected Doppler shift without aliasing. The next section will consider the quadrature demodulator, which is utilized in many phase measurement instruments such as the lock-in amplifier due to its simplicity, accuracy, and resistance to amplitude modulation and noise.
3.3 The Coherent Demodulator

A widely used demodulation scheme for extracting phase in a heterodyne system is the coherent (IQ) demodulator. The coherent demodulator provides complex phasor data, allowing the simultaneous measurement of phase and amplitude \[18\]. This technique is also widely used in radar and communications systems so the details are well established, but are briefly described here \[47,48\]. The basic premise of IQ demodulation is shown above in Figure 3-3. The measurement signal is compared to a reference signal of the same intermediate frequency. Both signals are digitized either by a dedicated analog-to-digital (A/D) converter or an oscilloscope, and each is further split into two identical copies. One copy of the reference signal is shifted by 90° using a Hilbert transform, converting a cosine into a sine or vice-versa. The reference signal cosine and sine images each multiply a copy of the measurement signal. This produces in-phase (I) and quadrature (Q) components at baseband and 2\(f_{IF}\). The high frequency terms are removed using a FIR low-pass filter and the phase is recovered by

\[
\phi = \arctan \frac{Q}{I}
\]  

(3.67)

To illustrate a virtue of this method, the phase is computed using noisy values for the measurement and reference signal, with the noisy measurement and reference signals

Figure 3-3: A conceptual drawing of the IQ demodulator for the coherent measurement of phase.
written as

\[ s_M(t) = [A_M + a_M(t)] \sin(2\pi ft + \phi_M(t)) + n_M(t) \]  \hspace{1cm} (3.68)
\[ s_R(t) = [A_R + a_R(t)] \sin(2\pi ft + \phi_R(t)) + n_R(t) \]  \hspace{1cm} (3.69)

The reference signal is copied and shifted 90° to generate a quadrature image such that

\[ s_R, I(t) = [A_R + a_R(t)] \sin(2\pi ft + \phi_R(t)) + n_R(t) \]  \hspace{1cm} (3.70)
\[ s_R, Q(t) = [A_R + a_R(t)] \cos(2\pi ft + \phi_R(t)) + n_R(t) \]  \hspace{1cm} (3.71)

Ignoring the AM noise components, the I and Q components are computed as

\[ I = s_{R,I}(t) \times s_M(t) = A_M A_R \sin(2\pi ft + \phi_R(t)) \sin(2\pi ft + \phi_M(t)) + n_R(t) A_M \sin(2\pi ft + \phi_M(t)) + n_M(t) A_R \sin(2\pi ft + \phi_R(t)) + n_M(t) n_R(t) \]  \hspace{1cm} (3.72)
\[ Q = s_{R,Q}(t) \times s_M(t) = A_M A_R \cos(2\pi ft + \phi_R(t)) \sin(2\pi ft + \phi_M(t)) + n_R(t) A_M \sin(2\pi ft + \phi_M(t)) + n_M(t) A_R \cos(2\pi ft + \phi_R(t)) + n_M(t) n_R(t) \]  \hspace{1cm} (3.73)

Applying the product-to-sum trigonometric identities to the above equations and low-pass filtering the terms at the fundamental and twice the fundamental, the in-phase and quadrature signal components are

\[ I = A_M A_R \cos(\phi_M(t) - \phi_R(t)) + n_M(t) n_R(t) \]  \hspace{1cm} (3.74)
\[ Q = A_M A_R \sin(\phi_M(t) - \phi_R(t)) + n_M(t) n_R(t) \]  \hspace{1cm} (3.75)

Applying Eq.(3.67) and assuming the product of the additive noise terms are negligible, the phase is given as

\[ \phi = \arctan \frac{\sin(\phi_M(t) - \phi_R(t))}{\cos(\phi_M(t) - \phi_R(t))} = \phi_M(t) - \phi_R(t) \]  \hspace{1cm} (3.76)
Because the product of the reference and measurement amplitude noise terms are present in both the I and Q components, the amplitude noise cancels and the phase noise is limited only by phase noise sources. The resulting phase is the difference between the noisy signal phases, hence the mean phase is the difference in the mean values of the signal phases and the phase variance is

\[
\sigma^2_\phi = \sigma^2_\phi_M + \sigma^2_\phi_R - 2 \text{Cov}(\phi_M, \phi_R)
\]  

(3.77)

Therefore, the resolution is limited by the phase noise, which for the IQ demodulator is determined by the quadrature phase noise sources and not the amplitude noise. In reality, the Hilbert transform modifies the spectrum of the noise itself, so cancellation in the ratios is not perfect. This residual additive noise adds to the quadrature phase noise. However, if there is high correlation between the measurement and reference phases, the total phase variance will be reduced, leading to lower total phase and position noise. Another benefit of the IQ demodulator is that unlike zero-crossing detection, where only signals near the origin are considered, the IQ demodulator can use any portion of the signal to determine the phase. Another ambiguity that exists in interferometric systems is resolving the direction of motion. These ambiguities are difficult to determine with non-coherent forms of demodulation, but are straightforward using coherent demodulation. The direction ambiguity of the phase and Doppler frequency shift, and how the IQ demodulator allows them to be resolved, is discussed in the following section.

### 3.3.1 Resolution of Doppler and Phase Sign Ambiguity

A difficulty that arises in the measurement of phase and Doppler frequency is an additional ambiguity related to the object’s direction of motion. For example, the expression for the Doppler shift in Eq. (2.44) has a sign that depends on whether the target is approaching (positive sign representing an increase in frequency) or receding (negative sign representing a decrease in frequency). In a heterodyne system, where the IF frequency is non-zero, it is easy to measure whether the Doppler shift
has caused a frequency increase or decrease with respect to the known IF. For a homodyne system operating near baseband, the even symmetry of the cosine creates an ambiguity, since \( \cos(-f_d) = \cos(f_d) \) and both conditions give identical spectra. A convenient way to resolve this ambiguity is to utilize the coherent (IQ) demodulator scheme described above. The IQ demodulator also provides the direction of the phase shift, allowing unambiguous determination of the direction of motion even when no Doppler shift is present. For Doppler measurements, the odd symmetry of the sine term allows the sign of the Doppler shift to be resolved, since \( \sin(-f_d) = -\sin(f_d) \).

By taking the arctangent of the quadrature to in-phase ratio, both the magnitude and direction of the phase shift are readily determined.

### 3.4 Conclusion

This chapter has explored the primary sources of noise in the heterodyne interferometer and demonstrated how they influence performance. The merits of zero-crossing detection versus coherent demodulation in mitigating these noise sources and improving the overall phase resolution were also discussed. The next chapter will describe the actual interferometer in detail and present results from nanometer-scale displacement measurements.
Chapter 4

The Integrated Heterodyne Interferometer

4.1 Background

In this chapter, the focus shifts from the theoretical background necessary to understand the integrated heterodyne interferometer to a detailed functional description. In particular, the design, operation, and experimental results of the interferometer in the displacement measurement mode are described in detail. These results were initially reported in [49], and only that theory which is necessary for the analysis of the integrated heterodyne interferometer will be repeated. The integrated heterodyne interferometer is shown conceptually in Figure 4-1. The device as currently designed is approximately 1 mm by 6 mm in size and is constructed of a series of on-chip beamsplitters, single-sideband modulators (SSBM), and germanium (Ge) detectors in a Michelson-like configuration. Because the heterodyne tones are confined to separate silicon waveguides, bulk optics are not required to maintain spatial separation; allowing the size to be significantly reduced. The device achieves a noise-limited position resolution on the order of 2 nm, defined as the Root Mean Square (RMS) deviation of the measured position from the commanded position. A laser with angular frequency $\omega_0$ is split between two SSBMs, which modulate their respective outputs to $\omega_1$ and $\omega_2$ (3a and 3b in Figure 4-1). The output of each modulator is further divided by
a 50:50 splitter with half the power from each modulator incident on a pair of Ge detectors operated in a balanced configuration. The splitters are adiabatic with a sum-difference ($180^\circ$-hybrid) transfer function given by

$$T_{\Sigma,\Delta} = \begin{bmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
-\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}
\end{bmatrix}$$

Therefore, one output gives the sum of the incident fields while the other gives the difference, which is well-suited to balanced detection. Descriptions of the adiabatic couplers and the adiabatic theorem are given in Appendix E. The first detector pair, shown at 3c in Figure 4-1, is designated as the reference pair. The signal produced by the reference pair is used to calibrate the input phase of the interferometer, enabling spurious drifts to be removed. If the detector coincident with the splitter sum port is designated as the sum detector and the other the difference detector, the respective
electric fields can be written as

\[ E_\Sigma = \frac{1}{\sqrt{2}} (E_{\text{Upper}} + E_{\text{Lower}}) \]  
(4.2)

\[ E_\Delta = \frac{1}{\sqrt{2}} (-E_{\text{Upper}} + E_{\text{Lower}}) \]  
(4.3)

Expressing the fields from the modulators (3a and 3b) into the reference path as

\[ E_{\text{Upper}} = E_1 e^{-j(\omega_1 t + \phi_1)} \]  
(4.4)

\[ E_{\text{Lower}} = E_2 e^{-j(\omega_2 t + \phi_2)} \]  
(4.5)

the electric fields and intensities at the reference detector sum and difference ports can be written as

\[ E_\Sigma = \left[ \frac{E_2}{\sqrt{2}} e^{-j(\omega_2 t + \phi_2)} + \frac{E_1}{\sqrt{2}} e^{-j(\omega_1 t + \phi_1)} \right] \]  
(4.6)

\[ E_\Delta = \left[ \frac{E_2}{\sqrt{2}} e^{-j(\omega_2 t + \phi_2)} - \frac{E_1}{\sqrt{2}} e^{-j(\omega_1 t + \phi_1)} \right] \]  
(4.7)

\[ I_\Sigma \propto \frac{|E_1|^2}{2} + \frac{|E_2|^2}{2} + E_1 E_2 \cos \left( (\omega_2 - \omega_1) t + (\phi_2 - \phi_1) \right) \]  
(4.8)

\[ I_\Delta \propto \frac{|E_1|^2}{2} + \frac{|E_2|^2}{2} - E_1 E_2 \cos \left( (\omega_2 - \omega_1) t + (\phi_2 - \phi_1) \right) \]  
(4.9)

Assuming an equal responsivity \( R \) for each detector, the balanced photocurrent from the reference pair is given by

\[ i_{\text{Reference}} = i_\Sigma - i_\Delta \propto 2R \sqrt{I_1 I_2} \cos (\omega_{IF} t + \delta \phi) \]  
(4.10)

where the frequency difference \( \omega_2 - \omega_1 \) in Eq.(4.10) is written as \( \omega_{IF} \), the intermediate frequency (IF) formed from the heterodyne mixing. The initial phase difference \( \phi_2 - \phi_1 \) in the reference pair is written \( \delta \phi \), and the electric field amplitudes \( E_1 \) and \( E_2 \) are given in terms of the square roots of their respective intensities \( I_1 \) and \( I_2 \). The
remaining signal from the upper modulator is directed into what is the functional equivalent of a bulk interferometer’s reference arm (4a). The remaining signal from the lower modulator (4b) is edge-coupled off-chip, collimated, and reflected from a corner cube reflector (CCR) mounted to a nano-positioner stage. The reflected measurement signal is coupled back into a separate waveguide where it combines with the reference arm signal. Assume the incident fields at (4c) are given by

\[
E_{\text{Reference}} = E_R e^{-j(\omega_1 t + \phi_R)} \\
E_{\text{Measurement}} = E_M e^{-j(\omega_1 t + \phi_M)}
\] (4.11)  (4.12)

These form a photocurrent in the measurement detectors given by

\[
i_{\text{Measurement}} \propto 2\mathcal{R}\sqrt{I_R I_M} \cos (\omega_{\text{IF}} t + \phi_R - \phi_M)
\] (4.13)

If the position of the CCR changes by a one-way distance \(\Delta d\), the phase of the measurement signal is shifted with respect to the reference signal by

\[
\Delta \phi_M = 4\pi \frac{\Delta d}{\lambda}
\] (4.14)

where \(\lambda\) is the optical wavelength of the measurement signal in the propagation medium. It is this phase difference that allows the relative displacement of the device under test to be measured. The next section describes the CMOS-compatible fabrication process and the primary components used to construct the interferometer, along with results of the displacement measurement mode.

### 4.1.1 The 300 mm SOI Process and the Integrated Interferometer

We take a brief detour to describe the CMOS-compatible fabrication technology platform and the interferometer’s on-chip layout. The interferometer is fabricated on a 300 mm CMOS-compatible Silicon on Insulator (SOI) process using 193 nm optical
immersion lithography at the College of Nanoscale Science and Engineering, University of Albany. This advanced 300 nm silicon photonics platform has already led to a number of demonstrations including ultra-low power silicon modulators [50], rapid, wide-angle-steered phased arrays [51], and the integrated SSBMs used in this design to name a few. The process utilizes a 220 nm silicon top layer for the fabrication of optical waveguides and a 2 \( \mu m \) buried oxide (BOX) layer as the under-cladding. The process supports two silicon etches, a shallow etch with a 110 nm depth and a full 220 nm silicon etch to define the geometry of photonic structures. There are also two doping types, n-type and p-type, which can be utilized to form p-n junctions for electro-optic modulators such as the SSBMs, integrated resistors, and thermo-optic heaters. Furthermore, each doping type has a light(-) and high(+) concentration level. The lightly-doped concentration is used to create low absorption loss electro-optic structures (p-n junctions for phase shifters), while the highly-doped concentration, with its low ohmic resistance, is utilized for contacts to other structures.

Two copper interconnection levels (Metal 1 and Metal 2) and two copper contact levels (Contact and Via) are used to provide on-chip electrical interconnections within and between layers and for external electrical testing (e.g. probe pads). An epitaxially grown germanium layer is available for the fabrication of optical photodetectors. The availability of integrated photodetectors is a significant advance that sets this design apart from earlier attempts at on-chip integration requiring external photodetectors. These layers, and several additional layers not described here, can be integrated on the same chip to form a nearly complete silicon photonics system. A key goal for future designs is to integrate a laser on-chip with the interferometer. This is possible due to the availability of two silicon nitride layers for the hybrid integration of on-chip Erbium-doped lasers, which have been demonstrated previously [52]. Another advantage gained from operating on a CMOS-compatible process is that CMOS electronic circuits can be integrated with photonic circuits utilizing methods such as wafer bonding to provide sophisticated electrical control over the photonic circuits. The integrated heterodyne interferometer fabricated using this process is shown in Figure 4-2.
A 300 mm wafer is divided into a number of identical reticles measuring 32 mm high by 26 mm wide. The reticle is further divided into columns of varying width. The most common widths are 3 mm and 6 mm, but can vary depending on fabrication needs. The integrated interferometer was placed on a 32 mm by 6 mm column, with the actual device dimensions being approximately 1 mm high by 6 mm wide. The 6 mm width is dictated primarily by the length of the SSBMs. The interferometer also uses edge-coupling for the transmitted and received measurement signals, so the waveguides must go all the way to the column’s edge. The frequency modulation to create the dual heterodyne tones is provided by a pair of SSBMs, shown conceptually in Figure 4-3.

Each SSBM measures approximately 2 mm in length and consists of two Mach-Zehnder (MZ) phase modulators operated in a push-pull configuration. Within a SSBM the phase modulators share a common DC bias voltage, but are driven by RF signals in phase quadrature. When operated in this manner, the frequency of the input is shifted by an amount equal to the RF signal such that

\[ f_{\text{Modulated}} = f_0 \pm f_{RF} \]  

(4.15)
Figure 4-4: Single sideband operation of the SSBM at 1570 nm. Carrier and sideband suppression >15-dB is seen at 1570 nm.

Total insertion loss is estimated to be approximately 7-dB for the unsuppressed carrier and the 3-dB frequency was measured at 10 GHz. Single-sideband operation at 1.55 μm showed more than 18-dB carrier suppression and more than 15-dB spurious sideband suppression. Whether the upper or lower sideband is obtained depends on the relative bias between the upper and lower arms. Thus suppression of the carrier and the unwanted sideband is sensitive to the relative phase between the phase shifters. Due to fabrication tolerances, each modulator must be calibrated to achieve the correct phase relationships. The phase compensation is accomplished using three thermo-optic heaters residing in the MZ arms. In total, each SSBM requires six control voltage: 2 RF voltages, 1 DC bias voltage, and 3 heater bias voltages. For more details on similar designs, the reader is referred to [53][54].

The Ge detectors, similar to one described in [55], measure 1.2 μm wide by 10 μm in length with an 800 nm high Ge trench. The detectors have a measured responsivity of approximately 0.5 A/W at 1.55 μm and a measured bandwidth of 35 GHz. Because the circuit operates at a fairly low intermediate frequency (500 kHz), and minimizing dark current is more important than reducing junction capacitance, the detectors are
operated with zero bias voltage. Images of the SSBM and detector sections are shown in Figure 4-5 with 3-D models of the detectors shown in Figure 4-6.

Figure 4-5: Images of the (a) SSBM and (b) detector sections of the fabricated device.
4.2 Results for the Displacement Mode

4.2.1 Experimental Setup

This section will describe the experimental results when the interferometer is operated in the displacement measurement mode with the experimental setup shown in Figure 4-7.

The measurement and reference photocurrents are converted to voltages via transimpedance amplifiers (TIAs), shown functionally in Figure 4-8. These produce two
voltage waveforms with the same intermediate frequency \( f_{IF} \), but in general different phases. The TIA outputs are analog band-pass filtered to minimize the noise bandwidth, then sampled by a high-speed oscilloscope. The oscilloscope is controlled from a PC-based MATLAB script to save the data for subsequent analysis. The circuit schematic and layout for the TIA receiver circuit are shown respectively in Figure 4-9 and Figure 4-10. The receiver input is AC coupled to prevent DC photocurrents from saturating the TIAs. The Texas Instruments (TI) OPA656, a wideband (500 MHz unity gain bandwidth), unity-gain stable, operational amplifier with a JFET input stage, was used as the TIA. The JFET input stage provides both low input bias current and low voltage noise. The differential amplifiers are TI INA157s operated in a Gain=2 feedback configuration. Because the TIAs are operated in a negative feedback configuration, their noise gain and signal gain are different. While both quantities are important, the noise gain ultimately determines the circuit stability. Detailed expressions for the signal and noise gains are given in Appendix B for different AC coupling networks.

Figure 4-8: Functional diagram of the TIA receiver circuit.
Figure 4-9: Schematic for the TIA receiver board.
Figure 4-10: TIA receiver circuit as fabricated on PCB.
4.2.2 Experimental Results

Figure 4-11 shows the measurement voltage for several CCR displacements. The measurement signal is compared to the reference signal and the phase is recovered via quadrature demodulation, as described in Chapter 3 by

$$\phi = \arctan \frac{Q}{I} \quad (4.16)$$

Because the arctangent is modulo $2\pi$, the phase shift has an ambiguous length of $\lambda/2$. The minimum detectable phase shift, and hence the minimum displacement, is limited primarily by system noise and ambient vibrations. To estimate the lower performance limit, the system noise is characterized by turning off all AC units and placing a tube lens between the collimating lens and the CCR to reduce phase noise due to air turbulence. Data was then taken over several periods, with the results for a quarter-second measurement and one-second measurement shown in Table 4.1.

<table>
<thead>
<tr>
<th></th>
<th>Average Phase (rad)</th>
<th>RMS Phase Noise (rad)</th>
<th>RMS Position Noise (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{1}{4}$ Sec</td>
<td>-2.5126</td>
<td>0.0037</td>
<td>0.4563</td>
</tr>
<tr>
<td>1 Sec</td>
<td>-2.5467</td>
<td>0.0043</td>
<td>0.5302</td>
</tr>
</tbody>
</table>

Table 4.1: Static phase noise statistics for a one-second measurement period.
The results are fairly consistent between periods. Although the RMS phase noise for one-second is slightly higher than for a quarter-second, there does not appear to be a strong time dependence in either the mean phase or variance, allowing stationarity to be assumed over most practical measurement periods. Figure 4-12 shows static noise measurements taken over one second, along with the probability distribution function (PDF). The static noise fits well to a Gaussian distribution, which was surmised in Chapter 3 for a large number of white noise contributions. The distribution shown here is based on 2 million sample points taken over a single measurement. Lastly, the phase noise spectrum, shown in Figure 4-13, shows no discernible higher frequency terms that might indicate vibration or other non-random sources.

Figure 4-12: Static noise and PDF for a one-second measurement period.
A great deal has been made about the benefits of balanced detection in reducing the correlated noise between photodetectors. This property is illustrated in Figure 4-14, where the phase noise is shown for the balanced configuration and for each measurement detector separately. Detector 1 has a standard deviation of 0.0064 and Detector 2 has a standard deviation of 0.0049, with a correlation coefficient of 88%. However, the phase difference has a standard deviation of 0.0031, which is a considerable improvement over the values from the individual detectors. That the correlation between the detectors is so high should come as no surprise, as the use of a single laser is expected to cause high correlation between the two detectors. The difference in the noise voltages, as measured by the standard deviation, may be attributed to unequal coupler splitting and/or unequal detector responsivities.

Once the static noise data were collected, a nano-positioner with 0.5 nm RMS position noise and 100 ppm accuracy over a 100 μm range was displaced in 5 nm steps from zero to 1.55 μm, with the wrapped phase shown in Figure 4-15. The phase ambiguity occurs at a displacement of 775 nm as expected for a free space wavelength of 1.55 μm.
Figure 4-14: Balanced and unbalanced phase noise from the measurement detectors taken over one second.

Figure 4-15: Wrapped phase versus displacement over a distance of 1.55 \( \mu m \).
However, there also exists a periodic modulation of the ideally linear phase. This modulation, more commonly known as a phase nonlinearity [9,56–59], is caused by frequency leakage between the SSBMs. The nonlinearity was corrected by deriving the equations of the ellipse described by the orthogonal IQ components. Plotting Q vs I produces the well-known Lissajous curves, shown in Figure 4-16 as the CCR was displaced ±2 μm by a 50 Hz triangular wave. The nonlinearity causes a non-zero eccentricity of the ellipse and a shift of the center away from the origin. A detailed derivation of the equations describing the ellipse in the presence of this nonlinearity is given in Appendix A. By fitting the IQ data using least squares methods, the ellipse can be “circularized” and the nonlinearity reduced, as shown in Figure 4-16.

Figure 4-16: Lissajous curves and phase for uncorrected and corrected 50 Hz triangular-wave displacement.
Table 4.2 summarizes the results of stepping the CCR from zero to 1.55 \( \mu m \) in steps of 5 nm, 10 nm, and 25 nm, respectively. The corrected measurements show a marked improvement in the position variance, and hence the RMS position noise. In fact, the corrected standard deviations are similar irrespective of step size.

<table>
<thead>
<tr>
<th>Step Size</th>
<th>Average</th>
<th>Std Deviation</th>
<th>Corrected Avg</th>
<th>Corrected Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 nm</td>
<td>4.94</td>
<td>3.56</td>
<td>4.88</td>
<td>2.52</td>
</tr>
<tr>
<td>10 nm</td>
<td>9.91</td>
<td>5.16</td>
<td>9.87</td>
<td>2.19</td>
</tr>
<tr>
<td>25 nm</td>
<td>25.04</td>
<td>11.76</td>
<td>25.07</td>
<td>2.36</td>
</tr>
</tbody>
</table>

Table 4.2: Summary of measurements for 5 nm, 10 nm, and 25 nm displacement steps.

Figure 4-17: Unwrapped corrected and uncorrected phase over a distance of 1.55 \( \mu m \).

4.3 Conclusion

In this chapter a silicon integrated heterodyne interferometer with on-chip splitters, modulators, and detectors was demonstrated that obtains a noise-limited position resolution of approximately 2 nm. This compares favorably with commercial interferometers such as the Keysight Technologies 10702A linear interferometer, which
obtains a 10 nm resolution at a wavelength of 632 nm. Compared to conventional interferometers based on bulk optics, this chip-scale device could offer a significant reduction in both size and cost, as well as increased stability due to the CMOS-compatible silicon photonic integration, thus creating opportunities for applications in 3-D inspection, handheld devices, and portable biomedical units. In the next chapter the vibrometer mode will be introduced.
Chapter 5

The Laser Doppler Vibrometer

5.1 Background

This chapter examines a particularly interesting application of the heterodyne interferometer, that of a Laser Doppler Vibrometer (LDV). The Doppler spectrum of a vibrating object can be exploited to make non-contact measurements of both the frequency and amplitude of vibration. This is particularly useful in a number of structural engineering and scientific applications, such as vibrational measurement and characterization of hard disk drives, where it is desirable to measure dynamical characteristics without mass loading the object or otherwise perturbing its physical characteristics \[17\]. LDVs have also been used to measure the vibration and deflection of MEMS structures such as Atomic Force Microscopy cantilever probes \[60,61\]. LDVs also have important applications as medical sensors, where they have been used to measure the flow rate of blood in arteries \[15\] and to examine the structure of the inner ear \[62\]. Because of the interest in this application, experiments were conducted using the integrated interferometer in which the object under test was a vibrating speaker with a retroreflector mounted to the front, as shown in Figure 5-1. This experiment also validated the operation of the interferometer in the first devices that were fabricated. Due to a process issue, the electrical contacts to the SSBMs were mostly open circuits, so there was minimal modulation of the fundamental carrier. Therefore, the device was essentially operating as a homodyne interferometer whose
response to displacement was a modulation of the amplitude. This is in contrast to a heterodyne interferometer that produces a beat frequency whose phase is shifted in response to a displacement. However, if an object is in motion, a Doppler frequency shift is imparted on the measurement signal in both the homodyne and heterodyne configurations. This frequency shift will generate a beat signal in the homodyne system equal to the Doppler frequency. Because of this, it was possible to determine that the interferometer did indeed work, even if only in the homodyne mode. But this allowed early issues with the SSBMs to be identified and corrected on subsequent fabrication runs.

Figure 5-1: Vibrometer measurements using a speaker-mounted retroreflector.

5.2 Homodyne Vibrometry using the Indirect Method

Freschi et.al [63] analyzed the dynamics of a loud speaker and developed an electrical analog of its damped harmonic oscillator behavior as a function of applied drive voltage and frequency, the magnetic field of the speaker’s magnet, and physical characteristics such as cone stiffness. The speaker used in this experiment was a Realistic 40-1354 with an open air resonance frequency of 51 Hz, which could be verified experimentally by observing the frequency response as modulation frequency was lowered.
below that point. The weight of the attached retroreflector (22 g) appeared to have minimal impact on the free air resonance. Although the specifics of this particular analysis are not critical, it sets the stage for the analysis of a general harmonic oscillator. First, the basic dynamical equations of a simple harmonic oscillator are derived before determining the vibrational spectrum of such a system. To start, the position of the retroreflector under simple harmonic motion can be expressed as

\[ x(t) = x_0 \sin(2\pi f_v t) \]  

(5.1)

Here \( x_0 \) is the maximum displacement of the speaker cone and \( f_v \) is the vibrational frequency. Because the initial displacement is zero at \( t = 0 \), the initial phase of the displacement is also zero. The speaker was driven by a function generator at frequencies starting from 30 Hz, which is below the free air resonance, to upwards of several hundred Hertz. The effect of this harmonic displacement can be considered as a sinusoidal phase modulation of the optical measurement signal with wavelength \( \lambda_m \), which for a double-pass interferometer is given by

\[ \phi_d(t) = \frac{4\pi x(t)}{\lambda_m} = \frac{4\pi x_0}{\lambda_m} \sin(2\pi f_v t) \]  

(5.2)

By adding the time-dependent Doppler frequency to that of the optical carrier, a frequency-based approach can be applied. Taking the first derivative of the position \( x(t) \) with respect to time, the instantaneous velocity \( v_d \) is found and the Doppler frequency \( f_d \) can be computed from

\[ f_d(t) = \frac{2v_d}{c} f_m = \frac{4\pi x_0}{\lambda_m} f_v \cos(2\pi f_v t) \]  

(5.3)

But the resulting frequency must be integrated to obtain the time-dependent phase, which gives the same expression as Eq. (5.2). This is simply a statement of the equivalence of frequency and phase modulation for sinusoidal phase modulation \[33\]. It is also easy to verify that \( \phi_d(t) = 2\pi \int f_d(t) \, dt \) as expected. Using the sinusoidal phase
modulation, the measurement and reference fields may be written respectively as

\[
E_M = A \cos \left( 2\pi f_M t + \frac{4\pi x_0}{\lambda_M} \sin (2\pi f_v t) + \phi_M \right) \quad (5.4)
\]

\[
E_R = B \cos (2\pi f_R t + \phi_R) \quad (5.5)
\]

The phase terms \( \phi_M \) and \( \phi_R \) account for any initial phase difference in the measurement and reference arms, and are assumed to be constant over the duration of the measurement. The intensity at the detectors due to mixing of the reference and measurement signals were derived previously, so only the intensity terms relevant to the interference between the signals is presented and can be written as

\[
P_{ac,het}(t) \propto \cos \left[ 2\pi (f_R - f_M) t - \frac{4\pi x_0}{\lambda_M} \sin (2\pi f_v t) + (\phi_R - \phi_M) \right] \quad (5.6)
\]

Defining \( f_{IF} = f_R - f_M \), \( \Delta \phi = \phi_R - \phi_M \), and applying the trigonometric identity for the cosine of a sum of two arguments, the intensity term can be expanded into a more useful form given by

\[
P_{ac,het}(t) \propto \left[ \cos (2\pi f_{IF} t + \Delta \phi) \cos \left( \frac{4\pi x_0}{\lambda_M} \sin (2\pi f_v t) \right) + \sin (2\pi f_{IF} t + \Delta \phi) \sin \left( \frac{4\pi x_0}{\lambda_M} \sin (2\pi f_v t) \right) \right] \quad (5.7)
\]

At this point it is instructive to consider the separate cases of a homodyne system where \( f_{IF} = 0 \), and a heterodyne system where \( f_{IF} \) is non-zero. Rewriting the equation above for the homodyne case yields the far simpler expression

\[
P_{ac,homo}(t) \propto \left[ \cos (\Delta \phi) \cos \left( \frac{4\pi x_0}{\lambda_M} \sin (2\pi f_v t) \right) + \sin (\Delta \phi) \sin \left( \frac{4\pi x_0}{\lambda_M} \sin (2\pi f_v t) \right) \right] \quad (5.8)
\]

The multiplicative terms involving the sine and cosine of the harmonic displacement can be written in terms of Bessel functions of the first kind of order \( n \) (\( J_n \)) via the
Jacobi-Anger expansion. Defining \( \beta = \frac{4\pi u}{\lambda_0} \), the Jacobi-Anger expansions yield

\[
\begin{align*}
\cos (\beta \sin (2\pi f_v t)) &= J_0 (\beta) + 2 \sum_{n=1}^{\infty} J_{2n} (\beta) \cos(2n \cdot (2\pi f_v t)) \\
\sin (\beta \sin (2\pi f_v t)) &= 2 \sum_{m=1}^{\infty} J_{2m-1} (\beta) \sin((2m - 1) \cdot (2\pi f_v t))
\end{align*}
\tag{5.9}
\]

Applying this expansion to the time-domain heterodyne and homodyne intensities incident at the measurement detectors gives the following expressions.

\[
P_{ac,het} (t) \propto \sum_{n=0}^{\infty} J_n (\beta) \left[ \cos (2\pi (f_{IF} + n f_v) t + \Delta \phi) + \right. \\
\left. \cos (2\pi (f_{IF} - n f_v) t + \Delta \phi) \right]
\tag{5.10}
\]

\[
P_{ac,homo} (t) \propto \left[ \cos(\Delta \phi) \sum_{n=0}^{\infty} J_{2n} (\beta) \cos (2\pi (2n f_v) t) + \right. \\
\left. \sin(\Delta \phi) \sum_{m=1}^{\infty} J_{2m-1} (\beta) \sin (2\pi ((2m - 1) f_v) t) \right]
\tag{5.11}
\]

The Fourier transform of the time-domain intensities can be computed for both the heterodyne and homodyne systems, which gives the spectrum of the photocurrent to within a factor of proportionality equal to the detector responsivity. Applying the modulation theorem of Fourier transforms to the product of two temporal functions yields the following expression for the heterodyne photocurrent spectrum.

\[
i_{het} (f) \propto \sum_{n=0}^{\infty} J_n (\beta) \left[ e^{j\Delta \phi} \{ \delta (f - f_{IF} - n f_v) + (-1)^n \delta (f - f_{IF} + n f_v) \} + \right. \\
\left. e^{-j\Delta \phi} \{ (-1)^n \delta (f + f_{IF} - n f_v) + \delta (f + f_{IF} + n f_v) \} \right]
\tag{5.12}
\]

Here Euler’s formula has been used to expand Eq.(5.10) by applying \( \cos \Delta \phi + j \sin \Delta \phi = e^{j\Delta \phi} \) and \( \cos \Delta \phi - j \sin \Delta \phi = e^{-j\Delta \phi} \). This shows a rather complex spectrum of sidebands around the intermediate frequency. There are, in theory, an infinite number of such sidebands on each side of the intermediate frequency, which is equivalent to a signal with infinite bandwidth. In reality, the amplitude of the Bessel functions becomes smaller as their order increase, so the summation can usually be truncated after several terms. As an aside, this expression also illustrates the concept of bandwidth as it is typically used in frequency modulation (FM) systems, in that the occupied
bandwidth must include both the IF and the modulation frequency of the oscillator. In a FM system, a quantity known as the modulation index $m_f$ serves the same role as $\beta$ in the above derivation, and determines the number of sidebands that contribute significantly to the total power. “Significant” is usually taken to be the sidebands containing approximately 98% of the total power. In FM systems, the sidebands that satisfy this criterion are orders zero through $\beta + 1$. This relationship is encapsulated in the well-known Carson’s Rule that expresses the effective bandwidth as

$$B_c = 2(\beta + 1)f_v = 2 \left( \frac{4\pi x_0}{\lambda M} + 1 \right) f_v \text{ (Hz)}$$

(5.13)

What this means is that for a given IF, the value of $\beta$ has to be such that the sidebands become insignificant for values of the index $n$ as $f_{IF} - nf_v$ approaches zero to avoid aliasing. Or viewed another way, depending on the maximum displacement and modulation frequency, the IF must support enough significant sidebands without aliasing. This is similar to the restrictions placed on the Doppler frequency for the non-sinusoidal case. The number of harmonics in the passband of a phase modulated signal is given in [33] as

$$M_c = 2 \lfloor \beta \rfloor + 3 = 2 \left\lfloor \frac{4\pi x_0}{\lambda M} \right\rfloor + 3$$

(5.14)

where $\lfloor \cdot \rfloor$ is the floor function. This result shows that the amplitude of oscillation $x_0$ plays an important role in determining not only the effective bandwidth $B_c$, but also the number of harmonics within that bandwidth. Going one step further, the power spectrum of the current can be written as

$$|i_{het}(f)|^2 \propto \sum_{n=0}^{\infty} J_n^2(\beta) \begin{bmatrix} \delta (f - f_{IF} - nf_v) + \delta (f - f_{IF} + nf_v) + \\ \delta (f + f_{IF} - nf_v) + \delta (f + f_{IF} + nf_v) \end{bmatrix}$$

(5.15)

Because all terms in Eq.(5.15) are multiplied by the same constants of proportionality, they have been dropped moving forward. What this expression shows is that given the power spectrum of the sinusoidally driven speaker, not only can the modulation
frequency \( f_v \) be determined by examining the harmonic composition, but also the vibrational amplitude \( x_o \) using the relationship between the power ratios of adjacent harmonics. Taking the ratios of adjacent harmonics eliminates the proportionality constants, justifying our decision to drop them in the expression for the power spectrum. The corresponding expressions for the homodyne current spectral density and the power spectral density are

\[
i_{\text{homo}}(f) \propto \cos(\Delta \phi) \left( \sum_{n=0}^{\infty} J_{2n} (\beta) [\delta (f + 2nf_v) + \delta (f - 2nf_v)] \right) - \]

\[
i_{\text{homo}}(f) \propto \left| \cos^2(\Delta \phi) \left( \sum_{n=0}^{\infty} J_{2n}^2 (\beta) [\delta (f + 2nf_v) + \delta (f - 2nf_v)] \right) + \cos^2(\Delta \phi) \left( \sum_{m=1}^{\infty} J_{2m-1}^2 (\beta) [\delta (f + (2m-1)f_v) + \delta (f - (2m-1)f_v)] \right) \right|
\]

Comparing the expressions for the homodyne and heterodyne power densities provides some insight on possible methods to gain information about the maximum displacement \( x_o \), which resides in the argument \( \beta \) of the Bessel function coefficients. In the heterodyne system, the constants of proportionality for both the even and odd harmonics are equal, such that the ratio of \( \frac{\text{abs}(J_n)}{\text{abs}(J_1)} \) for \( n \) odd or even can be used to determine \( \beta \), and hence determine \( x_o \) for a known wavelength. This frequency domain method is sometimes known as the indirect method of measuring the vibrational parameters. However, these ratios are usually only monotonic up to some small value. For example, \( \frac{\text{abs}(J_3)}{\text{abs}(J_1)} \) is only monotonic for values of \( \beta \) up to approximately 3.5. For \( \lambda_m = 1550 \) nm, this corresponds to only about 430 nm. This is more than three orders of magnitude below the maximum linear displacement of the speaker, which is approximately 1 mm. Even for a moderate driving voltage, it will be exceedingly difficult to measure the displacement without ambiguity since above \( \beta = 3.5 \), \( \frac{\text{abs}(J_3)}{\text{abs}(J_1)} \).
becomes multi-valued and additional information is needed to find $\beta$. However, if the spectra has even order harmonics such as $J_2$, the ratio of $\text{abs}(J_2/J_1)$ could be used to resolve the ambiguity. By using several sidebands, the value of $\beta$ can be recovered from the recurrence relation for adjacent Bessel functions.

$$J_{\nu-1}(\beta) + J_{\nu+1}(\beta) = \frac{2\nu}{\beta} J_\nu(\beta)$$

(5.18)

The homodyne case is more complicated because the even and odd harmonics are scaled by different proportionality constants corresponding to the cosine and sine of the initial phase difference between the measurement and reference signals. So unless this phase offset is zero or measured a priori, the mixed ratio of even and odd harmonics cannot be used to resolve the ambiguity problem. Figure 5-2 shows spectral plots for the speaker driven at 100 Hz and 200 Hz, respectively. It can be clearly seen that the fundamental and harmonics occur at the proper frequencies for the given driving voltage.

For the speaker experiment, the interferometer was driven in the homodyne configuration, and no attempt was made to determine the initial phase error. Therefore it was not possible to unambiguously determine the displacement. But even in the absence of displacement data, this result demonstrates a potentially useful application of interferometry with several practical applications. In the next section the same measurement is conducted using a sinusoidally driven nano-positioner in the heterodyne configuration to determine whether the displacement amplitude can be correctly measured, along with the vibrational frequency, using both the indirect and direct methods. The nano-positioner has a well understood relationship between the driving voltage and displacement that will make it possible to characterize its harmonic motion.
5.3 Heterodyne Vibrometry using the Indirect and Direct Methods

While the indirect method requires only the measurement signal and a spectrum analyzer, it also necessitates measuring several sidebands to determine the displacement amplitude. Also, any frequency selectivity such as filter roll-off or amplitude modulation that alters the spectra will distort the relationship between sidebands, compromising the applicability of this method. A more robust method is to measure the phase by means of the IQ demodulator. This method, often known as the direct method, uses the time domain phase measurements from the IQ demodulator.
to determine both the oscillation frequency and amplitude. This holds even if the amplitude is much larger than the ambiguous range. Such a signal, whose frequency is the structural vibrational frequency and whose amplitude is proportional to the displacement by a known relationship, can be directly applied to a lock-in amplifier to control the cantilever of an atomic force microscope [60, 61].

Several times throughout this thesis, in both the theory and the experimental data, the phase ambiguity has been shown to be a fundamental fact of life in interferometric distance measurements. But as will be shown here, and was alluded to in Chapter 2, it is possible to unambiguously measure displacements much greater than $\lambda/2$. This is possible because the ambiguity is a manifestation of insufficient sampling, and if measurements occur with sufficient frequency, the phase transitions that lead to the ambiguity can be captured. The rate at which the measurement must occur is tied to the object’s velocity. For a target undergoing simple harmonic motion the position and phase are sinusoidal, so to resolve the ambiguity at least two measurements should be made for every $2\pi$ of phase shift. In other words, at least once every $\pi$ of phase shift. For harmonic motion the phase is nonlinear in time, but the appropriate sampling rate can be determined at the maximum phase shift, which can be related to the Doppler frequency by

$$\frac{d\phi(t)}{dt} = \frac{4\pi dx(t)/dt}{\lambda_c} = \frac{4\pi v_d(t)}{c} f_c = 2\pi f_d(t)$$

(5.19)

The maximum rate of change occurs in the vicinity of maximum Doppler frequency shift, so Eq.(5.19) can be approximated by

$$\frac{\Delta \phi}{\Delta t} \bigg|_{Max} = 2\pi f_{d,Max}$$

(5.20)

For $\Delta \phi = \pi$, the minimum sampling period is given by

$$\Delta t_{Min} = \frac{\Delta \phi}{2\pi f_{d,Max}} = \frac{1}{2 f_{d,Max}}$$

(5.21)

Therefore, the ambiguity is resolvable if sampling is done in accordance with the
Nyquist-Shannon theorem using twice the peak Doppler shift as the sampling frequency. For the second part of the experiment a corner cube reflector (CCR) was mounted to a high precision nano-positioner with a range of $\pm 50 \, \mu m$. When operated in closed-loop feedback mode with PID (proportional,integral,derivative) control parameters set to achieve a critically-damped impulse response, the amplitude response has a 3-dB cutoff of 80 Hz. The nano-positioner has a peak-peak (pk-pk) amplitude of $5 \, \mu m/V_{pk-pk}$ when driven with an external voltage source, with the entire range corresponding to 20 Vpk-pk. The nano-positioner was driven sinusoidally at frequencies from 1 Hz to 75 Hz, just below the 3-dB amplitude cutoff. The corresponding voltage amplitudes were 0.1 Vpk-pk to 10 Vpk-pk, corresponding to a pk-pk range of 500 nm to 50 $\mu m$, a distance almost 65 times the ambiguous length. Figure 5-3 shows the phase computed directly from the IQ demodulator for a 1 Hz/50 $\mu m$ pk-pk modulation and a 50 Hz/10 $\mu m$ pk-pk modulation. The frequency can be easily determined from the period, and has the correct value for both cases. The pk-pk phase variation is measured to be 411 radians and 82 radians for the 1 Hz and 50 Hz signals, respectively. The pk-pk phase deviation is equal to $2 \beta$, where $\beta$ was previously defined as

$$\beta = \frac{4\pi x_0}{\lambda_M} \quad (5.22)$$

For an optical wavelength of $\lambda_m = 1530$ nm, this corresponds to a maximum displacement of $x_0 = 25 \, \mu m$ for the 1 Hz signal and $x_0 = 5 \, \mu m$ for the 50 Hz signal. Since $x_0$ is half the pk-pk value, the results are as expected. In fact, the results only deviate appreciably from the expected values at low voltages where position noise becomes a greater contributor to overall error, and as the frequency approaches the 3-dB amplitude roll-off.

But even near the 3-dB cutoff, the measured displacements approximate the amplitude transfer function well. For frequencies above several hundred Hz, the nano-positioner controller current output becomes the primary performance limitation, as the controller can no longer provide sufficient current to the piezo stack. But it is highly likely that the LDV will work at higher frequencies and larger displacement.
amplitudes, assuming the sampling rate is sufficient. The Doppler frequency can be computed from the time derivative of the phase, and from this the instantaneous velocity of the target can be obtained. Next, consider an example utilizing the indirect method. Figure 5-4 shows the first five sidebands of the measurement spectrum for a target driven at 5 Hz and 500 nm pk-pk amplitude.

The data was taken over a one-second period to obtain a theoretical resolution bandwidth of 1 Hz. However, the indirect method requires accurate measurement of the sideband amplitudes, so reducing spectral leakage is important. Hence, zero-padding was used to increase the FFT frequency bin size to 1/4 Hz, and a Blackman-Harris window was applied to further reduce spectral leakage, with a resulting increase in the 3-dB width of the sidebands. The sideband spacing about the intermediate
frequency is in multiples of 5 Hz as expected. Table 5.1 gives the sideband amplitudes and the ratio between the harmonics and the fundamental at 5 Hz.

<table>
<thead>
<tr>
<th></th>
<th>( J_1 ) (5 Hz)</th>
<th>( J_2 ) (10 Hz)</th>
<th>( J_3 ) (15 Hz)</th>
<th>( J_4 ) (20 Hz)</th>
<th>( J_5 ) (25 Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(dB)</td>
<td>-0.5747</td>
<td>-0.1275</td>
<td>-13.14</td>
<td>-20.08</td>
<td>-35.16</td>
</tr>
<tr>
<td>( J_x/J_1 )</td>
<td>1</td>
<td>1.05283</td>
<td>0.23536</td>
<td>0.10586</td>
<td>0.01865</td>
</tr>
</tbody>
</table>

Table 5.1: Vibrational amplitude computed using the relative amplitudes of the first five harmonic sidebands.

Rearranging the recurrence relation of Eq. (5.18), \( \beta \) can be solved in terms of the order \( \nu \) and the ratio of adjacent sidebands as

\[
\beta_\nu = \frac{2\nu}{J_{\nu-1}/J_{\nu} + J_{\nu+1}/J_{\nu}} \quad (5.23)
\]

The amplitudes ratios are computed from the data in Table 5.1. The relative signs of the sidebands are lost in the Fourier transform magnitude spectrum, but can be recovered from the phase spectrum. For \( \nu = 2, 3, \) and 4, the following values of \( \beta \) were obtained.

<table>
<thead>
<tr>
<th>( \beta_2 )</th>
<th>( \beta_3 )</th>
<th>( \beta_4 )</th>
<th>( \beta_{Avg} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.653</td>
<td>2.528</td>
<td>1.807</td>
<td>1.996</td>
</tr>
</tbody>
</table>

Table 5.2: \( \beta \)s computed using the relative sideband amplitudes.
Taking the average of the three values gives $\beta = 1.996$, which corresponds to a peak displacement of $x_0 = 243$ nm. This is in close agreement to the theoretical value of 250 nm. To determine $\beta$, all values of $x_0$ up to some maximum can be found that simultaneously satisfy $\beta_2$ through $\beta_4$ as determined from Eq. (5.23), then use the average to resolve any ambiguity. When discussing the limitations of the indirect method, it was mentioned that any frequency dependency in the passband that alters the relationship between the sidebands will limit the applicability of this method. Chapter 4 discussed the nonlinearity that occurs when frequency mixing is present, and the relevant equations describing this phenomenon are derived in Appendix A.

Before leaving this section, the nonlinearity is revisited and its effect on the vibrometer is examined. In obtaining the results for this section it was possible to minimize the nonlinearity to the point where the indirect method should be reasonably accurate. Appendix A shows that the equations of the in-phase and quadrature components in the presence of the linearity are

$$
I(L) = (AB' + A'B) + (AB + A'B') \cos \phi(L) \\
Q(L) = (AB - A'B') \sin \phi(L)
$$

(5.24)

where $A$ and $B$ are the desired field amplitudes in SSBM1 and SSBM2, and $B'$ and $A'$ are the leakage frequency terms in SSBM1 and SSBM2, respectively. If the $AB$ product terms in Eq. (5.24) are factored out and the constant origin offset in the in-phase component is ignored, the in-phase and quadrature components are

$$
I(L) = AB \left(1 + \frac{A'B'}{AB}\right) \cos \phi(L) = AB \left(1 + \Gamma\right) \cos \phi(L) \\
Q(L) = AB \left(1 - \frac{A'B'}{AB}\right) \sin \phi(L) = AB \left(1 - \Gamma\right) \sin \phi(L)
$$

(5.25)

where $\Gamma = \frac{A'B'}{AB}$. If the phase $\phi(L)$ is written in terms of the harmonic phase, Eq. (5.25) becomes

$$
I(L) = AB \left(1 + \Gamma\right) \cos (\beta \sin (2\pi f_v \cdot t)) \\
Q(L) = AB \left(1 - \Gamma\right) \sin (\beta \sin (2\pi f_v \cdot t))
$$

(5.26)
Referring back to Eq. (5.9), the in-phase component $I(L)$ corresponds to the even sidebands of the measurement signal and the quadrature component $Q(L)$ corresponds to the odd sidebands. The result is that the nonlinearity causes the even sidebands to be scaled by $(1 + \Gamma)$ and the odd sidebands by $(1 - \Gamma)$. This is indeed what was observed in Figure 4-16. This scaling obscures the true relationship between adjacent sidebands given by the Bessel function recurrence relationship in Eq. (5.23). However, if there are a sufficient number of sidebands, the relationship between several adjacent odd sidebands and/or several adjacent even sidebands can be used to find a solution to $\beta$. Still, the increased complexity introduced by the indirect method in the presence of a nonlinearity may not be worth the effort. However, the compensation method discussed in Chapter 4 and derived in Appendix A for correcting the phase nonlinearity is applicable when computing the phase using the direct method, and is thus resilient to the nonlinearity.

5.3.1 The Doppler Frequency Shift and Velocity

The vibrometer is brought full circle by examining the time-dependent Doppler frequency shift induced by the harmonic motion, and from this computing the time-dependent velocity. The Doppler frequency was defined in Eq. (5.3) and is directly proportional to the time-dependent phase given in Eq. (5.2), except that it is $90^\circ$ out of phase and multiplied by the vibrational frequency $f_v$. The velocity is computed from the Doppler shift via

$$v_d(t) = \frac{f_d(t)}{2\lambda_m}$$

(5.27)

where $\lambda_m$ is the optical wavelength in the propagation medium. The Doppler frequency and velocity for a target driven at 1 Hz/10 Vpk-pk and whose phase is shown in Figure 5-3 is compared with that of a target driven at 25 Hz/10 Vpk-pk. Because the peak-peak driving voltages are the same in both cases, the peak-peak phase excursions should be the same, but at different frequencies. The time-dependent phase for the 25 Hz/10 Vpk-pk case is shown in Figure 5-5, and the peak-peak values are indeed equal for both.
Now that the peak-peak phase excursions have been shown to be equal, it is possible to compare the time-dependent Doppler frequencies and velocities. The Doppler frequencies are shown in Figure 5-6. The maximum Doppler frequency for the 1 Hz case is equal to the maximum phase displacement, but shifted in time by $90^\circ$. Again, this is consistent because the scaling factor between the phase and the Doppler frequency is the vibrational frequency, which is 1 Hz in this case. The relationship also holds for the 25 Hz case, since $200 \text{ radians} \cdot 25 \text{ Hz/radian}$ is approximately 5 kHz, the peak Doppler frequency for the second case. The time-dependent velocities computed from Eq. (5.27) are shown in Figure 5-7.

Before closing, a technical challenge that arose when computing the derivatives (discrete differences) to find the Doppler frequency and velocity is briefly discussed. It is widely known that taking discrete derivatives is inherently noisy and amplifies high frequency noise. The result of doing a simple discrete difference is shown in Figure 5-8 for the 25 Hz/10 Vpk-pk case. While the velocity can be discerned in this example, in some situations the quantities obtained are so noisy as to be useless. This was solved by using a Savitzky-Golay (S-G) filter. A S-G filter is a digital smoothing filter that allows the SNR of a data set to be significantly increased without distorting the underlying signal. This is accomplished by fitting successive data frames of fixed size using a low-order polynomial by the method of linear least squares.
Figure 5-6: Derived Doppler frequency for 1 Hz and 25 Hz oscillations at 10 Vpk-pk.
Figure 5-7: Derived velocity for 1 Hz and 25 Hz oscillations at 10 Vpk-pk

Figure 5-8: Noisy velocity obtained without a Savitzky-Golay filter.
5.4 Conclusion

In this chapter the utility of the integrated heterodyne interferometer as a Laser Doppler Vibrometer was demonstrated. In this modality, the interferometer has the potential to facilitate a number of existing applications such as Atomic Force Microscopy that benefit from LDVs, but where conventional devices based on bulk optics are too large and expensive. The next chapter will focus on the final operational mode, LIDAR.
Chapter 6

The LIDAR Mode

6.1 Background

This chapter will introduce the LIDAR operational mode and present the theory and some early results. While LIDAR may invoke a negative image of police speed traps in the minds of many, it is one of the most important and ubiquitous of all sensing technologies. LIDAR systems are widely used in altimetry, imaging, surveying, and atmospheric remote sensing, to name just a few examples. One area seeing increased utilization of LIDAR is in autonomous vehicles, where it may be deployed as the sole sensor for obstacle detection and avoidance, or to augment other on-board sensors.

First developed in the 1960s following the invention of the laser, LIDAR borrowed heavily from its older sibling RADAR. Because RADAR systems theory is directly applicable to LIDAR systems, the analysis of LIDAR is relatively straightforward. Both LIDAR and RADAR systems may utilize either continuous wave (CW) or pulsed waveforms. In a pulsed waveform system, the distance to the object of interest is determined by the “time of flight” principle. Because the transmitted pulses have well-defined temporal edges, the distance can be measured using the round-trip time delay between the transmitted and the reflected pulses. In contrast to pulsed systems, CW systems do not utilize pulse modulation. Since there are no unique boundaries in a CW waveform, time of flight methods are not possible with an unmodulated CW signal.
CW waveforms are most often used in interferometers due to the simplicity of generation, but the measured distances are limited to half the laser wavelength due to phase ambiguities that occurs beyond this distance. In a pulsed RADAR system, the range resolution, or the ability to discern two closely-spaced objects, is one of the most important specifications. High range resolution necessitates narrow, high-bandwidth pulses. However, the ability to detect targets in background clutter and noise is closely tied to the Signal-to-Noise Ratio (SNR). The SNR increases with increased energy on the target, which is obtained with longer pulses and higher power. Therefore, there is a conflict between the need to resolve closely-spaced targets and the need to increase target detectability. One way to address these conflicting requirements is to apply a linear frequency sweep (chirp) to each pulse. Linear frequency chirp, also known as Frequency Modulated Continuous Wave (FMCW), has been widely utilized in LiDAR systems and has an extensive literature [64–66]. In a FMCW RADAR/LiDAR system, the frequency is swept linearly with an instantaneous frequency $f(t)$ given by

$$f(t) = f_0 + kt$$

where $k$ is the frequency sweep rate of the waveform in units of Hz/s. In general, the phase of a waveform is the integral of the frequency such that

$$\phi(t) = 2\pi \int f(t)dt$$

For a constant frequency signal, this simply gives the linear time-dependent phase

$$x(t) = \cos(2\pi f_0 t + \phi_0)$$

But for a linearly-swept signal, the phase is quadratic in time and given by

$$x(t) = \cos \left[ 2\pi \left( f_0 t + \frac{kt^2}{2} \right) + \phi_0 \right]$$

The chirp has the effect of increasing the effective bandwidth of the waveform as the
frequency is swept. Applying pulse compression techniques to chirped waveforms, long temporal pulses can obtain the same range resolution as much narrower pulses while maintaining their SNR advantage. As shown in Figure 6-1, one consequence of the linear chirp is to give the CW waveform a recognizable structure similar to that of a pulsed waveform, in that each sweep has a beginning and end. Fortunately, the linearly-chirped waveform can be utilized to overcome the inherent limitations of a CW LIDAR system.

![Figure 6-1: A linearly-chirped CW waveform in the time domain.](image)

The following analysis will demonstrate how a linearly-chirped CW waveform can be utilized in the integrated heterodyne interferometer to achieve a LIDAR system, allowing measurement distances far greater than the ambiguous range of a pure CW interferometer. Below, the heterodyne interferometer is reintroduced and its LIDAR operation is described when using a FMCW signal. Figure 6-2 shows a conceptual diagram of the LIDAR using a swept laser source. The swept laser source is evenly split with half the power passing into a pair of single-sideband modulators (SSBM). The SSBMs are driven by two RF signals that modify the input frequency according to

$$f_{Modulated} = f_0 + f_{RF}$$  \hspace{1cm} (6.5)

Assume the fields from the upper and lower SSBMs reach the two reference detectors ($R_1$ and $R_2$) and the upper measurement detector ($M_1$) with negligible time
Figure 6-2: The Integrated Heterodyne LIDAR

delay with respect to the $T = to$ reference plane, so that the fields from the upper and lower SSBMs at the $R_1, R_2$, and $M_1$ detectors can be written respectively as

$$S_{R1}(t_0) = A_{R1} \cos(2\pi(f_1 t_0 + kt_0^2/2) + \phi_{R1}),$$
$$S_{R2}(t_0) = A_{R2} \cos(2\pi(f_2 t_0 + kt_0^2/2) + \phi_{R2}),$$
$$S_{M1}(t_0) = A_{M1} \cos(2\pi(f_1 t_0 + kt_0^2/2) + \phi_{M1})$$

(6.6)

where $\phi_i$ are arbitrary initial phases. The signals $S_{R1}$ and $S_{R2}$ are non-linearly combined in a pair of reference detectors ($R_{D1}$ and $R_{D2}$) operated in a balanced configuration, with the resulting reference signal proportional to

$$S_R(t_0) \propto A_{R1}A_{R2}\cos(2\pi(f_2 - f_1)t_o + (\phi_{R2} - \phi_{R1}))$$

(6.7)

If $f_{IF} = f_2 - f_1$ is defined as the intermediate frequency (IF) and $\delta\phi_R = \phi_{R1} - \phi_{R2}$, the reference signal can be simplified to

$$S_R(t_0) \propto A_{R1}A_{R2}\cos(2\pi f_{IF}t_o + \delta\phi_R)$$

(6.8)
Next, the probe measurement signal that is coupled off-chip and reflected from the device under test (DUT), in this case a corner-cube reflector (CCR), is examined in more detail. This signal incurs an additional time delay versus the on-chip measurement signal $S_{M1}$ proportional to the round trip distance $2R$ in free space. Therefore, the total time delay of the probe measurement signal with respect to the on-chip measurement signal is

$$t_d = 2R/c \quad (6.9)$$

The signal $S_{M2}$ can therefore be written as

$$S_{M2}(t_0) = A_{M2} \cos \left( 2\pi \left[ f_2 \left( t_o - \frac{2R}{c} \right) + \frac{k}{2} \left( t_o - \frac{2R}{c} \right)^2 \right] + \phi_{M2} \right) \quad (6.10)$$

Combining the probe measurement signal $S_{M2}$ with $S_{M1}$ in the measurement detectors $(M_{D1}$ and $M_{D2}$), the following signal is obtained from the measurement detector pair.

$$S_M(t_0) \propto A_{M1}A_{M2} \cos \left( 2\pi \left[ f_{IF}t_o - \frac{2f_2R}{c} + \frac{k}{2} \left( \frac{4R^2}{c^2} - \frac{4Rt_o}{c} \right) \right] + \phi_{M2} - \phi_{M1} \right) \quad (6.11)$$

Defining $\delta\phi_M = \phi_{M2} - \phi_{M1}$ and rearranging terms, some useful observations are possible. By grouping all terms in $t_0$, the measurement signal can be written as

$$S_M(t_0) \propto A_{M1}A_{M2} \cos \left( 2\pi \left[ \left( f_{IF} - \frac{2f_2R}{c} \right)t_o - \frac{2f_2R}{c} + \frac{k}{2} \left( \frac{4R^2}{c^2} \right) \right] + \delta\phi_M \right) \quad (6.12)$$

Equation (6.12) illustrates a well-known and important property of linearly-chirped systems that makes them highly desirable. If the Fourier transform is taken with respect to the time variable $t_0$, the frequency spectrum is shifted by an amount directly proportional to the sweep rate $k$ and the round-trip propagation time $2R/c$. So by measuring the frequency offset from the IF, which is 500 kHz in these experiments, the displacement can be measured directly from the spectral information. Although the reference signal $S_R$ is not required for the LIDAR mode, it does provide a stable reference for measuring the frequency shift. An intuitive explanation is also provided.
by the following illustration.

![Sawtooth FMCW waveform](image)

**Figure 6-3:** Sawtooth FMCW waveform for range measurement. The transmitted waveform is shown in red and the received waveform is shown in green.

In Figure 6-3, the red waveform represents the transmitted signal and the green waveform represents the received signal at $t_1$. Because the transmitted signal must transverse an additional delay $\Delta t$ with respect to the on-chip reference, the return signal will lag the on-chip reference in frequency by an amount $\Delta f$ proportional to $\Delta t$ and the linear slope $k$. Equivalently, a strictly time domain analysis using the relative phase shift between the reference signal and the measurement signal should yield the same result. The reference signal $S_R$ and measurement signal $S_M$ are given respectively by

$$S_R(t_0) \propto A_{R1}A_{R2}\cos(2\pi f_{IF}t_o + \delta \phi_R)$$  \hspace{1cm} (6.13)  

$$S_M(t_0) \propto A_{M1}A_{M2}\cos \left(2\pi \left[f_{IF}t_o - \frac{2f_2R}{c} + \frac{k}{2}\left(\frac{4R^2}{c^2} - \frac{4Rt_o}{c}\right)\right] + \delta \phi_M \right)$$  \hspace{1cm} (6.14)  

Therefore, the measurement signal undergoes a displacement-dependent phase shift with respect to the reference signal equal to

$$\Phi(R; t_o) = -\frac{2f_2R}{c} + \frac{k}{2}\left(\frac{4R^2}{c^2} - \frac{4Rt_o}{c}\right) = -\frac{2Rkt_o}{c} + \frac{2R}{c}\left[kR - f_2\right]$$  \hspace{1cm} (6.15)

The phase in Eq. (6.15) can be grouped into two terms, one linear in time and the
second at a fixed offset that depends on the distance R, sweep rate k, and the initial frequency $f_2$. Again, it is worth noting that the benefit of using the integrated heterodyne interferometer as a LIDAR is that with the proper signal processing it is possible to measure the displacement R using either the time (phase) or Fourier (frequency offset) domains. The choice of domain is closely tied to the resolution limits on the lower end, and to the phase/frequency ambiguity on the upper end. As was shown previously, both the phase and frequency shifts are both directly proportional to $kR/c$. The integrated interferometer determines the phase using an IQ demodulation scheme, which not only allows the phase magnitude to be derived, but also the direction of motion. Due to the speed of light in the denominator of Eq. (6.15), large values of k are required to achieve sufficient phase or frequency shifts when measuring small values of R. The minimum measurable phase shift of an interferometer is dictated primarily by its phase noise, whereas the minimum frequency resolution of a Fourier transform is inversely proportional to the measurement (resolution) time. The maximum frequency offset is limited by the desire to avoid aliasing, which is related to the sampling frequency and the intermediate frequency. But regardless of whether phase or frequency is used, the maximum measurement range is always limited by the laser coherence length.

### 6.1.1 LIDAR Resolution

When characterizing the performance of the interferometer, one of the key performance metrics was the displacement resolution, defined as the RMS position noise. In pulsed radar systems, the range resolution is the ability to distinguish between two or more targets at different ranges but the same bearing. While there are several factors that influence the range resolution, such as the target radar cross section, the primary factor limiting range resolution is the pulse width $\tau$. The range resolution is related to the pulse width by the well-known equation

$$\Delta R \geq \frac{c\tau}{2}$$  \hspace{1cm} (6.16)
This equation implies that to avoid ambiguity, the targets must be separated by a distance greater than or equal to half the spatial width of the pulse, since the full pulse width corresponds to the full round trip. For example, a temporal pulse width of 1 $\mu$s is equivalent to a spatial width of 300 m. So the radar returns from two targets separated by more than 150 m will be discernible as two individual returns. For separations less than 150 m, the target returns will overlap to create one long pulse. In an FMCW LIDAR, measurements are typically made in the frequency domain. Because of the finite FFT measurement duration $T$, what is actually seen in the frequency domain is the convolution of a sinusoid at the distance-dependent beat frequency and a sinc function with its first null at a frequency equal to $1/T$. Correspondingly, the width of the main lobe is also inversely proportional to the FFT resolution time. One criterion for determining the resolution of two closely-spaced targets from the FFT is to measure the minimum distance where the two main lobes are just resolvable. This is similar to the Rayleigh criterion for determining the minimum resolvable detail in a diffraction-limited optical system. The Rayleigh criterion, illustrated in Figure 6-4, essentially defines the resolution limit as the separation where the first diffraction minimum (null) of a point source coincides with the maximum of the adjacent point source.

![Figure 6-4: Illustration of the Rayleigh criterion.](image)

For an arbitrary temporal measurement duration $T$, the first zero of the FFT, assuming no other windowing function is applied to the data, is $f = 1/T$. So to be minimally resolvable, two objects should be separated by a frequency offset such that $\Delta f \geq 1/T$. The minimum resolvable distance can be related to the minimum
frequency offset via
\[ \Delta f = \frac{2k\Delta R}{c} = \frac{2BW\Delta R}{cT_{sweep}} \] (6.17)

Here, the frequency sweep rate k is written as the ratio of the swept bandwidth (BW) and the sweep period \( T_{sweep} \). So the minimum resolvable distance can be written as

\[ \Delta R = \frac{c\Delta fT_{sweep}}{2BW} = \frac{cT_{sweep}}{2BW \cdot T} \] (6.18)

where \( \Delta f \), the minimum FFT frequency resolution, is replaced by the inverse of the acquisition (resolution) time \( 1/T \). Lastly, the FFT acquisition time \( T \) typically corresponds to the sweep duration \( T_{sweep} \) to avoid spectral leakage that could arise from computing the FFT over two or more sweep periods. Applying this assumption, the minimum displacement resolution is

\[ \Delta R = \frac{c}{2BW} \] (6.19)

When the FFT resolution time is equal to the sweep period, the minimum resolution is inversely proportional to the swept bandwidth. These expressions will be useful when analyzing the results of the diode laser experiments.

### 6.2 Results

#### 6.2.1 Ranging using an Agilent Tunable Laser

In this section ranging measurements are presented for three distances [10 in (0.254 m), 18 in (0.457 m), and 26 in (0.66 m)] using a linearly-swept Agilent 81600B tunable laser. The wavelength is swept from 1.54 \( \mu m \) to 1.55 \( \mu m \) at a rate of 5 nm/s, which corresponds to a frequency sweep rate k of approximately 630 GHz/s. By examining the Fourier spectrum for each measurement, the displacement is determined using the frequency offset from the 500 kHz IF. The same measurement is repeated using phase.

Figure 6-6 shows the spectrum of the measurement signal \( S_M \) for the three dis-
tances, with the resultant offsets from the 500 kHz reference. Table 6.1 shows the result of utilizing these offsets to compute the displacement. By determining the spectral offsets for the three positions, the distance to the CCR is measured to within a few percent of the true position. However, because the position of the CCR on the optical table was roughly determined by eye and a ruler, most of the error could be attributed to inaccuracies in target placement. For comparison, results for a 1.26 THz/s sweep (10 nm/s) are shown in Figure 6-7 for the same distances, with a summary of the results shown in Table 6.2.

![Offset vs. Distance for 630 GHz/s Frequency Sweep](image)

Figure 6-6: Offset versus Distance for a 630 GHz/s frequency sweep.
<table>
<thead>
<tr>
<th>Sweep Rate</th>
<th>Distance inches (m)</th>
<th>Measured Shift (Hz)</th>
<th>Theoretical Shift (Hz)</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>630 GHz/s</td>
<td>10 (.254)</td>
<td>1,010</td>
<td>1,058</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>18 (.457)</td>
<td>1,856</td>
<td>1,905</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>26 (.660)</td>
<td>2,686</td>
<td>2,752</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Table 6.1: Summary of measurement results for a 630 GHz/s sweep rate.

<table>
<thead>
<tr>
<th>Sweep Rate</th>
<th>Distance inches (m)</th>
<th>Measured Shift (Hz)</th>
<th>Theoretical Shift (Hz)</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.26 THz/s</td>
<td>10 (.254)</td>
<td>2,027</td>
<td>2,117</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>18 (.457)</td>
<td>3,724</td>
<td>3,810</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>26 (.660)</td>
<td>5,398</td>
<td>5,503</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Table 6.2: Summary of measurement results for a 1.26 THz/s sweep rate.

Comparing the results of the 630 GHz/s and 1.26 THz/s sweeps, the offsets for
the higher sweep rate are almost exactly twice those of the lower sweep rate at each
distance, which is good considering that the higher sweep rate is twice the lower rate.
Based on previous discussions of the range resolution, the theoretically achievable
resolution for a 630 GHz/s and 1.26 THz/s sweep is approximately 240 microns and
120 microns, respectively. In reality, the resolution may be limited by the linearity of
the linear sweep. If the sweep is nonlinear, the beat frequency will shift during the
sweep, even if the target is perfectly still. So the longer the FFT measurement time,
the more the main lobe is broadened.

Lastly, the above analysis is repeated using phase, which is the native measure-
ment of the integrated heterodyne interferometer. However, without some form of
modulation or wavelength diversity, the maximum displacement is limited to a dis-
tance equal to half the laser wavelength. For operation at 1.55 \( \mu m \), this would limit
the maximum displacement to 775 nm. Employing a linearly-swept laser, the ambigu-
ous length becomes an inverse function of \( k \), so again there is a choice between the
minimum measurable displacement and the maximum unambiguous displacement.
For a one-way displacement \( R \), the total phase shift as a function of the rate \( k \) is
given by

\[
\Phi (R; k) = \int_{\phi_0}^{\phi_1} d\phi = -4\pi R \int_{\lambda_0}^{\lambda_1} \frac{d\lambda}{\lambda^2} = \frac{4\pi R}{\lambda_0 \lambda_1} \left[ \frac{\lambda_0}{2} - \frac{\lambda_1}{2} \right] = \frac{4\pi R}{c/kt_{\text{sweep}}} \quad (6.20)
\]

This is identical to the time-dependent phase shift obtained for the chirped signal.
Defining an equivalent wavelength \( \Lambda_{EQ} = c/kt_{\text{sweep}} \), the total phase excursion can be
written as

\[
\Phi (R; k; t) = \frac{4\pi R}{\Lambda_{EQ}} \quad (6.21)
\]

The equivalent wavelength determines the displacement at which the phase becomes
ambiguous, and occurs when the one way displacement \( R \) is equal to \( \Lambda_{EQ}/2 \). This
yields

\[
R_{\text{Amb,Chirped}} = \frac{c}{2kt_{\text{sweep}}} \quad (6.22)
\]
For comparison, the unambiguous length for a single frequency $f_2$ is

$$R_{\text{Amb, Unmodulated}} = \frac{c}{2f_2} \quad (6.23)$$

As the swept bandwidth $kt_{\text{sweep}}$ can be several orders of magnitude smaller than the optical frequency (193 THz), the unambiguous range can be extended significantly. Assuming a one-second sweep, the ambiguous range is extended from 775 nm to 238 $\mu$m. But this is still far less than the distances required of a practical LIDAR system. But if the phase is measured frequently enough to capture the $2\pi$ phase changes, any displacement can theoretically be measured. Returning to the previous expression for the phase, the relative phase shift over the sweep period is

$$\frac{\Phi (R; (t_0 + t_{\text{sweep}}))}{2\pi} = \frac{-2Rk(t_0 + t_{\text{sweep}})}{c} + \frac{2R}{c} \left[ \frac{kR}{c} - f_2 \right] \quad (6.24)$$

The phase is linear in time, with an offset determined by $R$, $k$, and $f_2$. Therefore, the phase change over the linear sweep period is

$$\frac{\Phi (R; (t_0 + t_{\text{sweep}})) - \Phi (R; t_0)}{2\pi} = \frac{\delta\Phi}{2\pi} = \frac{-2Rkt_{\text{sweep}}}{c} \quad (6.25)$$

Taking the slope of the linear phase shift over some portion of the sweep, the distance can be determined via

$$R = \frac{-\delta\Phi}{4\pi kt_{\text{sweep}}/c} \quad (6.26)$$

Because the measurements occur frequently enough over the sweep time, the multiple phase ambiguities can be unwrapped. This is analogous to sampling above twice the Nyquist rate to avoid aliasing in the frequency domain. Figure 6-8 shows the phase measurements for the three distances used earlier as the laser frequency is swept at 630 GHz/s for a period of approximately one-second. Table 6.3 summarizes the results of range measurements using the phase. The distances obtained using phase were compared to those measured with Fourier analysis, since these values should be self-consistent. The distances computed with phase and frequency are in good
agreement, although the error is much larger for the 10 inch case. The results show that the interferometer can be utilized as a long-baseline LIDAR to measure distances significantly larger than the laser’s ambiguous length, limited only by the coherence length of the laser and receiver sensitivity, using either phase or frequency.

Figure 6-8: Phase Shift versus Sweep Time for a 630 GHz/s sweep rate.

<table>
<thead>
<tr>
<th>Sweep Rate</th>
<th>Distance inches (m)</th>
<th>Calculated R(m) (Phase)</th>
<th>Calculated R(m) (Frequency)</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>630 GHz/s</td>
<td>10 (.254)</td>
<td>0.223</td>
<td>0.240</td>
<td>-7.1</td>
</tr>
<tr>
<td></td>
<td>18 (.457)</td>
<td>0.441</td>
<td>0.442</td>
<td>-0.23</td>
</tr>
<tr>
<td></td>
<td>26 (.660)</td>
<td>0.636</td>
<td>0.640</td>
<td>-0.63</td>
</tr>
</tbody>
</table>

Table 6.3: Results of displacement measurement using phase shift for a 630 GHz/s sweep rate.
6.2.2 Ranging using a Compact Diode Laser

The previous section demonstrated that the integrated heterodyne interferometer can function as a FMCW LIDAR, at least over the half-meter range tested using the Agilent tunable laser. However, LIDAR systems are typically used to measure distances ranging from meters to a few kilometers, depending on the application and the laser coherence length. In this section, additional experiments are conducted over the longer ranges typical of autonomous vehicles. Unfortunately, the size of the Agilent tunable laser does not lend itself to consumer applications such as autonomous vehicles. A more compact solution is to use a diode laser such as that shown in Figure 6-9.

Figure 6-9: A compact swept diode laser for LIDAR measurements.

Compared to the Agilent, the diode laser is physically compact and the wavelength can be tuned by adjusting either the temperature via the thermo-optic effect, or by varying the laser’s injection current. Both methods modulate the effective index of refraction of the laser cavity, which has the effect of modulating the effective cavity length and hence modulating the laser wavelength. The thermo-optic effect exploits the temperature dependency of the refractive index, whereas the injection current controls the index via the plasma-dispersion effect. The plasma-dispersion effect modifies the refractive index via the addiction or removal of charge carriers into or out of the conduction band via the injection current. Of the two effects, the frequency sweep achievable via the thermo-optic effect is usually the larger. But while
the wavelength tuning range using current modulation is usually on the order of 1-2 nm, the current can be swept relatively fast, on the order of a few MHz, depending on the specific laser. This makes it possible to achieve large sweep rates in a much smaller footprint. The optical frequency versus injection current for the diode laser is shown in Figure 6-10.

Figure 6-10: Diode laser frequency versus injection current.

Figure 6-11 shows diode laser ranging results at distances from 10 inches to 25 inches (0.25 m to 0.64 m) over two acquisition periods (resolution times). The longer acquisition time, shown on the right, corresponds to a narrower FFT main lobe. As discussed in Section 6.1.1, the main lobe width, and hence the range resolution, is inversely proportional to the resolution time as expected. Plotting beat frequency versus range yields the results shown in Figure 6-12. The plots shows a good linear fit between distance and the beat frequency, with a slope of 14 kHz/m. This corresponds to a frequency sweep rate of 2.1 THz/s and a bandwidth of approximately 20 GHz when sweeping using a 50 Hz triangular wave. The theoretical resolution for this bandwidth is approximately 7.14 mm (0.28 in), assuming the entire data window is used to compute the FFT.
Figure 6-11: Diode laser ranging results from 10 to 25 inches over two acquisition periods. The longer acquisition time shows a narrower main lobe width.

Figure 6-13 shows a number of distance measurements ranging from 9.6 inches to 24.6 inches using a short-range track. The full-width at half-maximum (FWHM) serves as a resolution metric in many scientific measurements, or to specify the angular width of antenna patterns, for example. Since the -3dB point is easily identifiable in a Fourier transform, is there a reason not to use the FWHM to characterize the resolution versus the Rayleigh criterion? Applying FWHM to this data over the full measurement set yields a resolution of approximately one-half inch. The FFT peaks are still clearly discernible at these separations, so why can’t an even better value for the resolution be specified? The fallacy in how the data in Figure 6-13 was interpreted is that the data is comprised of measurements taken sequentially as the target is moved, such that the frequency spectra corresponding to two closely-spaced targets are created at different temporal instants. So the spectra are distinct and can be overlaid after processing. In reality, there is always a trade-off between the FFT spectral resolution and temporal resolution. If finer spectral resolution is desired, the resolution (measurement) time must be increased. The compromise is that the FFT will provide spectral information for all frequencies present in the data during the measurement period T, and it is impossible to determine exactly when
various frequencies occur in time. If the resolution time were shortened to determine when individual frequency events occur temporally, the frequency resolution would be degraded.

Figure 6-13: Ranging results demonstrating a half-inch FWHM resolution.
If separate returns are incident from two closely-spaced objects within the resolution time $T$, their FFT spectra will add via superposition, as the Fourier transform is a linear operation on the overlapping returns. These are the conditions in which a real LIDAR must operate. The distance between two FFT spectral peaks that just touch at their -3dB points has a FWHM of approximately $0.886/T$, whereas the distance given by the Rayleigh criterion occurs at $1/T$, where $T$ is the resolution time. Figure 6-14 demonstrates the resultant spectra when two FFT returns are separated by the FWHM and the Rayleigh criterion, respectively. For range separations corresponding to frequency offsets at the FWHM, the combined spectral peaks merge into essentially one continuous peak. For distances corresponding to frequency offsets obeying the Rayleigh criterion, the individual peaks are still discernible in the combined spectrum.

Figure 6-14: Illustration of frequency resolution based on FWHM and Rayleigh criterion.
6.2.3 Doppler Frequency Measurements

Range measurements constitute an important function of LIDAR and RADAR systems. But to most, the measurement of velocity via the Doppler frequency is the most familiar. Whether it is police speed measurements utilizing LIDAR, or measuring the speed of storms using RADAR, the measurement of velocity is an important modality for both systems. As discussed in Chapter 2, the frequency shift observed between the source and a target whose velocity vector makes an angle $\theta$ with the line of sight between the two is

$$f_d = \pm \frac{2v_t \cos(\theta)}{c} f_c \text{ (Hz)}$$

(6.27)

where $v_t$ is the velocity of the target in m/s, $\theta$ is the angle between the target’s velocity vector and the line of sight, $f_c$ is the frequency of the source in Hz, and $c$ is the speed of light in vacuum. The factor of two accounts for the round-trip to the target and back for a monostatic (single antenna) system. The truth is that the Doppler frequency is the more fundamental and easier measurement to make in a CW system. As discussed earlier in this chapter, the measurement of range requires the waveform to carry some timing information. In a pulsed system, the rising and falling edges demarcate the pulse boundaries. In a FMCW system, the frequency modulation provides the timing structure.

The measurement of the Doppler frequency shift requires no such timing information from the waveform, and can be determined simply using the unmodulated CW carrier. In a homodyne system, where the reference and measurement arm signals consist of equal frequency optical carriers, the resultant signal from the measurement detectors has a beat frequency of zero when the target is at rest. When the target is in motion, a beat frequency is generated equal to the Doppler frequency. But because the Doppler frequency can have a positive or negative sign as determined by the direction of relative motion, the homodyne system is unable to resolve the direction because of the symmetry about baseband ($f = 0$) of positive and negative frequencies. The heterodyne system overcomes this limitation, since the Doppler frequency shift occurs on top of the intermediate frequency. So the measured frequency
is $f = f_{IF} + f_D$ for positive frequency shifts and $f = f_{IF} - f_D$ for negative frequency shifts. Because the IF is known by design, the direction of motion can be unambiguously determined. Figure 6-15 is a plot of the Doppler frequency shift and the Doppler error versus target speed. The data was taken on a 1.5 m AeroTech Pro225-HS linear stage, shown in Figure 6-16, that is capable of a maximum velocity of 300 mm/s and an accuracy of $\pm 35.5 \mu m$. The results show excellent agreement with the theoretical Doppler shift with a maximum error of approximately 2 mm/s over the entire velocity range of the stage. This translates into an average relative error of less than 1%.

![Figure 6-15: Doppler frequency shift and Doppler error versus target velocity](image)

### 6.2.4 Simultaneous Ranging and Doppler Measurements using a Triangular Waveform

Thus far the utility of the integrated heterodyne interferometer when operating as a LIDAR has been the focus of this chapter. In particular, it was possible to measure either the range or velocity, but not both simultaneously. For a police LIDAR, only velocity is necessary for determining whether motorists are speeding, and a simple CW waveform is sufficient. For a LIDAR altimeter, range measurement is important and a pulsed or FMCW waveform must be utilized. An application that may require simultaneous range and velocity measurement is the adaptive cruise control system found in many modern automobiles. However, the application that stands to make the greatest use of compact LIDAR systems is autonomous vehicles. One method to
obtain both range and Doppler frequency is to make a range measurement using a FMCW waveform, then take a separate measurement of the Doppler frequency using a CW waveform. Another alternative is to take several sequential range measurements and then compute the velocity from the time rate of change. Both methods are valid, but taking multiple measurements is perhaps not the most efficient use of resources.

A better method is to measure both quantities simultaneously using a FMCW triangular waveform, shown in Figure 6-17. The triangular waveform is similar to the sawtooth waveform of Figure 6-3 but is symmetric about the maximum. When there is no Doppler shift, the IF frequency is equal for the upward and downward portions
of the sweep. Technically, the IF for the upward and downward portions have the opposite sign, which can be seen from the relative position of the red (transmitted) and green (received) waveforms during the upward and downward sweep. Because of the frequency symmetry about the origin, $f_{IF} = -f_{IF}$. Now consider the case of a positive Doppler shift, although the same result applies for a negative Doppler shift. For the upward and downward sweeps, the total frequency shifts are given by

$$f_{Up} = f_{IF} + f_D$$

$$f_{Down} = -f_{IF} + f_D$$

![Image by Christian Wolf. Used under GNU Free Documentation License](image)

Figure 6-17: A triangular FMCW waveform for simultaneous measurement of range and velocity.

Keeping in mind that the negative frequency is mirrored to positive frequencies, the frequency of the upward and downward chirps can be rewritten as

$$f_{Up} = f_{IF} + f_D$$

$$f_{Down} = f_{IF} - f_D$$

This can be seen in Figure 6-17 by examining $\Delta f_1$ and $\Delta f_2$, which correspond to $f_{Up}$ and $f_{Down}$, respectively. Equations (6.30) and (6.31) describe a system of two
equations and two unknowns that can be solved if the frequency offsets due to the upward and downward chirps are known. The Doppler frequency can then be solved from

\[ f_D = \frac{f_{Up} - f_{Down}}{2} \]  

(6.32)

Similarly, the range can be determined from the intermediate frequency introduced by the upward and downward chirps in the absence of a Doppler frequency shift.

\[ f_{IF} = \frac{f_{Up} + f_{Down}}{2} \]  

(6.33)

Figure 6-18 clearly shows the higher and lower frequencies that occur during the positive and negative portions of the sweep as a result of a moving target, demonstrating that it is indeed possible to simultaneously measure both the range and velocity. A further benefit of using triangular modulation is that the direction of the Doppler shift can be determined even if the system is operated in the homodyne mode, which is not true of CW operation. Figure 6-19 represents a large number of range-Doppler measurement sets taken with the 1.5 m track using a triangular waveform. The measurements were taken at four speeds (75, 150, 225, and 300 mm/s) over distances from 0.4 m to 2 m. Part (a) shows a series of 120 measurements over the 1.5 m travel of the long-range stage for each of the four velocity values. Each period of the triangular sweep provides two frequencies, representing the upward and downward chirp regions of Figure 6-17 and illustrated in the time domain in Figure 6-18. Region 1 is the down-shifted IF due to the decreasing sweep and Region 2 is the up-shifted IF due to the positive sweep. The average of the two frequencies is shown in red. Part (b) is the distance measured from the average frequencies in Region 1 and Region 2, and Part (c) shows the velocities computed from the differences between those frequencies.

Figure 6-20 summarizes the velocity measurements for the four cases, showing the measured velocities compared to the commanded velocities. Figure 6-21 summarizes the range and speed measurement errors over the measurement set. The average relative distance error was 0.39% and the average speed error was 1.93%. This demonstrates the ability of this device, using simple triangular modulation, to
Figure 6-18: Upward and downward frequency chirps taken of a moving target using the LIDAR mode.

Figure 6-19: Simultaneous ranging and velocity measurements using a triangular waveform.
simultaneously measure distance and velocity with low single-digit error. The primary impediments are the accuracy of the linear track and the nonlinearity of the diode laser sweep, the latter causing a time-varying IF.

Figure 6-20: Measured versus commanded velocity using a triangular waveform.

Figure 6-21: Distance and velocity measurement errors for 75 mm/s, 150 mm/s, 225 mm/s, and 300 mm/s.
6.3 Conclusion

In this chapter the operation of the integrated heterodyne interferometer as a LIDAR system was demonstrated. The flexibility of this mode to measure either range, velocity, or both simultaneously with excellent accuracy was also shown. This versatility of operation bodes well for the interferometer’s use in a large number of scientific and consumer applications including autonomous vehicles, altimetry, geographic survey, atmospheric physics, and long-baseline distance measurement.
Chapter 7

Conclusions and Future Work

This thesis has explored the design, fabrication, and characterization of the first chip-scale heterodyne interferometer on silicon that incorporates integrated splitters, modulators, and germanium detectors in a 1 mm by 6 mm footprint. The interferometer was capable of obtaining a noise-limited position resolution of approximately 2 nm when measuring relative displacement. This compares favorably with commercial interferometers such as the Keysight Technologies 10702A linear interferometer, which obtains a 10 nm resolution at a wavelength of 632 nm. Compared to conventional heterodyne interferometers based on bulk optical components, this chip-scale device offers significant reductions in size and potentially cost, as well as increased stability due to the CMOS-compatible silicon photonic integration. This could create opportunities for applications in 3-D inspection, photolithography, hand-held devices, and portable biomedical units. The utility of the integrated heterodyne interferometer as a Laser Doppler vibrometer (LDV) was also explored. This design has the potential to facilitate a number of existing applications such as Atomic Force Microscopy that benefit from LDVs, but where conventional bulk devices are too large and expensive; or in applications where it is necessary to measure the vibrational characteristics of a structure without mass loading it. Measurements of peak displacement amplitudes many times the ambiguous length by appropriate sampling of the harmonic phase were possible using coherent (IQ) demodulation. From these phase measurements, the instantaneous Doppler frequency and velocity were computed. Lastly, the LIDAR
operational mode was exploited to measure range, velocity, or both simultaneously with excellent accuracy, thus enabling scientific and consumer applications including autonomous vehicles, altimetry, geographic survey, and long-baseline distance measurement.

That such utility can be realized from such a compact device is a testament not only to the power of interferometric techniques, but also to the versatility that can be purposefully designed into photonic systems. But even at this level of integration, this demonstration has only scratched the surface of what is possible with this CMOS-compatible photonics platform. The availability of nitride layers means it is possible to integrate lasers directly with the interferometer on the same chip. Based on the results presented here, substantial improvements to this design have already been made. The spacing between the Mach-Zehnder arms of the single-sideband modulators (SSBMs) was reduced to approximately half the distance used in this device. This reduced the footprint of the overall interferometer to approximately 0.5 mm by 6 mm, improving upon a specification that was already considerably better than any existing chip-scale device. The RF signals driving the SSBMs were rerouted to minimize the frequency crosstalk, which should significantly reduce the phase nonlinearity described in Chapter 4. Signal lines carrying bias voltages to the thermo-optic heaters will have ground connections distinct from the RF lines and the photodetectors, which should further mitigate crosstalk.

The adiabatic splitters were replaced with a new design that shows more uniform splitting ratios across a wider bandwidth. This will increase the common-mode rejection of the interferometer, particularly in applications such as LIDAR or OCT where the laser frequency is swept over tens of nanometers, due to better cancellation of DC offset voltages and correlated noise in the detectors. This change will complement the use of higher-efficiency germanium detectors with measured responsivities near the theoretical maximum of approximately 1 A/W at 1550 nm [67], more than double the responsivity of the current detectors. One of the most difficult aspects of operating the interferometer is the edge-coupled design. While edge-coupling is widely used and fairly reliable, high precision is required to couple the reflected measurement signal
back into the input waveguide. It is often necessary to use an infrared camera to make this alignment, which is undesirable for a robust system designed for use in a range of applications. One way to overcome this limitation is the use of grating-based vertical couplers that act as nano-antennas, shown in Figure 7-1. These nano-antennas can capture reflected light over a broader range of angles than an edge-coupled design, particularly for non-specular reflection. For LIDAR applications where the relative phase is not as important, high-speed phased array antennas can be used that allow the transmitted and received beams to be quickly scanned to any desired angle. These phased arrays, designed specifically for LIDAR applications, have already been successfully demonstrated [51].

![Figure 7-1: Transmit and receive couplers at the output section of a vertically-coupled design variant.](image)

But perhaps one of the most exciting design changes is the direct 3-D integration of photonics circuits with CMOS electronics. Using the CMOS fabrication process, vertically-integrated photonic and electronic circuits have been successfully demonstrated on what is functionally a single chip. One of the electronic components that was validated as part of this effort was an integrated transimpedance amplifier (TIA) circuit. A TIA circuit designed specifically for integration with the heterodyne interferometer is shown in Figure 7-2 (a). Part (b) shows a GDS of the heterodyne interferometer 3-D integrated with the TIA. These TIAs have shown promising results in other 3-D integrated devices, and if they work successfully with the interferometer,
they will replace the current 6 inch by 6 inch TIA printed circuit board with a circuit less than 120 \( \mu m \) high by 50 \( \mu m \) wide. The integrated TIA will consume less power and yield better noise performance than what is possible with discrete components. Lower phase nonlinearities, higher detector responsivity, more balanced splitting ratios leading to improved common-mode rejection, and lower noise amplifiers should all work in concert to improve the already impressive performance obtained in this device.

Figure 7-2: (a) CMOS integrated TIA and (b) TIA and supporting circuitry combined with the integrated heterodyne interferometer.

Before the interferometer can be brought to bear against existing commercial products it must be properly packaged. But proper packaging is in itself a topic requiring considerable engineering design. Expensive and easily damaged signal probes that carry signals to and from the chip are cumbersome and difficult to use even un-
der the best laboratory conditions. Wire bonding the device to a chip carrier allows for increased flexibility in circuit layout and reduces the parasitic inductance and capacitance inherent with probes. Unfortunately, most commercial ceramic carriers are made for electronic circuits that can and must be hermetically sealed. To the contrary, the laser must be coupled into the chip via a fiber while simultaneously coupling the measurement signal on and off the chip without interference. Designing a suitable carrier and developing a proper packaging plan may require considerable development effort, but is a necessary step to realizing the full utility and potential of this platform.
Appendices
Appendix A

Interferometer Phase Nonlinearities

Numerous researchers have quantified the phase nonlinearities of heterodyne interferometers [9, 12, 56–58, 68–70]. In the simplest case, the nonlinearity in the measurement signal is assumed to be caused by the addition of two equal-frequency sinusoids. The addition of multiple sinusoids of the same frequency is governed by the Harmonic Addition Theorem, which is stated in Eqs. (A.1)-(A.3) for two sinusoids.

\[ S_{\text{Measurement}} = A \sin (2\pi ft + \phi_M + \phi(L)) + B' \sin (2\pi ft + \phi_{\text{Spur}}) \]
\[ = C \sin (2\pi ft + \delta) \]  
(A.1)

\[ C^2 = A^2 + B'^2 + 2AB' \cos (\phi_{\text{Spur}} - \phi_M - \phi(L)) \]  
(A.2)

\[ \tan \delta(L) = \frac{A \sin (\phi_M + \phi(L)) + B' \sin (\phi_{\text{Spur}})}{A \cos (\phi_M + \phi(L)) + B' \cos (\phi_{\text{Spur}})} \]
\[ = \frac{\sin (\phi_M + \phi(L)) + \frac{B'}{A} \sin (\phi_{\text{Spur}})}{\cos (\phi_M + \phi(L)) + \frac{B'}{A} \cos (\phi_{\text{Spur}})} \]  
(A.3)

In an ideal heterodyne interferometer, the two frequencies are completely separated as they traverse their respective arms, meeting again only at the photodetectors. In reality, elliptical polarization and non-orthogonality of the input laser and the finite extinction ratio of the polarizing components cause each arm to contain frequency
components intended for the other. This leads to a nonlinear relationship between
the measured phase and the displacement \[58\]. For the integrated heterodyne interferometer, the frequency mixing is due instead to RF crosstalk between the SSBMs, but the nonlinear phase effects are similar to those in bulk devices. Even as the achievable resolution of heterodyne interferometers has improved with better phase measuring techniques, the achievable accuracy is limited to a few nanometers by the phase nonlinearity. This nonlinearity does not affect repeatability and resolution, only accuracy. Forming the in-phase and quadrature components with the reference signal \( S_{\text{Reference}} = D \sin (2\pi ft + \phi_R) \) and low-pass filtering yields

\[
I = \frac{CD}{2} \cos(\delta - \phi_R) \\
Q = \frac{CD}{2} \sin(\delta - \phi_R)
\]  

(A.4)

Because the ratio of I to Q is of interest, the AC amplitudes are ignored and the cosine and sine are expanded to obtain

\[
I(L) = \cos(\delta(L)) \cos(\phi_R) + \sin(\delta(L)) \sin(\phi_R) \\
Q(L) = - \cos(\delta(L)) \sin(\phi_R) + \sin(\delta(L)) \cos(\phi_R)
\]  

(A.5)

If the angle \( \delta \) were linear, I and Q would describe a circle centered at the origin with unity semi-minor and semi-major axes. However, it is clear from Eq.(A.3) that the angle \( \delta \) is nonlinear, and this is at the heart of the dilemma. If the phase of Eq.(A.3) is written as

\[
\delta(L) = \arctan(x)
\]  

(A.6)

and substituted into Eq.(A.5), I and Q are given by

\[
I = \cos(\arctan(x)) \cos(\phi_R) + \sin(\arctan(x)) \sin(\phi_R) \\
Q = - \cos(\arctan(x)) \sin(\phi_R) + \sin(\arctan(x)) \cos(\phi_R)
\]  

(A.7)
Using the trigonometric identities
\[
\cos(\arctan(x)) = \frac{1}{\sqrt{1 + x^2}} \\
\sin(\arctan(x)) = \frac{x}{\sqrt{1 + x^2}}
\]  \hfill (A.8)

and applying Eqs. (A.3) and (A.8) to Eq. (A.7), revised expressions are obtained for the in-phase and quadrature components and are given by
\[
I(L) = \frac{\cos(\phi(L) + \phi_M) + \Gamma \cos(\phi_{Spur}) \cos(\phi_R) + [\sin(\phi(L) + \phi_M) + \Gamma \sin(\phi_{Spur})] \sin(\phi_R)}{\sqrt{1 + \Gamma^2 \cos(\phi(L) + \phi_M - \phi_R) + \Gamma^2}} \\
Q(L) = -\frac{\cos(\phi(L) + \phi_M) + \Gamma \cos(\phi_{Spur}) \sin(\phi_R) + [\sin(\phi(L) + \phi_M) + \Gamma \sin(\phi_{Spur})] \cos(\phi_R)}{\sqrt{1 + \Gamma^2 \cos(\phi(L) + \phi_M - \phi_R) + \Gamma^2}}
\]  \hfill (A.9)

Here \( \Gamma = B'/A \) and \( B' \) is the amplitude of the leakage term. Again, because the ratio of I to Q is of interest, the denominator of each term can be ignored. Finally, if the terms involving the sine and cosine of sums in Eq. (A.9) are expanded and the terms involving the sine and cosine of \( \phi_R \) are distributed, more useful equations for the in-phase and quadrature components are given by
\[
I(L) = \Gamma \cos(\phi_{Spur} - \phi_R) + \cos(\phi(L)) \cos(\phi) - \sin(\phi(L)) \sin(\phi) \\
Q(L) = -\Gamma \sin(\phi_{Spur} - \phi_R) + \cos(\phi(L)) \sin(\phi) + \sin(\phi(L)) \cos(\phi)
\]  \hfill (A.10)

The general equation for an ellipse in Cartesian coordinates centered at \((X_0, Y_0)\) with rotation angle \( \phi \) is
\[
X(L) = X_0 + a \cos(\phi(L)) \cos(\varphi) - b \sin(\phi(L)) \sin(\varphi) \\
Y(L) = Y_0 + a \cos(\phi(L)) \sin(\varphi) + b \sin(\phi(L)) \cos(\varphi)
\]  \hfill (A.11)

Therefore the ellipse described by Eq. (A.10) is one centered at \( X_0 = \Gamma \cos(\phi_{Spur} - \phi_R) \) and \( Y_0 = -\Gamma \sin(\phi_{Spur} - \phi_R) \), but with unity values for the semi-major and semi-minor axes. The terms involving \( \phi = \phi_M - \phi_R \), which is the initial phase difference between the measurement and reference arms, are assumed constant and
only contribute to a rotation of the ellipse. However, this shifting of the origin is indicative of a nonlinearity. If the uncorrected I and Q are plotted over at least one complete fringe (2\(\pi\) phase excursion), the Lissajous pattern should trace out an offset circle. But instead of an offset circle, the result is instead an offset ellipse with scaled values for the semi-minor and semi-major axes. This is inconsistent with the model that was just presented, which may be incomplete and too simplistic. Indeed, the initial assumption that the nonlinearity is due to a single leakage term that adds directly to the desired measurement signal is most surely incorrect.

In reality, the nonlinearity is due to the presence of both heterodyne frequency tones in each interferometer arm, a condition common in bulk interferometers. The presence of both tones in the reference arm allows a beat signal to exist in the measurement detectors even in the absence of a reflected measurement signal. If each arm consists of a desired frequency component and a leakage frequency component, the reference arm signal can be written as

\[
S_{\text{Reference}} = A \sin (2\pi f_1 t + \phi_A) + B' \sin (2\pi f_2 t + \phi_{B'})
\]  

(A.12)

and the measurement arm signal as

\[
S_{\text{Measurement}} = B \sin (2\pi f_2 t + \phi_B + \phi_2(L)) + A' \sin (2\pi f_1 t + \phi_{A'} + \phi_1(L))
\]  

(A.13)

The prime terms denote the undesired frequency component in each arm. The length-dependent phase measured by both frequencies of the measurement arm are also assumed equal, i.e. \(\phi_1(L) = \phi_2(L) = \phi(L)\). This is reasonable, as the heterodyne frequencies differ by only 500 kHz, approximately 3-billionths the value of the unmodulated frequency. Because the initial phase terms are only important to the origin and rotation angle, as shown previously, they are ignored going forward for the sake of simplicity. Considering only the in-band terms at the intermediate frequency, the
beat signal from the measurement detectors is

\[ S_{\text{Meas,Det}} \propto (AB' + A'B) \sin (2\pi f_{IF}t) \]
\[ + A'B' \sin (2\pi f_{IF}t - \phi(L)) \]
\[ + AB \sin (2\pi f_{IF}t + \phi(L)) \]  \hspace{1cm} (A.14)

Ignoring the reference signal amplitude, the in-phase and quadrature terms can be written respectively as

\[ I(L) = (AB' + A'B) + (AB + A'B') \cos \phi(L) \]
\[ Q(L) = (AB - A'B') \sin \phi(L) \]  \hspace{1cm} (A.15)

This ellipse has a semi-major axis proportional to \((AB + A'B')\) and semi-minor axis proportional to \((AB - A'B')\), in addition to the origin offset. In the absence of the leakage terms, \((A' = B' = 0)\), the in-phase and quadrature terms reduce to that of a circle centered at the origin. Using MATLAB to find the best fit ellipse, the parameters thus obtained \((a, b, \varphi)\) can be used to circularize the ellipse according to

\[ I'(L) = b \left[ (I(L) - I_0) \cos(\varphi) + (Q(L) - Q_0) \sin(\varphi) \right] \]
\[ Q'(L) = a \left[ -(I(L) - I_0) \sin(\varphi) + (Q(L) - Q_0) \cos(\varphi) \right] \]  \hspace{1cm} (A.16)

where \((I_0, Q_0)\) are the \((x,y)\) origin coordinates of the uncorrected ellipse. The expressions in Eq. (A.16) are the corrected in-phase and quadrature components calculated using a least-squares fit to the uncorrected ellipse.
Appendix B

Transimpedance Gain, Stability, and AC Coupling

B.1 Noise Gain and Stability

Even if there were ideal conversion in the SSBMs, the beating of the heterodyne frequencies will produce DC terms in the electrical output. If there is a significant unsuppressed carrier in the SSBMs, these terms will produce additional DC power in the electrical output in the form of homodyne mixing. These DC components can cause significant problems at the op-amp, particularly if the AC beat term produced by the heterodyne frequencies is small. The output of the transimpedance amplifiers (TIAs) is limited to within a few hundred millivolts of the power rails. For the OPA656, this is a maximum voltage difference of approximately 4.4 V. The heterodyne measurement current can be small, requiring a fairly large transimpedance gain to obtain a reasonable output voltage. Unfortunately, the large transimpedance gain also amplifies the DC signal, which can saturate the TIA and significantly reduce its dynamic range. In this appendix, the TIA performance is analyzed to better understand its gain and stability characteristics.

The TIA circuit is operated in the inverting feedback configuration shown in Figure B-1. The impedance of the feedback circuit, consisting of a parallel combination
of the transimpedance gain $R_F$ and a feedback capacitor $C_F$, is given by

$$\frac{1}{Z_F(\omega)} = \frac{1}{R_F} + j\omega C_F = \frac{1}{R_F} + j\omega R_F C_F$$

$$Z_F(\omega) = \frac{R_F}{1 + j\omega R_F C_F}$$

(B.1)

![Operational Amplifier TIA Circuit](image)

Figure B-1: An operational amplifier TIA circuit in the inverting configuration.

The transimpedance amplifier operated in the inverting configuration is defined by two gains, the signal gain $S_G$ and the noise gain $N_G$ as defined below.

$$N_G = 1 + \frac{Z_F}{Z_G}$$

(B.2)

$$S_G = -\frac{Z_F}{Z_G}$$

(B.3)

The signal gain is the gain seen between the input current or voltage source and the output voltage. For the TIA, the signal gain is the transimpedance gain between the input current and the output voltage. Contrary to normal instincts regarding an operational amplifier operated in feedback, the stability of the feedback circuit is determined not by the signal gain $S_G$, but rather by the noise gain $N_G$. The noise gain is the gain seen by the operational amplifier’s noise voltage $v_n$ and noise current $i_n$. The current noise $i_n$ is usually denoted by two current noise sources $i_{nn}$ and $i_{np}$.
due to shot noise generated by the transistor bias currents at the inverting and non-inverting inputs, respectively. These current noise sources are usually represented as originating from their respective inputs to ground. The voltage noise $v_n$ is used to model other internal noise sources that are not caused by the bias currents. In low source impedance op-amps such as the OPA656 that utilize a JFET front-end, the output voltage produced by the noise currents is usually insignificant compared to the internal voltage noise $v_n$ and can typically be ignored. A noisy op-amp is modeled in Figure B-2 as a noiseless operational amplifier combined with the current and voltage noise sources. The equivalent circuit for computing the noise gain, assuming only the voltage noise source, is shown in Figure B-3.

![Figure B-2: Noisy Op Amp Model from TI’s Op Amps for Everyone](image_url)
B.1.1 AC Coupling with a Shunt Inductor and Series Capacitor

This section will examine the impedance $Z_G$ for the case of AC coupling as implemented using the combination of an inductor and a capacitor. The diode capacitance, which is in the femto-Farad range, and the diode parallel resistance, which is in the mega-Ohm to giga-Ohm range, are ignored. The circuit for the LC coupled TIA is shown in Figure B-4.

Noise Gain

For the equivalent circuit in Figure B-5, the noise gain $Z_G$ can be written as

$$Z_G(\omega) = j\omega L_G + \frac{1}{j\omega C_G} = 1 - \frac{\omega^2 L_G C_G}{j\omega C_G} \quad (B.4)$$

Defining $K(\omega) = 1 - \omega^2 L_G C_G$, the noise gain can be rewritten as

$$N_G(\omega) = 1 + \frac{j\omega R_F C_G}{K(1 + j\omega R_F C_F)} = \frac{1 + j\omega R_F C_F \left(1 + \frac{C_G}{K C_F}\right)}{1 + j\omega R_F C_F} \quad (B.5)$$
The corresponding feedback factor $\beta(\omega)$ is the inverse of the noise gain.

$$\beta(\omega) = \frac{1}{N_G} = \frac{1 + j\omega R_F C_F}{1 + j\omega R_F C_F \left(1 + \frac{C_g K}{K C_F}\right)} \quad (B.6)$$

The magnitude and phase of the feedback factor $\beta$, when combined with the TIA open-loop gain $A_{OL}$, yields the closed-loop gain. The loop gain is defined as $A_{OL} \beta$ and the closed-loop gain is defined as $A_{CL} = \frac{A_{OL}}{1 + A_{OL} \beta}$. As was alluded to earlier, the $N_G$, and hence $\beta$, determine the stability of the TIA via its poles and zeros, along with those
of the open-loop gain $A_{OL}$. If $A_{OL}\beta \gg 1$, which is typically the case due to the large value of $A_{OL}$, then the closed-loop gain is approximately $A_{CL} = \frac{1}{\beta} = N_G$. Therefore, stability depends almost exclusively on the noise gain. It is then instructive to write the magnitude and phase of $\beta$, which determine the poles and zeros of the feedback system. To find the zeros and poles of $\beta$, it is more instructive to fully expand the above expression and to replace $j\omega$ with the Laplace notation $s = \sigma + j2\pi f = \sigma + j\omega$.

$$\beta(\omega) = \frac{1}{N_G} = \frac{s^3L_GC_GR_FC_F + s^2L_GC_G + sR_FC_F + 1}{s^3L_GC_GR_FC_F + s^2L_GC_G + sR_F[C_F + C_G] + 1} \quad (B.7)$$

Because the noise gain and the feedback factor are rational functions consisting of third-order polynomials in the numerator and denominator, both expressions consist of three zeros and three poles. From the relationship between the noise gain and the feedback factor, the poles and zeros of one quantity are the zeros and poles of the other, and are found from the roots of the numerator and denominator polynomial. Solving the above cubic equations symbolically is non-trivial, but can be accomplished via factoring by grouping.

$$|\beta_{Zero,1}(\omega)| = \sqrt{\frac{1 + (\omega R_FC_F)^2}{R_FC_F}}; \angle \beta_{Zero,1}(\omega) = \tan^{-1}\omega R_FC_F \quad (B.8)$$

The two remaining poles are complex-conjugate pairs that have equal magnitudes, but are out of phase by 180°.

$$|\beta_{Zero,2}(\omega)| = \sqrt{\frac{1}{L_GC_G} + \omega^2}; \angle \beta_{Zero,2}(\omega) = \frac{\pi}{2} \quad (B.9)$$

$$|\beta_{Zero,3}(\omega)| = \sqrt{\frac{1}{L_GC_G} + \omega^2}; \angle \beta_{Zero,3}(\omega) = -\frac{\pi}{2} \quad (B.10)$$

Determining a simple expression for the poles is complicated by the presence of the $R_F[C_F + C_G]$ term, which perturbs the symmetry of the cubic equation. However, if $C_G \ll C_F$, then the poles are essentially equal to the zeros and there is one-to-one pole-zero cancellation. But in many cases of practical interest, where $C_G$ is chosen to provide a specific AC response, it may be the case that $C_G \gg C_F$ and the poles occur
before the zeros. In that case stability must be determined by examining the closed-
loop gain. But it is still possible to write the magnitude and phase of the feedback
factor by determining the magnitude and phase of the numerator and denominator
separately.

\[
|\beta(\omega)| = \sqrt{\frac{1 + (\omega R_F C_F)^2}{1 + [\omega R_F C_F \left(1 + \frac{C_G}{K C_F}\right)]^2}}; \\
\angle\beta(\omega) = tan^{-1}\omega R_F C_F - tan^{-1}\omega R_F C_F \left(1 + \frac{C_G}{K C_F}\right) \tag{B.11}
\]

For \(A_{OL}\beta \gg 1\), the closed-loop gain is equal to the noise gain, which is the reciprocal
of \(\beta\). This means that the zeros of \(\beta\) become the poles of the closed-loop gain, and
the poles of \(\beta\) becomes the zeros of the closed-loop gain.

\[
|A_{CL}| = \frac{1}{|\beta|} = \sqrt{\frac{1 + (\omega R_F C_F)^2}{1 + (\omega R_F C_F)^2}}; \\
\angle A_{CL} = -tan^{-1}\omega R_F C_F + tan^{-1}\omega R_F C_F \left(1 + \frac{C_G}{K C_F}\right) \tag{B.12}
\]

The zero frequencies, where the numerator of the closed-loop magnitude vanishes, are
found by solving the quadratic equation resulting from setting the numerator to zero.
The pole frequencies, where the denominator of the closed-loop magnitude vanishes,
are given accordingly by

\[
F_P = \pm \frac{1}{2\pi R_F C_F} \tag{B.13}
\]

It is also important to remember that the pole and zero frequencies are actually
frequencies in the complex plane. In the notation of the Laplace Transform, the
complex frequency is given by \(s = \sigma + j2\pi f = \sigma + j\omega\). For \(C_F \gg C_G\), the above
relations are simplified accordingly to

\[
\beta = \frac{1}{1 + j\omega \frac{R_F C_G}{K}} \tag{B.14}
\]

\[
|\beta| = \frac{1}{\sqrt{1 + \left(\frac{\omega R_F C_G}{K}\right)^2}}; \angle\beta = -tan^{-1}\omega R_F C_G \frac{K}{K C_F} \tag{B.15}
\]
\[ |A_{CL}| = \frac{1}{|\beta|} = \sqrt{1 + \left(\frac{\omega R_F C_F}{K}\right)^2}; \angle A_{CL} = \tan^{-1} \omega \frac{R_F C_G}{K} \]  
\hspace{1cm} (B.16)

Signal (Transimpedance) Gain

The TIA noise gain and stability characteristics were discussed in the previous section, and it was shown that in the AC coupled configuration, the stability was a function of the feedback impedance, which includes \( R_F \) and \( C_F \), as well as the inductance \( L_G \) and capacitance \( C_G \) comprising the AC coupling section. Because the value of \( C_G \) is large with respect to the photodiode equivalent capacitance \( C_D \), \( C_D \) is ignored. The input capacitances of the TIA op-amps were also ignored, since they are much smaller than \( C_G \). But while these parameters must be chosen carefully to ensure the stability of the op-amp and to avoid oscillations, their effect on the signal gain must also be considered, which for the inverting configuration is

\[ S_G = -\frac{Z_F}{Z_G} \]  
\hspace{1cm} (B.17)

For the transimpedance amplifier, the transfer function of interest is given by the ratio of the output voltage to the diode photocurrent. For LC coupling, the circuit model for the signal gain is shown in Figure B-6.

![Figure B-6: Equivalent circuit for LC coupled signal gain.](image)

The photocurrent is in series with the parallel resonant circuit formed by the inductor
$L_G$ and the capacitor $L_G$, and whose impedance is given by

$$Z_G = \frac{j\omega L_G}{1 - \omega^2 L_G L_G} \quad (B.18)$$

The photocurrent creates a voltage drop $V = I_D Z_G$ between the biasing network and the virtual ground at the inverting input. The current in the capacitor $C_G$, which is also the current passing through the feedback impedance $Z_F$, is given by

$$I_F = \frac{V}{Z_{CG}} = \frac{I_D Z_G}{Z_{CG}} = \frac{I_D}{1 - \frac{1}{\omega^2 L_G C_G}} \quad (B.19)$$

The output voltage is then

$$V_0 = -I_F Z_F = -\frac{I_D}{1 - \frac{1}{\omega^2 L_G C_G}} \frac{R_F}{1 + j\omega R_F C_F} \quad (B.20)$$

If $K$ is defined as $K = \left(1 - \frac{1}{\omega^2 L_G C_G}\right)$, the transimpedance gain can be written as

$$S_G = \frac{-V_0}{I_D} = -\frac{R_F / K}{1 + j\omega R_F C_F} \quad (B.21)$$

So the signal gain depends not only on the feedback resistor $R_F$, but also on the reactive elements. The pole introduced in the transimpedance gain due to the feedback capacitor $C_F$, while key for maintaining stability, will also reduce the signal bandwidth. Finally, as $\omega$ becomes larger, the coupling inductor approaches an open circuit and the capacitor approaches a short circuit. In the limit as $\omega$ approaches infinity, or at frequencies well above resonance, $K$ approaches unity and the signal gain is not affected by the AC coupling.

### B.1.2 AC Coupling with a Shunt Resistor and Series Capacitor

LC coupling has the advantage that in theory, the capacitor and inductor are perfect reactive elements, and hence do not suffer from the thermal noise caused by the
resistor in a RC coupling circuit. The inductor also does not alter the photodiode DC bias voltage as a function of photocurrent. For an RC coupling circuit, the net bias voltage across the photodiode is a function of the applied DC bias voltage and the voltage across the resistor $R_G$. If not carefully controlled, the photodetector could go into forward bias for large values of the photocurrent and $R_G$. In reality, the reactive L and C elements form a resonant circuit which can cause gain peaking that leads to oscillations. One way to minimize the resonance is to insert a resistor in series with the inductor to spoil the Q of the resonance. But this essentially negates the primary advantage of the inductor, and still presents the problem of a photocurrent-dependent bias. In this section the stability parameters and transimpedance gain are derived for the RC coupled circuit shown in Figure B-7.

![Figure B-7: A TIA circuit with RC input coupling.](image)

### Noise Gain

To compute the noise gain the equivalent noise circuit in Figure B-8 is employed, and $Z_G$ for this circuit can be expressed as

$$Z_G(\omega) = R_G + \frac{1}{j\omega C_G} = 1 + \frac{j\omega R_G C_G}{j\omega C_G} \quad (B.22)$$

The expression for the noise gain can be expanded and the feedback factor written
Figure B-8: Equivalent circuit for RC coupled noise calculations.

according to

\[ N_G(\omega) = 1 + \frac{j\omega R_F C_G}{(1 + j\omega R_G C_G)(1 + j\omega R_F C_F)} = \frac{(1 + j\omega R_G C_G)(1 + j\omega R_F C_F) + j\omega R_F C_G}{(1 + j\omega R_G C_G)(1 + j\omega R_F C_F)} \]  

(B.23)

\[ \beta(\omega) = \frac{1}{N_G} = \frac{(1 + j\omega R_G C_G)(1 + j\omega R_F C_F)}{(1 + j\omega R_G C_G)(1 + j\omega R_F C_F) + j\omega R_F C_G} \]  

(B.24)

The roots of the numerator in Eq. (B.24) are the zeros of \( \beta \), and have magnitude and phase given by

\[ |\beta_{Zero,1}(\omega)| = \sqrt{1 + (\omega R_F C_F)^2}; \quad \angle\beta_{Zero,1}(\omega) = \tan^{-1} \omega R_F C_F \]  

(B.25)

\[ |\beta_{Zero,2}(\omega)| = \sqrt{1 + (\omega R_G C_G)^2}; \quad \angle\beta_{Zero,1}(\omega) = \tan^{-1} \omega R_G C_G \]  

(B.26)

As was the case for the LC coupled circuit, determining a simple expression for the poles is complicated by the presence of the \( R_F[C_F + C_G] \) terms, which perturb the symmetry of the cubic equation. However, if \( C_G \ll C_F \), then the poles are essentially equal to the zeros and there is one-to-one pole-zero cancellation. But in many cases of practical interest where \( C_G \) is chosen to provide a specific AC response, it may be the case that \( C_G \gg C_F \) and the poles occur before the zeros. In that case stability must
be determined by examining the closed-loop gain. But the magnitude and phase of the feedback factor can still be written by determining the magnitude and phase of the numerator and denominator separately, which gives

$$|\beta(\omega)| = \sqrt{\frac{[1 + (\omega R_G C_G)^2] \left[ 1 + (\omega R_F C_F)^2 \right]}{[1 - \omega^2 R_G C_G R_F C_F]^2 + [\omega (R_F [C_F + C_G] + R_G C_G)]^2}}$$  \hspace{1cm} (B.27)

$$\angle \beta(\omega) = \tan^{-1} \omega R_F C_F + \tan^{-1} \omega R_F C_F - \tan^{-1} \frac{\omega (R_F [C_F + C_G] + R_G C_G)}{1 - \omega^2 R_G C_G R_F C_F}$$  \hspace{1cm} (B.28)

For $A_{OL} \beta \gg 1$, the closed-loop gain is equal to the noise gain, which is the reciprocal of $\beta$. As was shown previously, the zeros of $\beta$ become the poles of the closed-loop gain, and the poles of $\beta$ becomes the zeros of the closed-loop gain.

$$|A_{CL}| \equiv \frac{1}{|\beta|} = \sqrt{\frac{[1 - \omega^2 R_G C_G R_F C_F]^2 + [\omega (R_F [C_F + C_G] + R_G C_G)]^2}{[1 + (\omega R_G C_G)^2] \left[ 1 + (\omega R_F C_F)^2 \right]}}$$  \hspace{1cm} (B.29)

$$\angle A_{CL} = \tan^{-1} \frac{\omega (R_F [C_F + C_G] + R_G C_G)}{1 - \omega^2 R_G C_G R_F C_F} - \tan^{-1} \omega R_F C_F - \tan^{-1} \omega R_G C_G$$  \hspace{1cm} (B.30)

The zero frequencies, where the numerator of the closed-loop magnitude vanishes, are found by solving the quadratic equation resulting from setting the numerator to zero. The pole frequencies, where the denominator of the closed-loop magnitude vanishes, can be determined by inspection and are given by

$$F_{P,1} = \frac{1}{2 \pi R_F C_F}$$  \hspace{1cm} (B.31)

$$F_{P,2} = \frac{1}{2 \pi R_G C_G}$$  \hspace{1cm} (B.32)

**Signal (Transimpedance) Gain**

The equivalent circuit for computing the signal gain is shown in Figure B-9. As shown for the LC case, the resistor $R_G$ and capacitor $C_G$ form a parallel resonant circuit in
series with the photocurrent, with an impedance given by

\[ Z_G = \frac{R_G}{1 + j\omega R_G C_G} \quad (B.33) \]

The photocurrent creates a voltage drop \( V = I_D Z_G \) between the biasing network and the virtual ground at the inverting input. The current through the capacitor \( C_G \), which is also the current through the feedback impedance \( Z_F \), is found according to

\[ I_F = \frac{V}{Z_C} = \frac{I_D}{1 + \frac{1}{j\omega R_G C_G}} \quad (B.34) \]

The output voltage is given by

\[ V_0 = -I_F Z_F = -\frac{I_D}{\left(1 + \frac{1}{j\omega R_G C_G}\right)} \frac{R_F}{\left(1 + j\omega R_F C_F\right)} \quad (B.35) \]

This yields a transimpedance gain equal to

\[ S_G = \frac{-V_0}{I_D} = -\frac{R_F}{\left(1 + \frac{1}{j\omega R_G C_G}\right) \left(1 + j\omega R_F C_F\right)} \]
\[ = -\frac{-j\omega R_F R_G C_G}{\left(1 + j\omega R_G C_G\right) \left(1 + j\omega R_F C_F\right)} \quad (B.36) \]
Appendix C

A Jones Matrix Analysis of a Forward and Reverse Pass through a Quarter Waveplate

In the analysis of the bulk heterodyne interferometer, a polarizing beam splitter (PBS) was utilized to separate two orthogonal polarizations. Polarization is typically used to isolate the reference and measurement paths, which are then recombined at the PBS before impinging on a photodetector. In Figure C-1, the two polarizations enter the PBS through Port 1 and are split into Port 2 and Port 3, depending on their polarization. Assuming the polarization state of the reference and measurement arms are not altered, they will reenter the PBS at Ports 2 and 3 after reflection. Due to the reciprocal nature of the PBS, the composite beam will exit the PBS at Port 1 where it entered, rather than the photodetector at Port 4.

This can be solved by rotating the polarization of both the reference and measurement paths by 90°, thereby allowing both paths to exit through Port 4 upon their recombination in the PBS. This can be done easily in a single pass using a half waveplate (HWP) or by cascading two quarter waveplates (QWP). However, the method employed must accomplish this 90° phase rotation in two passes through any such phase rotation: once when leaving the PBS outbound and again before reentering the PBS. The purpose of this section is to demonstrate that this operation, when using a
QWP, is equivalent to a single pass through a HWP. This simple analysis is carried out using the Jones matrix formulation, and it is assumed that both the reference and measurement paths see perfectly reflecting surfaces. Depending on the application this may not be true of the measurement path, however, this analysis assumes that polarization is maintained. Consider a QWP that has its fast and slow axes oriented at an arbitrary angle theta ($\theta$) with respect to a right-handed Cartesian reference frame shown in Figure C-2.

$$ R(\theta) = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \quad (C.1) $$
The Jones matrix describing the transformation of the input polarization vector in the \((x, y)\) frame by the QWP in the forward direction is given by

\[
J_{\text{Forward,Ref}}(\theta) = R(-\theta) J_{\text{QWP,Fast}} R(\theta)
\]

\[
= \begin{bmatrix}
\cos(\theta) & -\sin(\theta) \\
\sin(\theta) & \cos(\theta)
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
0 & j
\end{bmatrix}
\begin{bmatrix}
\cos(\theta) & \sin(\theta) \\
-\sin(\theta) & \cos(\theta)
\end{bmatrix}
\]

\[(\text{C.2})\]

In the Jones matrix formulation, the multiplications representing polarization devices take place from right to left in order of occurrence, rather than left to right. In Eq.(\text{C.2}), the first operation is a rotation of the input polarization vector from the \((x, y)\) reference frame into the \((f, s)\) frame of the QWP, represented by the rotation matrix \(R(\theta)\). This is followed by multiplication with the Jones matrix of a QWP with the fast axis oriented along the horizontal axis of the right-handed \((f, s)\) frame.

At this point the input polarization vector has been rotated into the local reference frame of the QWP, and a relative phase shift of \(\pi/2\) applied between the fast and slow axes. Lastly, the system is rotated back into the \((x, y)\) frame by applying the rotation matrix \(R(-\theta)\).

Next, consider the Jones matrix for reflection from a perfect mirror at normal incidence. When analyzing reflection and transmission at a material interface, the convention is to define a fixed coordinate system at the interface and then to analyze the incident, reflected, and transmitted field in that frame. However, when considering polarization there are some differences in convention with regards to how the coordinate systems are defined. Rather than employing the global coordinate system used in analyzing reflection and transmission at a boundary, polarization analysis is based on a coordinate system local to each wave.

One convention used in optics is to define the local z-axis of each wave in the direction of its propagation. The sense of the polarization is determined by looking at the wave from the front. In other words, the polarization is defined by looking at
the wave in the direction toward the source along the local negative z-axis. In some areas of electromagnetics, such as antenna theory, the polarization sense is usually defined by looking along the positive z-axis in a direction propagating away from the source. Based on these considerations, the coordinate systems for the incident wave and the reflected wave can be defined as in Figure C-3.

![Figure C-3: Local coordinate frames for the reflected and incident fields.](image)

With the positive $z'$-axis defined in the direction of propagation, reversing either the x or y axis with respect to the incident frame will ensure a right-handed coordinate system for the reflected field. Without loss of generality, the x-axis is reversed. The phase reversal required to satisfy the boundary conditions of the reflected waves at normal incidence must also be considered. Because both polarization components experience a $\pi$ phase shift, neglecting the negative sign has no net impact on the polarization. However, for the sake of completeness these components were negated when constructing the Jones matrix for the mirror. With these caveats in mind, the Jones matrix for the mirror at normal incidence can be written as:

$$J_{\text{Mirror}} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad \text{(C.3)}$$

Given $J_{\text{Mirror}}$ and $J_{\text{Forward,Ref}}$, the polarization vector of the backward propagating
field following reflection from the mirror can be expressed as

\[
J_{\text{Backward, Ref}}(\theta) = J_{\text{Mirror}} J_{\text{Forward, Ref}}(\theta) = \begin{bmatrix}
\cos^2(\theta) + j\sin^2(\theta) & \cos(\theta)\sin(\theta)(1-j) \\
-\cos(\theta)\sin(\theta)(1-j) & -\sin^2(\theta) - j\cos^2(\theta)
\end{bmatrix}
\] (C.4)

Approaching the QWP from the backward direction, the relationship between the reference frame, (now described by \(x', y'\)), and the QWP frame must take the reversal of the backward propagating wave's local reference into account. The reference frames are now related as in Figure C-4.

Figure C-4: Reference frame \((x', y')\) and QWP \((f', s')\) frame for the backward propagating wave.

As was done for the forward propagating wave, the polarization matrix for the backward wave is obtained by rotating \(J_{\text{Backward, Ref}}(\theta)\) into the QWP frame, applying the QWP Jones matrix, then rotating back into the \((x', y')\) reference frame. It should be noted that the rotation necessary to convert the reference frame into the QWP frame was \(+\theta\) in the forward direction, but is \(-\theta\) in the reverse direction. All the necessary matrix components are now in place to compute the overall Jones matrix.
of a two-pass QWP as

\[
J_{\text{2Pass QWP}}(\theta) = R(\theta) J_{\text{QWP Fast}} R(-\theta) J_{\text{Backward,Ref}}(\theta)
\]

\[
= \begin{bmatrix}
\cos(\theta) & \sin(\theta) \\
-\sin(\theta) & \cos(\theta)
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
0 & j
\end{bmatrix}
\begin{bmatrix}
\cos(\theta) & -\sin(\theta) \\
\sin(\theta) & \cos(\theta)
\end{bmatrix}
\times
\begin{bmatrix}
cos^2(\theta) + jsin^2(\theta) & \cos(\theta) \sin(\theta) (1 - j) \\
-\cos(\theta) \sin(\theta) (1 - j) & -\sin^2(\theta) - j\cos^2(\theta)
\end{bmatrix}
\]

\[
= \begin{bmatrix}
cos^2(\theta) - \sin^2(\theta) & 2\cos(\theta) \sin(\theta) \\
-2\cos(\theta) \sin(\theta) & \cos^2(\theta) - \sin^2(\theta)
\end{bmatrix}
\]

We can immediately observe that the elements of this Jones matrix are real. Therefore, any linearly polarized input vector will remain linearly polarized, but with a possible \( \theta \)-dependent rotation. For the two-pass QWP to act as a HWP, the following relationships must hold.

\[
\begin{bmatrix}
0 \\
\pm 1
\end{bmatrix} = [J_{\text{2Pass}}] \begin{bmatrix}
1 \\
0
\end{bmatrix}
\text{ and } \begin{bmatrix}
\pm 1 \\
0
\end{bmatrix} = [J_{\text{2Pass}}] \begin{bmatrix}
0 \\
1
\end{bmatrix}
\]

This constrains the elements of the two-pass QWP to be

\[
[J_{\text{2Pass}}] = \begin{bmatrix}
0 & \pm 1 \\
\pm 1 & 0
\end{bmatrix}
\]

Using the expressions for the elements of the two-pass QWP and the conditions on those elements given in Eq.\((C.7)\), \( \theta \) can be solved by simultaneously satisfying

\[
\cos(\theta) \sin(\theta) = \frac{1}{2} \text{ and } \sin^2(\theta) = \cos^2(\theta)
\]

For \( 0 < \theta < \pi/2 \), this is equal to \( \pi/4 \) radians, or \( 45^\circ \). This is the value of \( \theta \) required for two QWPs in cascade, and shows the equivalence to a double pass through a single QWP. Before concluding, we can examine the Jones matrix for \( \theta = 0 \), which is given
by

\[
J_{2\text{Pass}}(\theta = 0) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}
\]

For this case, the fast and slow axes of the QWP are aligned with the \(x\) and \(y\) axes respectively, and a horizontally or vertically polarized wave will only experience the influence of either the fast or slow axis. Therefore no rotation is introduced between components and the horizontal or vertical polarization of the input field is maintained.
Appendix D

Utilization of Amplitude Modulators in the Heterodyne Interferometer

An amplitude modulator emits two sidebands at $f_0 - f_{RF}$ and $f_0 + f_{RF}$. The SSBMs used in the integrated heterodyne interferometer are constructed of a pair of amplitude modulators in a Mach-Zehnder configuration, and are biased such that one sideband cancels. One interesting modification is to replace both SSBMs with amplitude modulators. The expressions obtained from doing so will be shown to reduce to the case of two SSBMs when one of the sidebands at each modulator is zero.

First, the fields from an amplitude modulator, modulated with a frequency $f_{RF}$, are described by

$$E_1 = E_{U1} \cos(2\pi f_{U1} + \phi_{U1}) + E_{L1} \cos(2\pi f_{L1} + \phi_{L1})$$
$$E_2 = E_{U2} \cos(2\pi f_{U2} + \phi_{U2}) + E_{L2} \cos(2\pi f_{L2} + \phi_{L2})$$

(E.1)

$E_U$ and $E_L$ are the field amplitudes of the upper and lower sidebands, $\phi_U$ and $\phi_L$ are arbitrary initial phases, and $f_U$ and $f_L$ are the upper and lower sideband frequencies defined by $f_U = f_0 + f_{RF}$ and $f_L = f_0 - f_{RF}$. Going forward the factors of $2\pi$ in the sine and cosine arguments are dropped for convenience. The power from each modulator is evenly split, with half directed to the reference detectors and the other half directed toward the measurement detectors. For the purposes of examining the resulting
heterodyne mixing components, factors of proportionality due to power division and detector responsivity are ignored. Therefore, the rather complex photocurrent from the reference detectors is proportional to

\[ I_R \propto E_U^1 E_L^1 \cos(2f_{RF1}t + (\phi_{U1} - \phi_{L1})) + E_U^1 E_U^2 \cos(f_{IF}t + (\phi_{U2} - \phi_{U1})) \]

\[ + E_U^1 E_L^2 \cos((f_{RF1} + f_{RF2})t + (\phi_{U1} - \phi_{L2})) + E_U^2 E_L^2 \cos(2f_{RF2}t + (\phi_{U2} - \phi_{L2})) \]

\[ + E_{L1} E_L^2 \cos(f_{IF}t + (\phi_{L2} - \phi_{L1})) + E_{L1} E_U^2 \cos((f_{RF1} + f_{RF2})t + (\phi_{U2} - \phi_{L1})) \]

(D.2)

To model the measurement photocurrent, assume the measurement signal originates from the second modulator. The phase of the second modulator, which corresponds to the measurement arm, is indexed with \( \tau \) to denote its dependence on the round-trip delay. Also, assume the phase from the first modulator (reference arm) at the measurement detector is equal to its value at the reference detectors, so that phase differences can be ignored.

\[ I_M \propto E_U^1 E_L^1 \cos(2f_{RF1}t + (\phi_{U1} - \phi_{L1})) + E_U^1 E_U^2 \cos(f_{IF}t + (\phi_{U2}(\tau) - \phi_{U1})) \]

\[ + E_U^1 E_L^2 \cos((f_{RF1} + f_{RF2})t + (\phi_{U1} - \phi_{L2}(\tau))) + E_U^2 E_L^2 \cos(2f_{RF2}t + (\phi_{U2}(\tau) - \phi_{L2}(\tau))) \]

\[ + E_{L1} E_L^2 \cos(f_{IF}t + (\phi_{L2}(\tau) - \phi_{L1})) + E_{L1} E_U^2 \cos((f_{RF1} + f_{RF2})t + (\phi_{U2}(\tau) - \phi_{L1})) \]

(D.3)

By setting \( E_{L1} \) and \( E_{L2} \) to zero at the reference detectors, it can be seen that the frequency components of the photocurrent reduce to those of dual SSBMs.

\[ I_R \propto E_U^1 E_U^2 \cos(f_{IF}t + (\phi_{U2} - \phi_{U1})) \]

(D.4)

For a single SSBM and a single amplitude modulator, where the amplitude modulator corresponds to the measurement (lower) arm, setting \( E_{L1} \) to zero gives the photocurrents at the reference and measurement detectors.

\[ I_R \propto E_U^1 E_U^2 \cos(f_{IF}t + (\phi_{U2} - \phi_{U1})) \]

\[ + E_U^1 E_L^2 \cos((f_{RF1} + f_{RF2})t + (\phi_{U1} - \phi_{L2})) \]

\[ + E_U^2 E_L^2 \cos(2f_{RF2}t + (\phi_{U2} - \phi_{L2})) \]

(D.5)
\[ I_M(\tau) \propto E_{U1}E_{U2}\cos(f_{IF}t + (\phi_{U2}(\tau) - \phi_{U1})) \\
+ E_{U1}E_{L2}\cos((f_{RF1} + f_{RF2})t + (\phi_{U1} - \phi_{L2}(\tau))) \\
+ E_{U2}E_{L2}\cos(2f_{RF2}t + (\phi_{U2}(\tau) - \phi_{L2}(\tau))) \]  

(D.6)

D.1 Synthetic Wavelength Interferometry utilizing an Amplitude Modulator

Equation (D.3) shows that the measurement signal for dual amplitude modulators contains two terms at \( f_{IF} \), two terms at \( f_{RF1} + f_{RF2} \), and a single term at \( 2f_{RF1} \) and \( 2f_{RF2} \). Two equal-frequency terms with different phases will add via the harmonic addition theorem to produce a sinusoid of the same frequency, but with a phase that is nonlinear with displacement. This is not ideal and will complicate attempts to accurately measure the displacement. However, if a single amplitude modulator were used to feed the measurement(lower) arm, the target could be simultaneously illuminated with two wavelengths. By using a SSBM in the reference (upper) arm, the duplicate frequency terms in Eq.(D.3) are eliminated. This is seen by examining Eq.(D.6), where there are three non-degenerate frequency terms that carry displacement information when referenced to their equal-frequency counterpart in the reference signal. The term at \( 2f_{RF2} \) consists of the phase shifts measured simultaneously by both frequencies of the measurement arm, and this produces a phase equivalent to that obtained with an equivalent synthetic wavelength.

\[
\Lambda_{eq} = \frac{\lambda_2\lambda_1}{|\lambda_2 - \lambda_1|} = \frac{\lambda_2\lambda_1}{\Delta \lambda} \quad (D.7)
\]

As discussed in Section 2.3.1, the synthetic measurement permits longer unambiguous measurements, but with less resolution than using individual wavelengths. However, in this scenario the individual phase measurements of the sidebands at frequencies equal to the intermediate frequency \( f_{IF} \) and the sum of the two RF frequencies \( (f_{RF1} + f_{RF2}) \) are still maintained. By proper filtering, the synthetic phase mea-
surement and the individual phase measurements can be obtained simultaneously. This provides the longer range possible using two wavelengths, with the accuracy and resolution of a single wavelength.
Appendix E

Input Waveguides, Adiabatic Couplers, and the Adiabatic Theorem

The input to the interferometer consists of two silicon waveguides separated by 30 microns. Either input may be used to couple to the interferometer and the use of two input waveguides provides some redundancy. The input and output waveguides are 150 nm wide and 110 nm high to accommodate the mode of a 6 micron cleaved fiber. In the SOI process, the standard height of the silicon waveguides are 220 nm, but are etched down vertically to 110 nm in this part of the structure. The dimensions of the input waveguide is constant over a length of approximately 105 microns. The waveguide is then adiabatically tapered to 400 nm in width over a distance of 150 microns while maintaining the etched height of 110 nm. Because the term adiabatic is used throughout this document, it makes sense to discuss it briefly.

To those familiar with thermodynamics, an adiabatic process is one that takes place without the transfer of heat or matter between a system and its surroundings, and over a time scale considered slow compared to the time it takes to reach equilibrium. The use of the term adiabatic as it applies to photonic devices is actually based on the adiabatic theory of quantum mechanics first proposed by Born and Fock in 1928 \[72,73\]. According to this theorem, a physical system will remain in its current eigenstate if the perturbation acting on it occurs slowly enough, and if its eigenvalues are distinct and can be ordered. To explain it differently, the change must occur
slowly enough that the system has an opportunity to adjust, and the eigenstate must maintain the same relative order in the new Hamiltonian as it was under the initial Hamiltonian. The corresponding process in thermodynamics is called a quasi-static process. The consequence in an electromagnetic structure such as a waveguide is that if we start with an eigenmode basis corresponding to a particular cross-section, then change the cross-section of the waveguide to one with a different set of eigenmodes, then the mode orders of the final configuration will be equivalent to those of the initial configuration if the change occurs adiabatically. So if the waveguide is initially in the fundamental mode, the final waveguide configuration will also be in its fundamental mode, without exciting any higher-order modes during the transition. For waveguide modes, the ordering is made according to the propagation constant $\beta$.

When compared to their non-adiabatic counterparts such as directional couplers, adiabatic structures tend to be much longer, but are less frequency sensitive once the adiabatic condition is satisfied. This lends itself to more broadband devices. For example, directional couplers are sensitive to wavelength and their coupling coefficient varies as a function of distance along the coupler. This is because both the fundamental even and odd modes are simultaneously excited at the input [74–77]. These modes propagate at different group velocities and interfere with each other along the length of the coupler. Thus the power distribution oscillates back and forth between the two waveguides over the length of the coupler. Adiabatic couplers, on the other hand, provide stable coupling over a wide bandwidth because only one mode ever exists within the structure at any point. Therefore, the couplers and splitters used in these designs tend to be adiabatic. One particularly useful property of adiabatic couplers is that depending on which of the two inputs is excited, the output of an adiabatic 3-dB coupler consists of either symmetric modes or anti-symmetric modes. If two fields $E_A$ and $E_B$ are incident on the upper and lower input port respectively,
then the output fields are given by

\[ E_{1,\text{Out}} = \frac{E_A}{\sqrt{2}} + \frac{E_B}{\sqrt{2}} \]  
\[ E_{2,\text{Out}} = \frac{E_A}{\sqrt{2}} - \frac{E_B}{\sqrt{2}} \]  

(E.1)

(E.2)

The adiabatic coupler then acts similarly to a 180° hybrid coupler. This quality is useful in implementing balanced detection in the interferometer. Continuing the description of the input taper, once the width of the guide is tapered adiabatically from 150 nm wide to 400 nm wide, the height is then adiabatically increased from 110 nm to 220 nm over a length of approximately 200 microns. The final dimensions of the silicon waveguide for internal guiding are 400 nm by 220 nm, which supports a single TE mode at 1550 nm.
Appendix F

The Effect of an Under-suppressed Carrier and Sidelobes in the Single-Sideband Modulator

In the analysis of the heterodyne interferometer, it was implicitly assumed that the Single-Sideband Modulator (SSBM) accepts an optical carrier $f_0$ and produces a shifted output frequency at $f = f_0 - f_{RF}$ or $f = f_0 + f_{RF}$. But by examining the equations governing the SSBM [53, 54], it can be seen that the suppression of the carrier $f_0$ and odd harmonics at $f_{RF}$ depends on how strongly the SSBM is driven electrically, and in particular, how well the bias points are set. The most important of the harmonics is at $f_0 + 3f_{RF}$. Under normal operating conditions the carrier and third-order harmonic can be suppressed on the order of 20-dB or more at the output, so they are usually not of concern. However, due to problems during a CMOS fabrication run, contacting issues prevented the modulators from being driven sufficiently to achieve the expected carrier and third-order harmonic suppression. The question that arose was whether this would limit the interferometer’s performance, and if so, how is this degradation quantified?

The following analysis considers these effects in the SSBMs. The earlier analysis of the interferometer assumed that the modulators were driven with RF frequencies $f_{RF1}$ and $f_{RF2}$ such that the difference between them produces the desired intermediate
frequency $f_{IF} = f_{RF2} - f_{RF1}$. Instead, assume the modulators generate composite signals given by

$$S_{SSBM1} = A \sin(f_0 t + \phi_1) + B \sin(f_1 t + \phi_2)$$  \hspace{1cm} (F.1)  

$$S_{SSBM2} = C \sin(f_0 t + \phi_3) + D \sin(f_2 t + \phi_4)$$  \hspace{1cm} (F.2)  

where going forward the $2\pi$ terms in the arguments of the sines and cosines are dropped for convenience. The coefficients $A$ through $D$, which represent the relative electric field amplitudes, are assumed real and would ideally satisfy $A \ll B$ and $C \ll D$. The phase terms are somewhat arbitrary and will be left unspecified. The output of a pair of balanced detectors such as those shown in Figure F-1 was previously examined.

Assuming a 50:50 split at the adiabatic coupler, the fields at the sum and difference ports can be written as

$$S_{\Sigma} = \frac{S_{SSBM1}}{\sqrt{2}} + \frac{S_{SSBM2}}{\sqrt{2}}$$  \hspace{1cm} (F.3)  

$$S_{\Delta} = \frac{S_{SSBM1}}{\sqrt{2}} - \frac{S_{SSBM2}}{\sqrt{2}}$$  \hspace{1cm} (F.4)
The resulting photocurrent from each detector is proportional to the power of the incident field, with the constant of proportionality equal to the detector responsivity \( R \) in units of A/W. Therefore the photocurrents are related to the modulator fields by

\[
I_{\Sigma} \propto S_{\Sigma}^2 = \frac{S_{SSBM1}^2}{2} + \frac{S_{SSBM2}^2}{2} + S_{SSBM1} \times S_{SSBM2}  
\]

\[
I_{\Delta} \propto S_{\Delta}^2 = \frac{S_{SSBM1}^2}{2} + \frac{S_{SSBM2}^2}{2} - S_{SSBM1} \times S_{SSBM2}  
\]

Expanding the terms explicitly yields

\[
S_{SSBM1}^2 = \frac{A^2}{2} (1 + \cos [2 (f_0 t + \phi_1)]) + \frac{B^2}{2} (1 + \cos [2 (f_1 t + \phi_2)]) + AB \cos [(f_1 - f_0)t + \phi_2 - \phi_1]  
\]

\[
S_{SSBM2}^2 = \frac{C^2}{2} (1 + \cos [2 (f_0 t + \phi_3)]) + \frac{D^2}{2} (1 + \cos [2 (f_2 t + \phi_4)]) + CD \cos [(f_2 - f_0)t + \phi_4 - \phi_3]  
\]

\[
S_{SSBM1} \times S_{SSBM2} = \frac{AC}{2} [\cos (\phi_3 - \phi_1) - \cos (2f_0 t + \phi_3 + \phi_1)] + \frac{AD}{2} [\cos ((f_2 - f_0)t + \phi_4 - \phi_1) - \cos ((f_2 + f_0)t + \phi_4 + \phi_1)] + \frac{BC}{2} [\cos ((f_1 - f_0)t + \phi_3 - \phi_2) - \cos ((f_1 + f_0)t + \phi_3 + \phi_2)] + \frac{BD}{2} [\cos ((f_2 - f_1)t + \phi_4 - \phi_2) - \cos ((f_2 + f_1)t + \phi_4 + \phi_2)]  
\]

High frequency terms of the form \(2f_i \) and \( f_i + f_j \) can be ignored, since they are outside the detector passband and are thereby low-pass filtered. The following definitions can also be applied: \( f_{IF} = f_2 - f_1 \), \( f_{RF2} = f_2 - f_0 \), and \( f_{RF1} = f_1 - f_0 \). Applying these identities to Eq.(F.7) and Eq.(F.8) gives

\[
S_{SSBM1}^2 = \frac{A^2}{2} + \frac{B^2}{2} + AB \cos (f_{RF1} t + \delta \phi_{21})  
\]

\[
S_{SSBM2}^2 = \frac{C^2}{2} + \frac{D^2}{2} + CD \cos (f_{RF2} t + \delta \phi_{43})  
\]
\[
S_{SSBM1} \times S_{SSBM2} = \frac{AC}{2} \cos (\delta \phi_{31}) + \frac{AD}{2} \cos (f_{RF2} t + \delta \phi_{41})
\]
\[
+ \frac{BC}{2} \left[ \cos (f_{RF1} t + \delta \phi_{32}) \right] + \frac{BD}{2} \left[ \cos (f_{IF} t + \delta \phi_{42}) \right]; \phi_{ij} = \phi_i - \phi_j
\]

Applying these expressions to the detector sum and difference currents and ignoring the DC terms, which will eventually be filtered, yields

\[
I_\Delta \propto \frac{AB}{2} \cos (f_{RF1} t + \delta \phi_{21}) + \frac{CD}{2} \cos (f_{RF2} t + \delta \phi_{43}) - \frac{AD}{2} \cos (f_{RF2} t + \delta \phi_{41}) - \frac{BC}{2} \cos (f_{RF1} t + \delta \phi_{32}) - \frac{BD}{2} \cos (f_{IF} t + \delta \phi_{42})
\]

\[
I_\Sigma \propto \frac{AB}{2} \cos (f_{RF1} t + \delta \phi_{21}) + \frac{CD}{2} \cos (f_{RF2} t + \delta \phi_{43}) + \frac{AD}{2} \cos (f_{RF2} t + \delta \phi_{41}) + \frac{BC}{2} \cos (f_{RF1} t + \delta \phi_{32}) + \frac{BD}{2} \cos (f_{IF} t + \delta \phi_{42})
\]

Using the trigonometric identify \(\cos(a+b) = \cos(a)\cos(b) - \sin(a)\sin(b)\), the currents can be expanded in terms of sines and cosines.

\[
I_\Delta = R_\Delta \times \left\{ \begin{array}{l}
\cos (f_{RF1} t) \left[ \frac{AB}{2} \cos (\delta \phi_{21}) - \frac{BC}{2} \cos (\delta \phi_{32}) \right] \\
- \sin (f_{RF1} t) \left[ \frac{AB}{2} \sin (\delta \phi_{21}) - \frac{BC}{2} \sin (\delta \phi_{32}) \right] \\
+ \cos (f_{RF2} t) \left[ \frac{CD}{2} \cos (\delta \phi_{43}) - \frac{AD}{2} \cos (\delta \phi_{41}) \right] \\
- \sin (f_{RF2} t) \left[ \frac{CD}{2} \sin (\delta \phi_{43}) - \frac{AD}{2} \sin (\delta \phi_{41}) \right] \\
- \cos (f_{IF} t) \left[ \frac{BD}{2} \cos (\delta \phi_{42}) \right] + \sin (f_{IF} t) \left[ \frac{BD}{2} \sin (\delta \phi_{42}) \right]
\end{array} \right\} (F.15)
\]

\[
I_\Sigma = R_\Sigma \times \left\{ \begin{array}{l}
\cos (f_{RF1} t) \left[ \frac{AB}{2} \cos (\delta \phi_{21}) + \frac{BC}{2} \cos (\delta \phi_{32}) \right] \\
- \sin (f_{RF1} t) \left[ \frac{AB}{2} \sin (\delta \phi_{21}) + \frac{BC}{2} \sin (\delta \phi_{32}) \right] \\
+ \cos (f_{RF2} t) \left[ \frac{CD}{2} \cos (\delta \phi_{43}) + \frac{AD}{2} \cos (\delta \phi_{41}) \right] \\
- \sin (f_{RF2} t) \left[ \frac{CD}{2} \sin (\delta \phi_{43}) + \frac{AD}{2} \sin (\delta \phi_{41}) \right] \\
+ \cos (f_{IF} t) \left[ \frac{BD}{2} \cos (\delta \phi_{42}) \right] - \sin (f_{IF} t) \left[ \frac{BD}{2} \sin (\delta \phi_{42}) \right]
\end{array} \right\} (F.16)
\]
Assuming equal responsivities, $\mathcal{R}_\Sigma = \mathcal{R}_\Delta = \mathcal{R}$, the difference current from the balanced detector pair is given by

$$I_{\text{Difference}} = \mathcal{R} \times \left\{ \begin{array}{l} \cos (f_{RF1}t) [BC \cos (\delta \phi_{32})] - \sin (f_{RF1}t) [BC \sin (\delta \phi_{32})] \\ + \cos (f_{RF2}t) [AD \cos (\delta \phi_{41})] - \sin (f_{RF2}t) [AD \sin (\delta \phi_{41})] \\ + \cos (f_{IF}t) [BD \cos (\delta \phi_{42})] - \sin (f_{IF}t) [BD \sin (\delta \phi_{42})] \end{array} \right\}$$

(F.17)

The sine and cosine terms in the above expression can be recombined to yield a simpler expression, given by

$$I_{\text{Difference}} = \mathcal{R} \times \left\{ \begin{array}{l} BC \cos (f_{RF1}t + \delta \phi_{32}) + \\ AD \cos (f_{RF2}t + \delta \phi_{41}) + \\ BD \cos (f_{IF}t + \delta \phi_{42}) \end{array} \right\}$$

(F.18)

For complete carrier suppression in both modulators, A and C are zero, leaving only the desired IF term. Notice that for equal responsivities $\frac{AB}{2} \cos (f_{RF1}t + \delta \phi_{21})$ and $\frac{CD}{2} \cos (f_{RF2}t + \delta \phi_{41})$ cancel. If the responsivities are not equal, then a more accurate expression, ignoring DC terms, is given by

$$I_{\text{Difference}} = \left( \mathcal{R}_\Sigma + \mathcal{R}_\Delta \right) \left[ \frac{BC}{2} \cos (f_{RF1}t + \delta \phi_{32}) + \frac{AD}{2} \cos (f_{RF2}t + \delta \phi_{41}) + \right] +$$

$$\left( \mathcal{R}_\Sigma - \mathcal{R}_\Delta \right) \left[ \frac{AB}{2} \cos (f_{RF1}t + \delta \phi_{21}) + \frac{CD}{2} \cos (f_{RF2}t + \delta \phi_{41}) \right]$$

(F.19)

This demonstrates that if the carrier is not sufficiently suppressed, then matching detector responsivities becomes even more important to avoid adding additional spurious RF terms. But while an unsuppressed carrier can potentially cause problems in the heterodyne interferometer, there is a fairly simple way to minimize their influence. Choosing values of $RF_1$ and $RF_2$ that are far higher than the desired IF, but close in value to each other situates the IF well below the modulating frequencies. By placing a low-pass filter at the output of the balanced detectors that includes the IF frequency
in its passband, but is well above cutoff at $RF_1$ and $RF_2$, the device can still perform
as a heterodyne interferometer, albeit with diminished performance. Any power in
the carrier and its various beat terms will also take power away from the desired IF
signal, and hence have an adverse effect on overall system performance.

The impact of including the sidelobe at $f_0 + 3f_{RF}$ is also examined, as the beating
of this term with the desired sideband at $f_0 - f_{RF}$ can produce in-band harmonics.
Assume that only these two sidebands are important and that the carrier and any
higher-order harmonics are insignificant. As in the previous analysis the composite
signals in SSBM1 and SSBM2 can be written as

$$S_{SSBM1} = A \cos(f_1 t) - B \cos(f_3 t)$$  \hspace{1cm} (F.20)
$$S_{SSBM2} = C \sin(f_2 t) - D \sin(f_4 t)$$  \hspace{1cm} (F.21)

Here again the $2\pi$ terms are dropped, the initial phases are ignored, and the frequency
combinations are defined as $f_1 = f_0 - f_{RF1}$, $f_2 = f_0 - f_{RF2}$, $f_3 = f_0 + 3f_{RF1}$, and
$f_4 = f_0 + 3f_{RF2}$. This yields

$$S_{SSBM1}^2 = \frac{A^2}{2} + \frac{B^2}{2} - AB \cos(4f_{RF1}t)$$  \hspace{1cm} (F.22)
$$S_{SSBM2}^2 = \frac{C^2}{2} + \frac{D^2}{2} - CD \cos(4f_{RF2}t)$$  \hspace{1cm} (F.23)

$$S_{SSBM1} \times S_{SSBM2} = \frac{AC}{2} \cos(f_{IF} t) - \frac{AD}{2} \cos[(3f_{RF2} + f_{RF1}) t] +$$
$$\frac{BD}{2} \cos(3f_{IF}) - \frac{BC}{2} \cos[(3f_{RF1} + f_{RF2}) t]$$  \hspace{1cm} (F.24)

Applying these expressions to the detector sum and difference currents and ignoring
the DC terms gives

\[
I_\Delta \propto -\frac{AB}{2} \cos \left(4f_{RF1} t\right) - \frac{CD}{2} \cos \left(4f_{RF2} t\right) - \frac{AC}{2} \cos \left(f_{IF} t\right) + \frac{AD}{2} \cos \left[(3f_{RF2} + f_{RF1}) t\right] + \frac{BC}{2} \cos \left[(3f_{RF1} + f_{RF2}) t\right]
\]

\[-\frac{BD}{2} \cos \left(3f_{IF} t\right)
\]

\[
I_\Sigma \propto -\frac{AB}{2} \cos \left(4f_{RF1} t\right) - \frac{CD}{2} \cos \left(4f_{RF2} t\right) + \frac{AC}{2} \cos \left(f_{IF} t\right) - \frac{AD}{2} \cos \left[(3f_{RF2} + f_{RF1}) t\right] - \frac{BC}{2} \cos \left[(3f_{RF1} + f_{RF2}) t\right]
\]

\[+\frac{BD}{2} \cos \left(3f_{IF} t\right)
\]

Assuming equal responsivities, \( \mathcal{R}_\Sigma = \mathcal{R}_\Delta = \mathcal{R} \), the difference current from the balanced detector pair is

\[
I_{\text{Difference}} = \mathcal{R} \times \left\{ AC \cos \left(f_{IF} t\right) - BC \cos \left[(3f_{RF1} + f_{RF2}) t\right] - AD \cos \left[(3f_{RF2} + f_{RF1}) t\right] + BD \cos \left(3f_{IF} t\right) \right\}
\]

\[
(F.27)
\]

If the responsivities are not equal, then a more accurate expression, again ignoring DC terms, is

\[
I_{\text{Difference}} = \frac{\mathcal{R}_\Sigma + \mathcal{R}_\Delta}{2} \left[ AC \cos \left(f_{IF} t\right) - BC \cos \left[(3f_{RF1} + f_{RF2}) t\right] - AD \cos \left[(3f_{RF2} + f_{RF1}) t\right] + BD \cos \left(3f_{IF} t\right) \right] - \frac{\mathcal{R}_\Sigma - \mathcal{R}_\Delta}{2} [AB \cos \left(4f_{RF1} t\right) + CD \cos \left(4f_{RF2} t\right)]
\]

\[
(F.28)
\]

Lastly, for the sake of completeness, the photocurrents are derived for the case where both the carrier and third-order harmonic are under-suppressed. In this derivation, as in the previous derivations, the assumption is that the SSBM operates ideally and
produces a single sideband. Assume the signal produced by each SSBM is

\[ S_{SSBM1} = A_0 \cos(f_0t) + A_1 \cos(f_1t) - A_2 \cos(f_3t) \quad \text{(F.29)} \]
\[ S_{SSBM2} = B_0 \cos(f_0t) + B_1 \cos(f_2t) - B_2 \cos(f_4t) \quad \text{(F.30)} \]

Here, we ignore the initial phases and define the frequencies as \( f_1 = f_0 - f_{RF1}, \)
\( f_2 = f_0 - f_{RF2}, \) \( f_3 = f_0 + 3f_{RF1}, \) and \( f_4 = f_0 + 3f_{RF2}. \) For the sake of brevity, and to
avoid the tedious algebra, the results for the squared signals and their products are
given below.

\[ S_{SSBM1}^2 = \frac{A_0^2}{2} + \frac{A_1^2}{2} + \frac{A_2^2}{2} + A_0 A_1 \cos(f_{RF1}t) - \]
\[ A_0 A_2 \cos(3f_{RF1}t) - A_1 A_2 \cos(4f_{RF1}t) \quad \text{(F.31)} \]
\[ S_{SSBM2}^2 = \frac{B_0^2}{2} + \frac{B_1^2}{2} + \frac{B_2^2}{2} + B_0 B_1 \cos(f_{RF2}t) - \]
\[ B_0 B_2 \cos(3f_{RF2}t) - B_1 B_2 \cos(4f_{RF2}t) \quad \text{(F.32)} \]

\[ S_{SSBM1} \times S_{SSBM2} = \frac{A_0}{2} [1 + B_1 \cos(f_{RF2}t) - B_2 \cos(3f_{RF2}t)] + \]
\[ \frac{A_1}{2} [B_0 \cos(f_{RF1}t) + B_1 \cos(f_{1ft}) - B_2 \cos[(3f_{RF2} + f_{RF1})t]] + \]
\[ \frac{A_2}{2} [B_0 \cos(f_{RF1}t) + B_1 \cos(f_{1ft}) - B_2 \cos[(3f_{RF2} + f_{RF1})t]] + \]
\[ \frac{A_1 B_1}{2} \cos(4f_{RF1}t) - \frac{A_1 A_2}{2} \cos(4f_{RF1}t) - \frac{A_1 B_2}{2} \cos(4f_{RF2}t) + \]
\[ \frac{A_2 B_2}{2} \cos[(3f_{RF2} + f_{RF1})t] + \frac{A_1 B_2}{2} \cos[(3f_{RF1} + f_{RF2})t] \quad \text{(F.33)} \]

Ignoring DC terms allow the sum and difference currents to be written as

\[ I_\Delta = R_\Delta \times \begin{cases} A_1 \cos(f_{RF1}t) \left[ \frac{A_0}{2} - \frac{B_0}{2} \right] + B_1 \cos(f_{RF2}t) \left[ \frac{B_0}{2} - \frac{A_0}{2} \right] - \frac{A_1 B_1}{2} \cos(f_{1ft}) - A_2 \cos(3f_{RF1}t) \left[ \frac{A_0}{2} - \frac{B_0}{2} \right] - B_2 \cos(3f_{RF2}t) \left[ - \frac{A_0}{2} + \frac{B_0}{2} \right] - \frac{A_1 B_2}{2} \cos(3f_{RF2}t) - \frac{A_1 A_2}{2} \cos(4f_{RF1}t) - \frac{B_1 B_2}{2} \cos(4f_{RF2}t) + \frac{A_1 B_2}{2} \cos[(3f_{RF2} + f_{RF1})t] + \frac{A_1 B_1}{2} \cos[(3f_{RF1} + f_{RF2})t] \end{cases} \quad \text{(F.34)} \]
\[ I_\Sigma = \mathcal{R}_\Sigma \times \left\{ \begin{array}{l}
A_1 \cos (f_{RF1}t) \left[ \frac{A_0}{2} + \frac{B_0}{2} \right] + B_1 \cos (f_{RF2}t) \left[ \frac{B_0}{2} + \frac{A_0}{2} \right] \\
+ \frac{A_1B_1}{2} \cos (f_{IF}t) - A_2 \cos (3f_{RF1}t) \left[ \frac{A_0}{2} + \frac{B_0}{2} \right] \\
- B_2 \cos (3f_{RF2}t) \left[ \frac{A_0}{2} + \frac{B_0}{2} \right] \\
+ \frac{A_2B_2}{2} \cos (3f_{IF}t) - \frac{A_1A_2}{2} \cos (4f_{RF1}t) - \frac{B_1B_2}{2} \cos (4f_{RF2}t) \\
- \frac{A_1B_2}{2} \cos [(3f_{RF2} + f_{RF1})t] - \frac{A_2B_1}{2} \cos [(3f_{RF1} + f_{RF2})t]
\end{array} \right\} \tag{F.35}
\]

If the responsivities are not equal, the difference current is

\[ I_{\text{Difference}} = A_1 \cos (f_{RF1}t) \left[ (\mathcal{R}_\Sigma - \mathcal{R}_\Delta) \frac{A_0}{2} + (\mathcal{R}_\Sigma + \mathcal{R}_\Delta) \frac{B_0}{2} \right] \\
+ B_1 \cos (f_{RF2}t) \left[ (\mathcal{R}_\Sigma - \mathcal{R}_\Delta) \frac{B_0}{2} + (\mathcal{R}_\Sigma + \mathcal{R}_\Delta) \frac{A_0}{2} \right] \\
+ (\mathcal{R}_\Sigma + \mathcal{R}_\Delta) \frac{A_1B_1}{2} \cos (f_{IF}t) \\
- A_2 \cos (3f_{RF1}t) \left[ (\mathcal{R}_\Sigma - \mathcal{R}_\Delta) \frac{A_0}{2} + (\mathcal{R}_\Sigma + \mathcal{R}_\Delta) \frac{B_0}{2} \right] \\
- B_2 \cos (3f_{RF2}t) \left[ (\mathcal{R}_\Sigma + \mathcal{R}_\Delta) \frac{A_0}{2} + (\mathcal{R}_\Sigma - \mathcal{R}_\Delta) \frac{B_0}{2} \right] \\
+ (\mathcal{R}_\Sigma + \mathcal{R}_\Delta) \frac{A_2B_2}{2} \cos (3f_{IF}t) - (\mathcal{R}_\Sigma - \mathcal{R}_\Delta) \frac{A_1A_2}{2} \cos (4f_{RF1}t) \\
- (\mathcal{R}_\Sigma - \mathcal{R}_\Delta) \frac{B_1B_2}{2} \cos (4f_{RF2}t) \\
- (\mathcal{R}_\Sigma + \mathcal{R}_\Delta) \frac{A_2B_1}{2} \cos [(3f_{RF2} + f_{RF1})t] \\
- (\mathcal{R}_\Sigma + \mathcal{R}_\Delta) \frac{A_1B_2}{2} \cos [(3f_{RF1} + f_{RF2})t] \tag{F.36}
\]

Assuming equal responsivities, \( R_\Sigma = R_\Delta = R \), the difference current from the balanced detector pair reduces to the much simpler expression given by

\[ I_{\text{Difference}} = \mathcal{R} \times \left\{ \begin{array}{l}
A_1B_0 \cos (f_{RF1}t) + A_0B_1 \cos (f_{RF2}t) + A_1B_1 \cos (f_{IF}t) \\
- A_2B_0 \cos (3f_{RF1}t) - A_0B_2 \cos (3f_{RF2}t) + A_2B_2 \cos (3f_{IF}t) \\
- A_2B_1 \cos [(3f_{RF1} + f_{RF2})t] - A_1B_2 \cos [(3f_{RF2} + f_{RF1})t]
\end{array} \right\} \tag{F.37}
\]

Below is a summary of the difference current spectral components for both equal and unequal responsivities and their amplitude coefficients. The power in each term is proportional to the square of the amplitude coefficients.
<table>
<thead>
<tr>
<th>Spectral Component</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{RF1}$</td>
<td>$A_1 B_0$</td>
</tr>
<tr>
<td>$f_{RF2}$</td>
<td>$A_0 B_1$</td>
</tr>
<tr>
<td>$f_{IF}$</td>
<td>$A_1 B_1$</td>
</tr>
<tr>
<td>$3f_{RF1}$</td>
<td>$A_2 B_0$</td>
</tr>
<tr>
<td>$3f_{RF2}$</td>
<td>$A_0 B_2$</td>
</tr>
<tr>
<td>$3f_{IF}$</td>
<td>$A_2 B_2$</td>
</tr>
<tr>
<td>$3f_{RF1} + f_{RF2}$</td>
<td>$A_2 B_1$</td>
</tr>
<tr>
<td>$3f_{RF2} + f_{RF1}$</td>
<td>$A_1 B_2$</td>
</tr>
</tbody>
</table>

Table F.1: Spectral components for equal responsivities when analyzing the effect of an unsuppressed carrier and sidebands in the SSBM.

<table>
<thead>
<tr>
<th>Spectral Component</th>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{RF1}$</td>
<td>$A_1 B_0, A_0 A_1$</td>
</tr>
<tr>
<td>$f_{RF2}$</td>
<td>$A_0 B_1, B_0 B_1$</td>
</tr>
<tr>
<td>$f_{IF}$</td>
<td>$A_1 B_1$</td>
</tr>
<tr>
<td>$3f_{RF1}$</td>
<td>$A_0 A_2, A_2 B_0$</td>
</tr>
<tr>
<td>$3f_{RF2}$</td>
<td>$A_0 B_2, B_0 B_2$</td>
</tr>
<tr>
<td>$3f_{IF}$</td>
<td>$A_2 B_2$</td>
</tr>
<tr>
<td>$3f_{RF1} + f_{RF2}$</td>
<td>$A_2 B_1$</td>
</tr>
<tr>
<td>$3f_{RF2} + f_{RF1}$</td>
<td>$A_1 B_2$</td>
</tr>
<tr>
<td>$4f_{RF1}$</td>
<td>$A_1 A_2$</td>
</tr>
<tr>
<td>$4f_{RF2}$</td>
<td>$B_1 B_2$</td>
</tr>
</tbody>
</table>

Table F.2: Spectral components for non-equal responsivities when analyzing the effect of an unsuppressed carrier and sidebands in the SSBM.
Bibliography


