



A review of demodulation techniques for amplitude-modulation atomic force microscopy

Michael G. Ruppert^{*1}, David M. Harcombe¹, Michael R. P. Ragazzon², S. O. Reza Moheimani³ and Andrew J. Fleming¹

Review

Open Access

Address:

¹School of Electrical Engineering and Computing, The University of Newcastle, Callaghan, NSW, 2308, Australia, ²Department of Engineering Cybernetics, NTNU, Norwegian University of Science and Technology, Trondheim, Norway and ³Department of Mechanical Engineering, The University of Texas at Dallas, Richardson, TX, USA

Email:

Michael G. Ruppert^{*} - Michael.Ruppert@newcastle.edu.au;
David M. Harcombe - David.Harcombe@uon.edu.au;
Michael R. P. Ragazzon - ragazzon@itk.ntnu.no;
S. O. Reza Moheimani - Reza.Moheimani@utdallas.edu;
Andrew J. Fleming - Andrew.Fleming@newcastle.edu.au

* Corresponding author

Keywords:

amplitude estimation; atomic force microscopy; amplitude modulation; digital signal processing; field-programmable gate array

Beilstein J. Nanotechnol. **2017**, *8*, 1407–1426.

doi:10.3762/bjnano.8.142

Received: 23 December 2016

Accepted: 07 June 2017

Published: 10 July 2017

This article is part of the Thematic Series "Advanced atomic force microscopy".

Guest Editor: T. Glatzel

© 2017 Ruppert et al.; licensee Beilstein-Institut.

License and terms: see end of document.

Abstract

In this review paper, traditional and novel demodulation methods applicable to amplitude-modulation atomic force microscopy are implemented on a widely used digital processing system. As a crucial bandwidth-limiting component in the *z*-axis feedback loop of an atomic force microscope, the purpose of the demodulator is to obtain estimates of amplitude and phase of the cantilever deflection signal in the presence of sensor noise or additional distinct frequency components. Specifically for modern multifrequency techniques, where higher harmonic and/or higher eigenmode contributions are present in the oscillation signal, the fidelity of the estimates obtained from some demodulation techniques is not guaranteed. To enable a rigorous comparison, the performance metrics tracking bandwidth, implementation complexity and sensitivity to other frequency components are experimentally evaluated for each method. Finally, the significance of an adequate demodulator bandwidth is highlighted during high-speed tapping-mode atomic force microscopy experiments in constant-height mode.

Introduction

Amplitude modulation is one of the oldest forms of modulation in analog communication systems, mostly due to its simplicity of implementation [1]. Not only is the modulation principle

used in many forms of scientific instruments and sensors [2-4], but numerous fields of research also rely on the improved detection sensitivity made available by this technique [5-7].

While the invention of the atomic force microscope (AFM) [8] in the late 1980s had little to do with modulation to begin with, a fundamental prerequisite was given by the nonlinear tip–sample interaction force. With the advent of dynamic imaging modes [9], in which the microcantilever is excited at one of its resonance frequencies, the foundation for transmitting information via modulation was established. These imaging modes are especially suitable for the investigation of delicate matter and biological samples because of the low tip–sample forces [10] and have led to the instrument establishing itself as a key enabling technology for the nanoscale analysis of objects and materials properties for both research and industry [11,12].

Dynamic operating modes of the AFM can map the surface topography of a specimen with high spatial resolution by scanning a sharp tip located at the end of an actively driven microcantilever over the surface of a sample. Due to the nonlinear tip–sample forces acting on the cantilever, a feedback loop has to be employed in order to maintain a fixed setpoint with respect to the sample; the controller performs disturbance rejection by commanding a nanopositioner in its vertical direction. As the high-frequency cantilever deflection signal cannot be controlled directly, low-frequency measurables such as the change in oscillation amplitude in amplitude-modulation AFM [11] have to be employed. Other feedback variables such as the shift in cantilever resonance frequency in frequency-modulation AFM [13] or the phase shift in phase-modulation AFM [14] have also been used. Situated at the heart of these dynamic methods, a demodulator is employed to estimate amplitude and phase of the cantilever deflection signal.

A number of demodulation techniques can be found in the existing literature, some of which have found regular use in commercial AFM systems. The performance metrics, tracking bandwidth and sensitivity to other frequency components, are especially important in high-speed [15–18] and multifrequency AFM [19] applications. As the tracking bandwidth directly affects the achievable scan rate, it should be maximized. However, this also increases the noise bandwidth. On the other hand, in multifrequency AFM applications, the sensitivity to other frequency components is of greatest concern. These applications may include multiple eigenmode contributions [20–22], higher harmonics [23–25], and multi-tone near-resonance frequency components [26–28].

For instance, RMS-to-DC conversion [29] is low in implementation complexity and can achieve high tracking bandwidth, but it is sensitive to other frequency components. In contrast, the lock-in amplifier [30–32] is a narrow-band technique that has been adopted as the industry-wide standard in commercial

AFMs, since it is insensitive to other frequency components but is limited in tracking bandwidth.

Inspired by image-rejection mixers [33] and modulated–demodulated control [34], a high-bandwidth lock-in amplifier was recently proposed and implemented to improve upon this constraint [35]. However, the method is still ultimately limited by the low-pass filters that are required to account for residual phase mismatches.

For high-speed AFM applications, as required for the study of fast biological processes [36,37], the above methods are not suitable and have led to the development of fast single-wave detectors in the form of the peak-hold method [38,39] and coherent demodulator [40–43]. The latter is an all-digital lock-in amplifier where the characteristic low-pass filter is replaced by a precise numerical integration scheme. While these methods can yield fast estimates with low latency, they may not be suitable for multifrequency AFM methods where non-integer multiples of the fundamental frequency are present in the deflection signal.

The demand for a high tracking bandwidth while maintaining insensitivity to additional frequencies in the signal has motivated the development of filters such as the time-varying Kalman filter [44] and Lyapunov filter [45,46]. These methods are based on a linear parametric model of the cantilever deflection signal and were shown to be extendable for the estimation of multiple frequencies for multifrequency AFM [47–49].

Observer-based approaches have also been investigated to provide an alternative feedback signal other than the estimated amplitude. For instance, if an observer is constructed from the free-air model of the cantilever, the innovation signal (error signal between measurement and model output) will contain information of the disturbance profile during the transient response of the cantilever [50–53]. In addition, it was shown that the tip–sample force can be estimated directly by assuming it takes the form of an impulse train [54]. In this way, the tip–sample force is estimated directly, thus potentially enabling high-bandwidth z -axis control by relying on feedback from the force estimate instead of from the cantilever oscillation amplitude.

This article aims to provide a rigorous experimental comparison of the most commonly used demodulation methods for amplitude-modulation AFM over their entire tracking bandwidth range. The methods considered are the lock-in amplifier, high-bandwidth lock-in amplifier, Lyapunov filter, Kalman filter, RMS-to-DC conversion (moving-average filter and mean absolute deviation computation), peak detector and coherent

demodulator. To make a fair comparison, a widely used digital signal processing system (LabVIEW) is used and the implementations are unified to a common sample rate. The performance metrics are tracking bandwidth, implementation complexity, sensitivity to other frequency components and total integrated noise of the amplitude estimate as a function of the tracking bandwidth. The experimental analysis is concluded by high-speed constant-height tapping-mode AFM experiments which highlight the case where the demodulator is the bandwidth bottleneck in the z -axis feedback loop.

Fundamentals of amplitude modulation and demodulation

Modulation

A basic amplitude-modulated (double-sideband full carrier) signal is obtained by mixing a modulating signal $y_m(t)$ at a modulation index M and frequency $\omega_m = 2\pi f_m$ with a carrier signal $y_c(t)$ with (for the sake of brevity) unity amplitude, phase ϕ , and frequency $\omega_c = 2\pi f_c$ such that

$$\begin{aligned} y_m(t)y_c(t) &= [1 + M \sin(\omega_m t)] \sin(\omega_c t + \phi) \\ &= \sin(\omega_c t) + \frac{M}{2} \cos[(\omega_c - \omega_m)t + \phi] \\ &\quad - \frac{M}{2} \cos[(\omega_c + \omega_m)t + \phi]. \end{aligned} \quad (1)$$

In the time domain, this process is shown in Figure 1a, where the minimum and the maximum levels attained by the amplitude-modulated signal are $1 - M$ and $1 + M$, respectively. It can be seen from Equation 1 that the modulation process creates distinct frequency components located at f_c and $f_c \pm f_m$. The latter components are termed the upper and lower sidebands and are centered symmetrically around the carrier frequency for $f_m < f_c$, illustrated in Figure 1b. As the modulating frequency increases, these sidebands move away from the carrier until they appear at DC and at $2f_c$ for the limit where $f_m = f_c$. For the case where $f_m > f_c$, $y(t)$ resembles a distorted wave with sidebands located at $f_m \pm f_c$ and can therefore no longer be considered an amplitude-modulated signal because the sidebands are no longer symmetrically located around the carrier frequency. For the application in AFM, this case is practically irrelevant as it corresponds to amplitude changes appearing faster than the tapping frequency.

Demodulation

The process of demodulation always requires a nonlinear operation on a signal in order to estimate a baseband signal proportional to the modulation of the carrier. Based on this nonlinearity, the demodulation methods can be broadly classified as

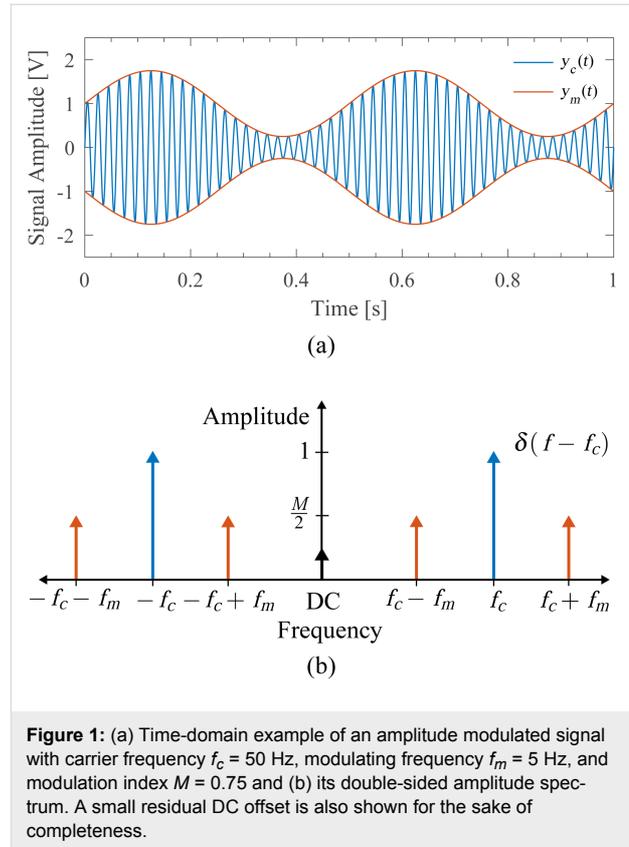
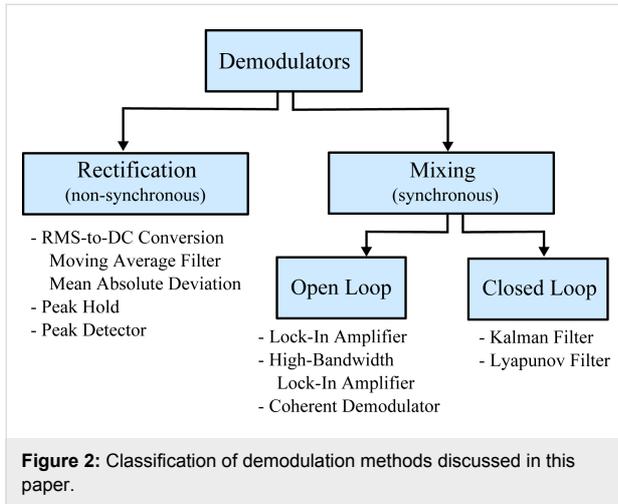


Figure 1: (a) Time-domain example of an amplitude modulated signal with carrier frequency $f_c = 50$ Hz, modulating frequency $f_m = 5$ Hz, and modulation index $M = 0.75$ and (b) its double-sided amplitude spectrum. A small residual DC offset is also shown for the sake of completeness.

methods using rectification (non-synchronous detection) and methods using mixing with a reference oscillator signal (synchronous detection). For demodulators of the latter class, the reference signal can be either a square wave, most commonly used for analog implementations, or a sinusoid, most commonly used for digital implementations as is the case in this paper. Within the class of demodulators using mixing, further classification can be made based on how the $2f_c$ component from the mixing process is filtered out. While the open-loop methods rely on either general or numerically precise low-pass filters, the closed-loop methods employ feedback of the parameterized signal states to eliminate this component. An overview of the demodulator classification is shown in Figure 2. As will be discussed in the course of this paper, each class has distinct properties with regards to tracking bandwidth, implementation complexity and sensitivity to other frequency components.

The linear parameterization used by the demodulation techniques based on mixing is derived from a sine wave with known carrier frequency ω_c , unknown amplitude $A = A(t)$ and unknown phase $\phi = \phi(t)$ of the form

$$y(t) = A(t) \sin[\omega_c t + \phi(t)]. \quad (2)$$



The signal can be rewritten as a sum of its quadrature and in-phase components by applying trigonometric identities in order to obtain a linear parameterization (the time dependency for slowly changing parameters are left out for the sake of readability)

$$\begin{aligned}
 y(t) &= \underbrace{A \cos(\omega_c t) \sin(\phi)}_{\text{quadrature component}} + \underbrace{A \sin(\omega_c t) \cos(\phi)}_{\text{in-phase component}} \\
 &= \begin{bmatrix} \cos(\omega_c t) & \sin(\omega_c t) \end{bmatrix} \begin{bmatrix} A \sin(\phi) & A \cos(\phi) \end{bmatrix}^T \quad (3) \\
 &= \mathbf{c}(t) \mathbf{x}.
 \end{aligned}$$

As we will make frequent reference to this parameterization, the entries of the vector $\mathbf{c}(t)$ are termed the quadrature and in-phase sinusoids and the entries of the state vector

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$$

are termed the quadrature and in-phase states. In this form, amplitude and phase can be directly calculated as

$$\begin{aligned}
 A &= \sqrt{x_1^2 + x_2^2} \\
 \phi &= \arctan\left(\frac{x_1}{x_2}\right). \quad (4)
 \end{aligned}$$

Performance metrics

The performance metrics used for the comparison of the demodulation methods are implementation complexity, tracking bandwidth, sensitivity to other frequency components and total integrated noise of the amplitude estimate as a function of the tracking bandwidth. The implementation complexity is qualitatively evaluated based on the maximum free-running sample rate achieved by the digital signal processing system. Where

applicable, latencies arising from fixed time-delays in the implementation of the methods are highlighted. The tracking bandwidth is defined as the frequency $f_{-3\text{dB}}$, at which the amplitude estimate drops by -3 dB. This figure of merit is important to determine both the speed of convergence and the amount of noise suppression in the estimate. This relationship is clearly identified by plotting the total integrated noise of the amplitude estimate against the tracking bandwidth for a known input noise density. Lastly, the sensitivity to other frequency components is evaluated to determine the ability of each method to filter out any signal at frequencies other than the carrier frequency of interest.

Review of demodulation methods

Lock-in amplifier

The lock-in amplifier [30-32] mixes the input signal (Equation 2) with in-phase and quadrature sinusoids to obtain

$$\begin{aligned}
 y_i(t) &= y(t) \sin(\omega_c t) \\
 &= \frac{1}{2} A \left[\cos(\phi) - \cos(2\omega_c t + \phi) \right] \quad (5)
 \end{aligned}$$

and

$$\begin{aligned}
 y_q(t) &= y(t) \cos(\omega_c t) \\
 &= \frac{1}{2} A \left[\sin(\phi) + \sin(2\omega_c t + \phi) \right]. \quad (6)
 \end{aligned}$$

From Equation 5 and Equation 6, it can be seen that the mixing process generates harmonics at $2f_c$, which need to be removed by employing a low-pass filter with $f_{lp} \ll 2f_c$, as illustratively shown in Figure 3b. Further, any residual DC offset in the input signal will generate a harmonic at f_c , which is the reason why lock-in amplifiers should always be AC-coupled. The order and cut-off frequency of the low-pass filter directly determines the tracking bandwidth and hence the noise performance. For instance, in order to limit the ripple to 1% of the signal, a -40 dB suppression of the $2f_c$ component is required. A 2nd-order low-pass filter would achieve this by limiting the bandwidth to approximately a decade below the carrier frequency.

A lock-in amplifier described by Equations 4–6 can be realized digitally with a direct digital synthesizer (DDS) to create the reference sine and cosine signals, two multipliers, two low-pass filters, and an output block with square-root functionality and an arctan calculation method such as polynomial approximation or the CORDIC algorithm [55] to calculate the phase [43]. Such an implementation is schematically shown in the block diagram in Figure 3a, where the output block represents Equations

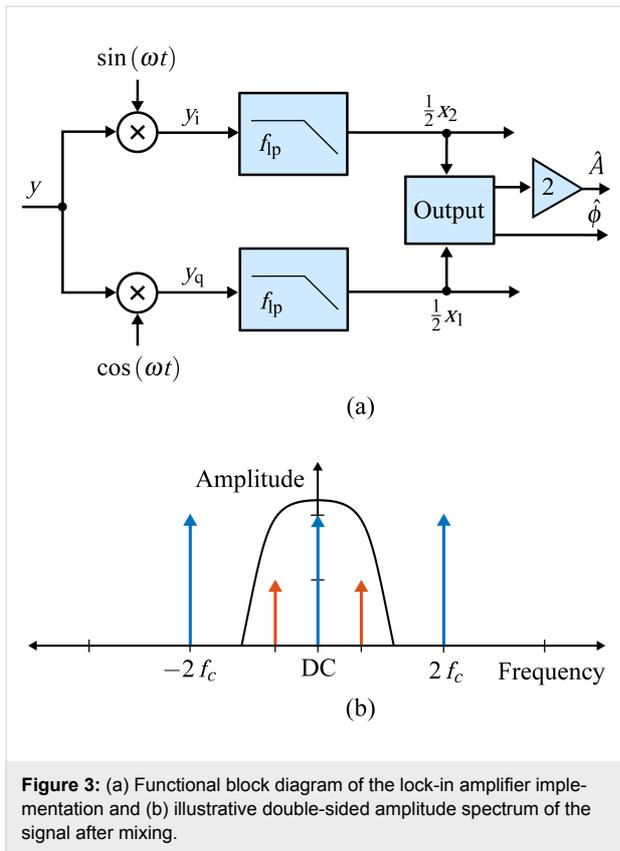


Figure 3: (a) Functional block diagram of the lock-in amplifier implementation and (b) illustrative double-sided amplitude spectrum of the signal after mixing.

tion 4. A factor of two is necessary to obtain the correct amplitude scaling.

High-bandwidth lock-in amplifier

The high bandwidth lock-in amplifier is a novel extension to the standard lock-in amplifier technique, which employs phase cancellation to precisely cancel the $2f_c$ term [35]. The method is inspired by radio frequency image rejection mixers [33] and modulated–demodulated control [34,56]. Compared to the standard lock-in amplifier, this demodulation scheme essentially requires two additional multipliers, which operate on the input signal shifted by 90°

$$y_{\perp}(t) = A \cos(\omega_c t + \phi) \tag{7}$$

to form the respective output products

$$\begin{aligned} y_{\perp,i}(t) &= \frac{1}{2} A [\cos(\phi) + \cos(2\omega_c t + \phi)] \\ y_{\perp,q}(t) &= \frac{1}{2} A [-\sin(\phi) + \sin(2\omega_c t + \phi)] \end{aligned} \tag{8}$$

which are added to the output products of the LIA mixers to exactly cancel the $2f_c$ components

$$\begin{aligned} x_1 &= y_q(t) - y_{\perp,q}(t) \\ x_2 &= y_i(t) + y_{\perp,i}(t). \end{aligned} \tag{9}$$

Although an analog implementation is possible [35], in practice imperfect phase cancellation due to circuit mismatches still requires post-mixing low-pass filters. However, as the $2f_c$ terms are heavily attenuated, the bandwidth of the filters can be increased. This is illustrated in Figure 4b.

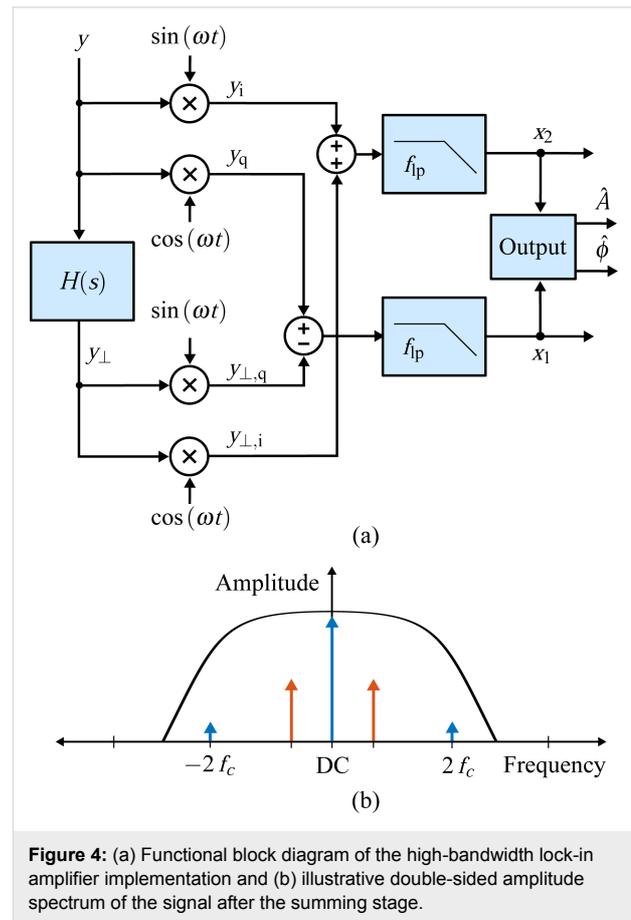


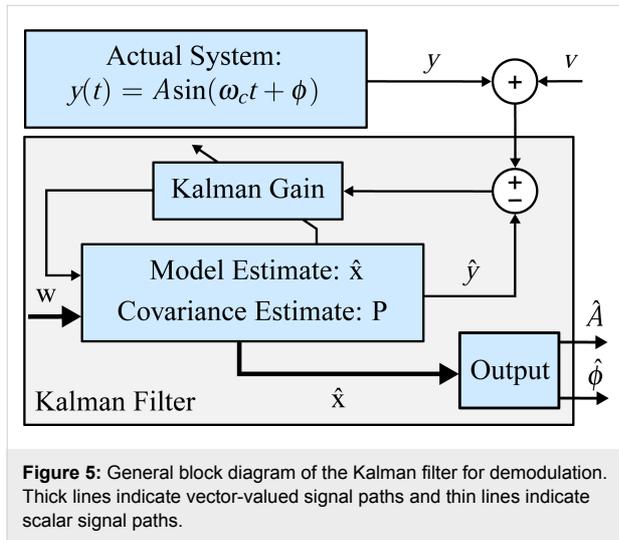
Figure 4: (a) Functional block diagram of the high-bandwidth lock-in amplifier implementation and (b) illustrative double-sided amplitude spectrum of the signal after the summing stage.

In a digital implementation, the standard lock-in amplifier implementation has to be extended with two additional multipliers, two summing stages and a 90° phase-shift block $H(s)$ as shown in Figure 4a. Such an operation can be realized with a Hilbert transform filter or an all-pass filter tuned to the carrier frequency [57]. Amplitude and phase are recovered by employing the output Equation 4 without an additional scaling factor.

Kalman filter

The Kalman filter [58] can generally be considered as a recursive algorithm that makes the best possible trade-off between modeled and measured information to estimate unknown variables of a process. Specifically, if the modeling error and the

noise in the measurement can be considered to have a Gaussian distribution, the Kalman filter is the minimum variance estimator in the least-squares sense [59]. Typical uses of the Kalman filter include sensor fusion, smoothing noisy data and estimation of internal states in numerous applications ranging from navigation, economics and signal processing [60]. Characteristic and fundamental to its working principle is the existence of a linear system model that describes the dynamics to be estimated and the presence of feedback generated from the Kalman gain, which dictates the rate of convergence. This structure is schematically shown in Figure 5.



A linear time-invariant (LTI) state-space model of the signal described by Equation 2 can be obtained by choosing $\tilde{x}_1 = y$ and $\tilde{x}_2 = \dot{y}$ as the state variables and $y = \tilde{x}_1$ as the output to yield

$$\begin{aligned} \dot{\tilde{\mathbf{x}}} &= \mathbf{A}\tilde{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -\omega_c^2 & 0 \end{bmatrix} \tilde{\mathbf{x}} \\ y &= \mathbf{c}\tilde{\mathbf{x}} = [1 \quad 0] \tilde{\mathbf{x}}. \end{aligned} \quad (10)$$

However, due to the sparse nature of the resulting dynamic matrix \mathbf{A} , especially when modeling higher resonance frequencies, this model is generally ill-conditioned for the use in an observer such as the Kalman filter, which requires an inversion. This problem is circumvented by applying a time-variant transformation $\tilde{\mathbf{x}} = \mathbf{T}(t)\mathbf{x}$ with [44]

$$\mathbf{T}(t) = \begin{bmatrix} \cos(\omega_c t) & \sin(\omega_c t) \\ -\omega_c \sin(\omega_c t) & \omega_c \cos(\omega_c t) \end{bmatrix} \quad (11)$$

such that a time-varying but numerically well-conditioned state-space representation is obtained. This constitutes the process

model of the Kalman filter, which in its discretized form is given as

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k + \mathbf{w} \\ y_k &= \mathbf{c}_k \mathbf{x}_k + v \end{aligned} \quad (12)$$

$$E[\mathbf{w}\mathbf{w}^T] = \mathbf{Q}, \quad E[vv^T] = R, \quad E[\mathbf{w}v^T] = 0 \quad (13)$$

where $\mathbf{c}_k = [\cos(\omega_c k T_s) \quad \sin(\omega_c k T_s)]$, T_s is the sample period, and \mathbf{w} and v are the process noise and the measurement noise with their respective covariance matrices \mathbf{Q} and R . A similar system description was successfully used in tracking power system voltage phasors [61]. In this representation, the states $x_{1,k}$ and $x_{2,k}$ are assumed to be random variables describing the quadrature and in-phase states of Equation 3. Moreover, \mathbf{Q} determines the amount of uncertainty in the model (Equation 12) and R the amount of noise in the measurement. If R is chosen to be the standard deviation σ of the Gaussian noise in the sensor signal $y(t)$ ($R = \sigma^2$), then \mathbf{Q} remains the only tuning variable to set the tracking bandwidth of the estimated amplitude and phase. The recursive implementation follows the standard equations of the Kalman filter [62,63] and are stated in Appendix A. Amplitude and phase are recovered by employing the output equations in Equation 4.

Lyapunov filter

The Lyapunov filter is conceptually related to the Kalman filter in the sense that it uses feedback to correct the estimated quadrature and in-phase states of Equation 3 of the linear parameterization of the signal (Equation 2). Compared to the Kalman filter, it is significantly less computational expensive as it does not require the computation of a covariance matrix to determine the feedback gain. Instead, the gain is a predetermined constant parameter that is related to the Kalman gain for certain conditions [46].

In the literature about adaptive control, the Lyapunov filter is regarded as an online adaptive estimator for which the estimation error relates to the parameter error through a strictly positive real (SPR) transfer function $W(s)$ [64]. This SPR property is exploited in designing the update law via a Lyapunov stability proof to show boundedness of the error, hence the name. An additional persistency of excitation property [64] guarantees exponential convergence of the parameters. The estimator can be written in the compact form [45,46]

$$\begin{aligned} \dot{\mathbf{x}} &= \gamma \mathbf{c}(t)^T W(s)(y - \hat{y}) \\ \hat{y} &= \mathbf{c}(t) \mathbf{x}, \end{aligned} \quad (14)$$

where γ is the constant gain parameter for tuning the bandwidth and \hat{y} is the estimated signal. For simplicity, $W(s)$ can be assumed to be a constant 1 as any other assumption will limit the tracking bandwidth [46]. A digital implementation requires a DDS, four multipliers, two gain blocks, a discrete filter and two discrete integrators but can also be realized with scalar operations as shown in Figure 6. The amplitude and phase are recovered by employing the output equations in Equation 4.

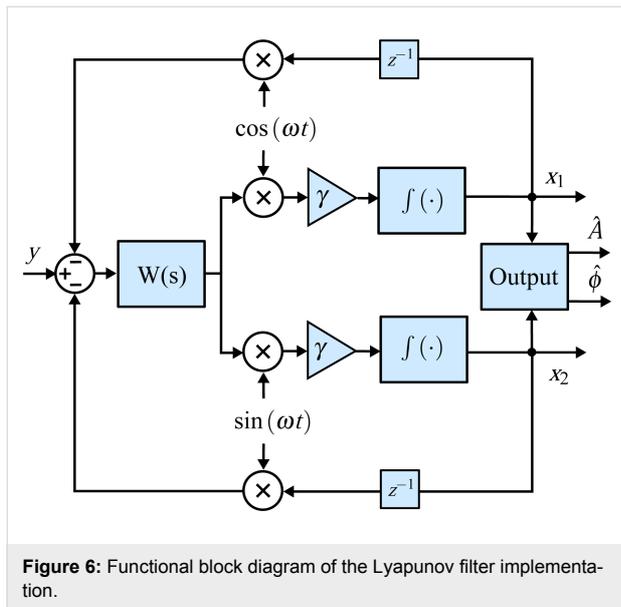


Figure 6: Functional block diagram of the Lyapunov filter implementation.

RMS-to-DC conversion methods

One of the easiest forms of amplitude estimation is RMS-to-DC conversion. The root-mean-square (RMS) value y_{rms} of a sinusoidal signal $y(t)$ with period T is proportional to the amplitude of the signal and is defined as

$$y_{rms} = \sqrt{\frac{1}{T} \int_t^{t+T} y^2(t) dt} = \frac{A}{\sqrt{2}}. \quad (15)$$

True RMS-to-DC conversion can be achieved in analog using either direct or indirect computation. The direct method performs the squaring, averaging and square-root functionality using multipliers and operational amplifiers. While the benefit of direct computation is a large bandwidth, it has a limited dynamic range due to the squaring stage [29]. Indirect computation alleviates this problem by employing feedback and division of the average output, which significantly improves the dynamic range but comes at the expense of tracking bandwidth. A number of direct and indirect analog true RMS-to-DC converters are available commercially. For a purely sinusoidal signal, the output of these methods are proportional to the oscillation amplitude. However, biased amplitude estimates are ob-

tained when additional frequency components are present in the signal.

Moving average filter

Equation 15 can be implemented digitally by obtaining the moving average using a finite impulse response (FIR) filter [57] of the squared input signal and passing it through a subsequent square-root stage. The integration period T in Equation 15 is related to the length of the $(n + 1)$ -tap moving average FIR filter as $T = n/f_s$, which dictates the tracking bandwidth of this approach. The number of samples n should be a half-period integer multiple of the sample rate f_s . A functional block diagram of this implementation is shown in Figure 7a. As a true RMS-to-DC converter, the output needs to be scaled by $\sqrt{2}$ to obtain the amplitude as evident from Equation 15. This method has increasing latency for decreasing tracking bandwidth.

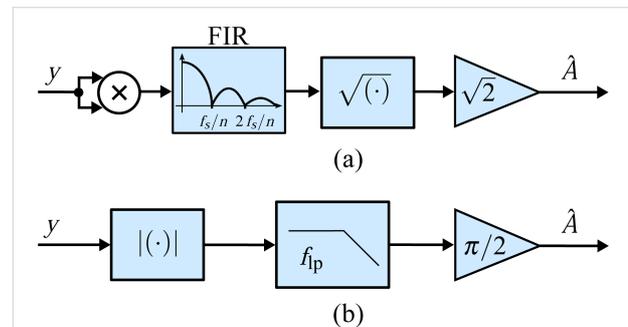


Figure 7: Functional block diagram of (a) moving average filter and (b) mean absolute deviation measurement to perform digital RMS-to-DC conversion.

Mean absolute deviation

In tapping-mode AFM, RMS-to-DC conversion was typically performed using a precision rectifier circuit and a low-pass filter [36]. Strictly, this is not RMS-to-DC conversion but mean absolute deviation [29], which calculates the AC average of the waveform $2A/\pi$. In a digital implementation, this circuit can be realized with an absolute value block representing the rectifier, a low-pass filter and an output scaling factor of $\pi/2$. The functional block diagram of this implementation is shown in Figure 7b.

Peak hold and peak detector method

The peak hold technique [36,38] was specifically developed for high-speed tapping-mode AFM, enabling video-rate imaging of Myosin V [15]. The analog implementation of this method comprises two sample and hold circuits to hold both the positive and negative peaks of the carrier signal for the duration of a cycle triggered by using a zero-cross comparator on the phase-shifted signal. By calculating the arithmetic mean of the outputs of the two sample and hold circuits and passing it through a

low-pass filter to set the bandwidth, the output represents the amplitude of the input signal. The functional block diagram of this implementation is shown in Figure 8a.

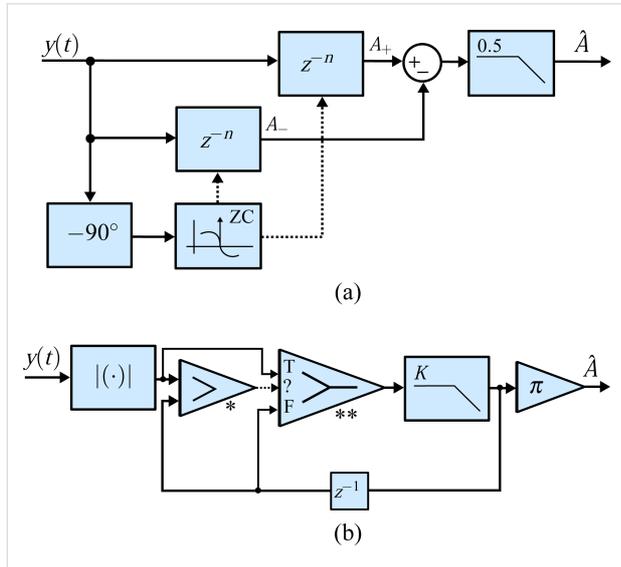


Figure 8: Functional block diagram of (a) the peak hold method and (b) the modified peak hold method based on a peak detector implementation alleviating the sample frequency limitation. The blocks labeled * and ** follow the LabVIEW-specific layout and represent the “greater” and triggered selector functionality.

While this technique offers high tracking bandwidth, it is more susceptible to measurement noise and other frequency components in the signal as there are only two measurements per cycle. Furthermore, the presence of harmonics will result in biased amplitude estimates and complicates the synchronization of the sample and hold circuitry. In a digital implementation, the triggering of the sample and hold blocks can be simplified with a single zero-cross comparator and knowledge of the sample frequency and carrier frequency. However, this approach requires a sufficiently high sample rate to carrier frequency ratio $m = f_s/f_c$ such that the zero-crossing can be detected accurately. Then, knowing that the negative peak will appear at $m/4$ samples after the zero-crossing and the positive peak will appear at $3m/4$ samples after the zero-crossing, m must be at least 4, or any integer multiple.

For the digital system used in this work and the chosen carrier frequency, detecting the zero-crossing with only 6 samples per cycle is infeasible. As such, a modified peak detection method is implemented that does not rely on accurate timing [45]. The block diagram is shown in Figure 8b. The method quickly tracks rising amplitudes due to the comparator and then slowly decreases the estimate based on the low-pass filter gain $0 < K < 1$. We chose $K = 0.5$ throughout this paper and the low-pass filter is used to set the tracking bandwidth.

Coherent demodulator

A digital low-latency, coherent demodulation method has been proposed based on mixing and post-integration over a fixed time window [40-43]. Conceptually, it is an all-digital lock-in amplifier implementation that mixes the signal to be demodulated with in-phase and quadrature sinusoids

$$y_i(t) = y(t) \sin(\omega_c t)$$

$$y_q(t) = y(t) \cos(\omega_c t), \tag{16}$$

and implements the low-pass filtering of the harmonic content with a precise fixed-length numerical integration [41]. If the input signal is a pure sinusoid and the integration period T is chosen to be an integer multiple of the drive signal period, $T = mT_c$, the integrals over $y_i(t)$ and $y_q(t)$ evaluate exactly to the in-phase and quadrature states

$$\frac{1}{2} x_1 = \frac{1}{mT_c} \int_0^{mT_c} y_i(t) dt$$

$$\frac{1}{2} x_2 = \frac{1}{mT_c} \int_0^{mT_c} y_q(t) dt. \tag{17}$$

The functional block diagram of this implementation is shown in Figure 9. Of particular importance is the timing and integration length of this approach. For Equation 17 to hold, the integration period must be an integer multiple of the sampling period $nT_s = mT_c$, where n is the number of samples in the integration. However, for an arbitrary carrier frequency the ratio f_s/f_c is rarely an integer making this condition hard to meet. Therefore, a practical solution is to find the smallest n such that $nT_s \leq mT_c \leq (n + 1)T_s$ and performing a partial integration over the last sampling interval [41]. Such precise control over the integration period is achievable in digital systems, however, the implementation of this method is still challenging.

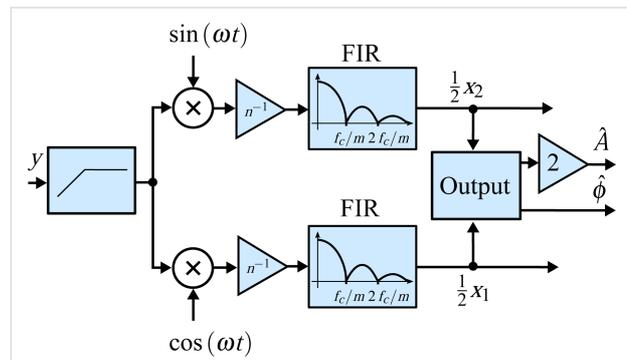
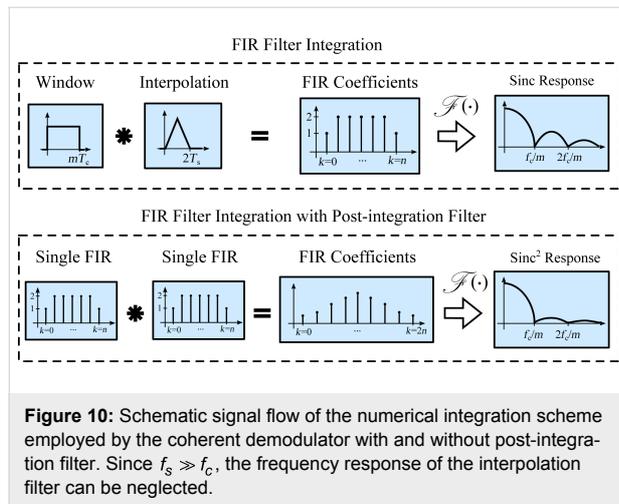


Figure 9: Functional block diagram of the coherent demodulator implementation.

By writing out the sum of the discrete-time integral using a trapezoidal interpolation method [41], it can be seen that the full-period integration can be directly realized with an $(n + 1)$ -tap FIR filter with coefficients $[1, 2, 2, \dots, 2, 1]$ as schematically shown in Figure 10. The impulse response of this FIR filter is naturally obtained from the convolution of the rectangular integration window of length mT_c with a first-order hold element of length $2T_s$. This is equivalent to passing the mixed signal through a sinc filter with side-lobes located at integer multiples of f_c/m . Since f_s is much higher than f_c , the frequency response of the interpolation filter can be neglected.



It is clear that if the integration window length is chosen to be integer multiples of the oscillation period, the sinc filter will have zeros at the harmonics of the oscillation frequency. If the integration period can be made infinitely precise, this approach will remove harmonics from the output. However, as $nT_s = mT_c$ rarely holds, the partial integration is difficult to do precisely and the input signal may contain a DC offset, harmonics will still appear in the output of the integration method. As such, this method can be improved by employing a high-pass filter on the

input and post-integration filters [43], either in the form of a direct notch filter at the second harmonic or by passing the output through another numerical integrator [41]. Intuitively, as Figure 10 illustrates, this method can be viewed as a FIR filter with a triangular impulse response obtained from the convolution of the two FIR integration filters, resulting in a sinc^2 frequency response with significantly reduced side-lobes (for a direct comparison see Appendix B).

While simulation results show that low latency and high tracking bandwidth can be achieved for a pure sinusoid by integrating over one period $M = 1$, in order to reject white noise, multiple oscillation periods must be integrated which reduces the tracking bandwidth and increases the latency [42]. In this work, the trapezoidal numerical integration method with post-integration filters described in [42,43] is directly implemented by cascading two FIR integration filters. Alternatively, the second FIR filter can be replaced by a notch filter at the second harmonic. The computational efficiency of this method can be increased by computing the integral cumulatively, as described in [41,42].

The original work presenting the coherent demodulator integrates over a full period of the fundamental frequency to achieve the highest tracking bandwidth (corresponding to $M = 1$ and $n = 6$ in this work). However, the highest possible tracking bandwidth can be achieved by setting $n = 3$, which still guarantees that the component at $2f_c$ is exactly canceled. A comparison of the original and half-period coherent demodulator is presented in Appendix B.

Summary

Table 1 compares the amplitude estimation techniques discussed in this section. From the classification shown in Figure 2, methods based on rectification can only obtain amplitude estimates while methods based on mixing with an internal reference oscillator can recover both amplitude and phase. Ad-

Table 1: Qualitative summary of amplitude estimation methods stating the tracking bandwidth tuning parameter, ability to determine a phase estimate and timing requirements.

| method | tuning parameter | phase estimate | timing requirement | references |
|-------------------------|------------------|----------------|--------------------|------------|
| lock-in amplifier | LPF f_p | yes | no | [30-32] |
| HBW lock-in amplifier | LPF f_p | yes | no | [35] |
| Kalman filter | Q | yes | no | [44,47,48] |
| Lyapunov filter | γ | yes | no | [45,46,49] |
| moving average filter | # of samples | no | yes | [29,57] |
| mean absolute deviation | LPF f_p | no | yes | [29] |
| peak hold | LPF f_p | no | yes | [36,38] |
| peak detector | LPF f_p | no | no | [45] |
| coherent demodulator | # of samples | yes | yes | [40-43,65] |

ditionally, some of these methods require precise synchronization between the sampling frequency and reference signal. In practice, this requires a single system clock for the sampling time and signal generation. While this property is not a disadvantage when using FPGA-based processing, it does affect the choice of carrier frequencies for the coherent demodulator if the integral is to be precise.

Experimental evaluation

Experimental setup

The aforementioned demodulation techniques were implemented digitally on a common DSP system (National Instruments USB-7855R with Kintex-7 70T FPGA) using dedicated LabVIEW blocks and simple scalar operations. This system was chosen due to its system-oriented graphical design approach, which makes it an accessible FPGA tool without the need for knowledge of hardware description languages.

For a fair comparison and to rule out varying amounts of quantization noise, all demodulation methods are run at a normalized sample frequency of $f_s = 300$ kHz. However, this may not do full justice to the fastest running methods as these techniques might benefit from noise reduction due to oversampling. Additionally, the methods requiring accurate timing will also benefit from more samples per oscillation period.

Implementation complexity

The sample rate achieved by any FPGA implementation, irrespective of the hardware, is a function of the sequential computations which are carried out during each sample period [66]. Therefore the maximum free-running sampling rates, listed below in Table 2, are used to qualitatively compare the implementation complexities.

Due to their simple implementations, the mean absolute deviation method, the peak detector and the moving average filter achieve the highest sampling rates with the mean absolute deviation method approaching the maximum achievable rate of the FPGA system of 1 MHz. The lock-in amplifier, high-bandwidth lock-in amplifier and coherent demodulator achieve the next highest sample rates, while the Lyapunov filter and Kalman filter run at around 300 kHz. Although the Kalman filter is significantly more complex than the Lyapunov filter, the small difference of only 27 kHz can be associated with the highly hardware-optimized implementation of the Kalman filter [44,48], which does not use any continuous states or LabVIEW specific blocks.

Tracking bandwidth

The tracking bandwidth of each demodulator is determined by measuring the amplitude tracking frequency response. This was

performed using a laboratory function generator (Agilent 33521A Waveform Generator) to provide a carrier frequency of $f_c = 50$ kHz, which is amplitude-modulated by a frequency-swept sine signal using the external modulation input. The -3 dB modulation bandwidth of the waveform generator was experimentally verified to be 103.9 kHz, surprisingly low compared to the 30 MHz generator bandwidth but large enough for the carrier frequency used in this experiment. This FM–AM concept directly reveals the low-pass filter characteristic of the demodulators and allows for a direct extraction of the -3 dB tracking bandwidth.

The results are presented in Figure 11 where four different tracking bandwidths are plotted. The maximum achievable tracking bandwidth for each technique is stated below in Table 2. Apart from the tracking bandwidth, the equivalent demodulator filter order (determined from the amplitude reduction per decade for the slowest bandwidth setting) can also be determined from this plot and is stated below in the Summary subsection in Table 2. However, as every demodulator operates nonlinearly, such a classification is only an approximation. As the moving average filter and coherent demodulator are effectively sinc and sinc² filters, we have approximated these by fitting to the local maxima of the side-lobes. From this experiment the linear relationship between the demodulator tuning variable and resulting tracking bandwidth can be obtained, which is discussed in more detail in Appendix C.

It can be seen that the lock-in amplifier and the mean absolute deviation method followed by the peak detector achieve the highest bandwidth, however, at the expense of passing through large $2f_c$ components, which are visible in Figure 21a,f,g. Comparing Figure 21a with Figure 21b, the elimination of the peaks due to the phase cancellation of the high-bandwidth lock-in amplifier is clearly visible.

The Kalman filter achieves a maximum bandwidth of around 50 kHz, which corresponds to tracking within one cycle of the carrier signal without any distortion. However, the Lyapunov filter achieves a slightly higher maximum bandwidth of around 59 kHz but at the expense of peaking at the carrier frequency. This fact is due to the filter recovering the sum of the carrier and the modulating frequency, hence a gain of 2 (6 dB) is measured at 50 kHz.

The FIR filters in the coherent demodulator and the moving average filter implementation cause a characteristic sinc/sinc² frequency response, mathematically originating from the Fourier transform of the integration window. The maximum tracking bandwidth of the coherent demodulator is 39.0 kHz without post-integration filter and 28.6 kHz with post-integration filter.

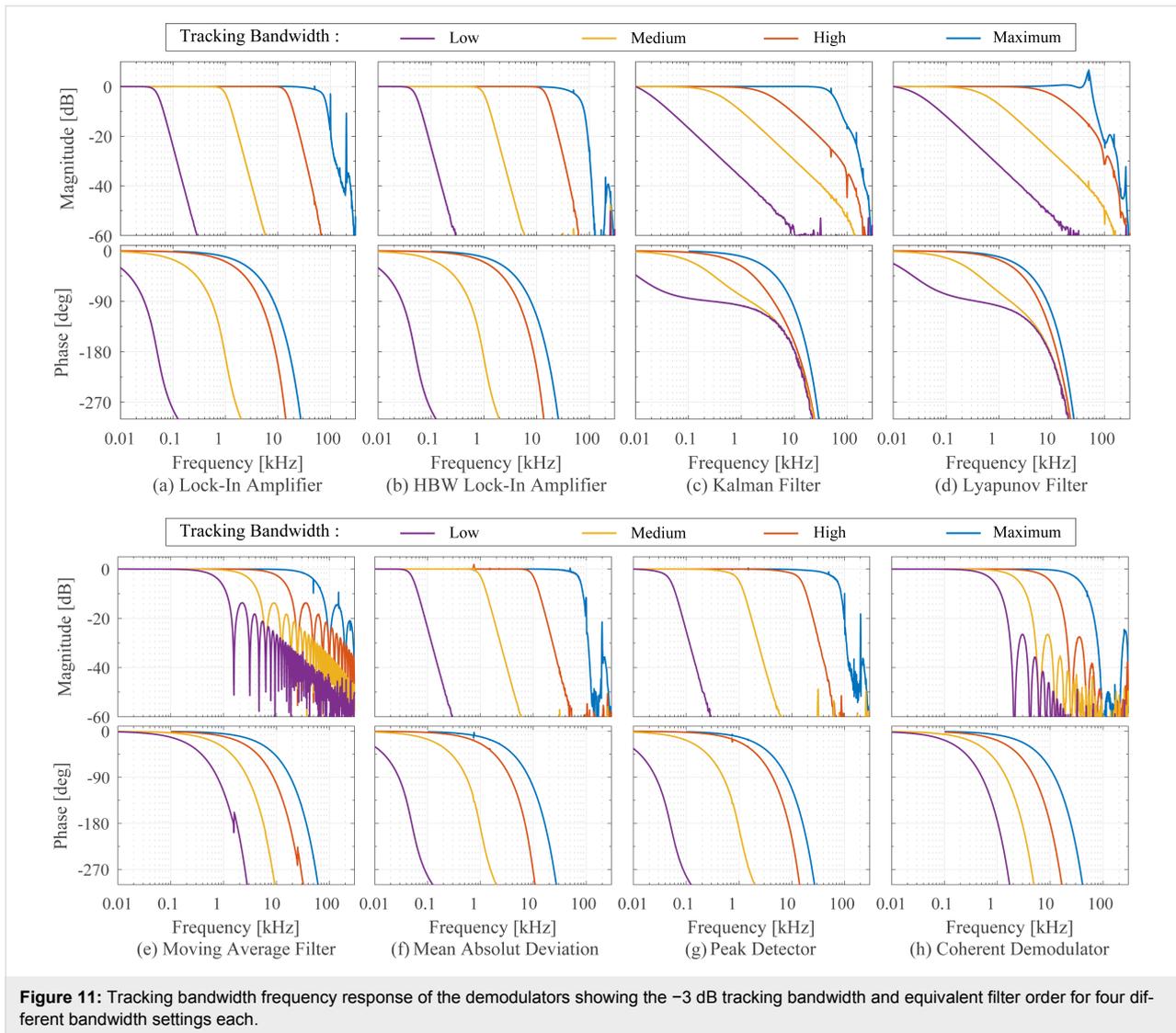


Figure 11: Tracking bandwidth frequency response of the demodulators showing the -3 dB tracking bandwidth and equivalent filter order for four different bandwidth settings each.

These values correlate with the time-domain simulation in [42], which show a convergence after around 1–2 cycles.

Sensitivity to other frequency components

In order to determine the sensitivity to other frequency components present in the signal to be demodulated, a frequency sweep on the carrier signal is performed while the demodulators (where possible) are set to a specific frequency ($f_c = 50$ kHz) and the demodulation bandwidth is set to a fixed value of 1 kHz using the relationships plotted in Appendix C. The resulting plot in Figure 12 shows the attenuation of frequencies other than the modeled carrier frequency and is therefore termed off-mode rejection (OMR). As a quantitative comparison parameter, the OMR is calculated as the gain difference at the modeled frequency (0 dB) and at 40 kHz as highlighted in Figure 12 and stated below in Table 2. In this experiment, the noise floor far away from the modeled frequency

is limited by the residual DC-offset caused by the finite quantization of the digital-to-analog converter (DAC) of the amplitude estimator. However, as all methods are measured with the same hardware, the relative difference is a good indication of maximum achievable off-mode rejection values.

The rectification methods that do not make any assumption on the carrier frequency such as the moving average (MA) filter, mean absolute deviation (MAD) method and peak detector (PD) show a constant gain across all frequencies. In other words, these methods are very sensitive to additional frequency components in the signal as they recover all frequencies equally and are hence impractical for multifrequency AFM.

On the other hand, the lock-in amplifier (LIA) and the high-bandwidth lock-in amplifier (HBW LIA) yield the best off-mode rejection of around -52.0 dB owing to the fourth-order

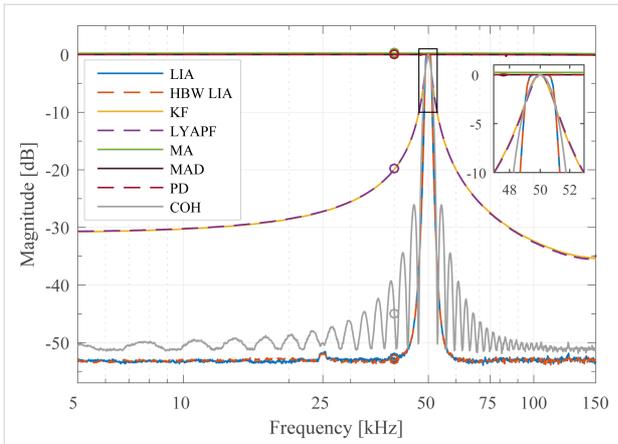


Figure 12: Off-mode rejection of the demodulators for a carrier frequency of $f_c = 50$ kHz and a tracking bandwidth of 1 kHz. The zoom box highlights the intersection at the points $(50 \pm 1$ kHz, -3 dB) and the different filter shapes at the modeled carrier frequency.

Butterworth low-pass filters employed. This result emphasizes the fact that these methods are very insensitive to additional frequency components in the signal and should be used when maximum suppression of these components is of priority.

The Lyapunov filter (LYAPF) and the Kalman filter (KF) yield an off-mode rejection of around -20 dB, significantly lower than the two lock-in amplifier implementations. This fact is due to the equivalent first order response of these filters as shown in Figure 11c,d and stated below in Table 2.

The equivalent order of the coherent demodulator (COH) follows from the envelope of the sinc^2 frequency response. It can be seen that the off-mode rejection is maximized at frequencies corresponding to the zeros of the sinc^2 function. This in turn means that broadband white noise or noise at frequencies other than at these zeros cannot be sufficiently suppressed. This is in contrast to the lock-in amplifier and high-bandwidth lock-in amplifier which show a constant large off-mode rejection away from the carrier frequency.

The off-mode rejection of the Kalman filter and Lyapunov filter can be significantly improved by lowering the tracking bandwidth as shown in Figure 13. In order to achieve a rejection of greater than -40 dB, the bandwidth must be reduced to 100 Hz. On the other hand, the lock-in amplifier only significantly loses its off-mode rejection property at large tracking bandwidths of around 10 kHz. For these large tracking bandwidths, the Butterworth nature of the post-mixing low-pass filters is clearly evident in Figure 13a. The flat region around the modeled frequency where the amplitude is within -3 dB corresponds to twice the tracking bandwidth. The tuning for the Kalman filter is described in Appendix C.

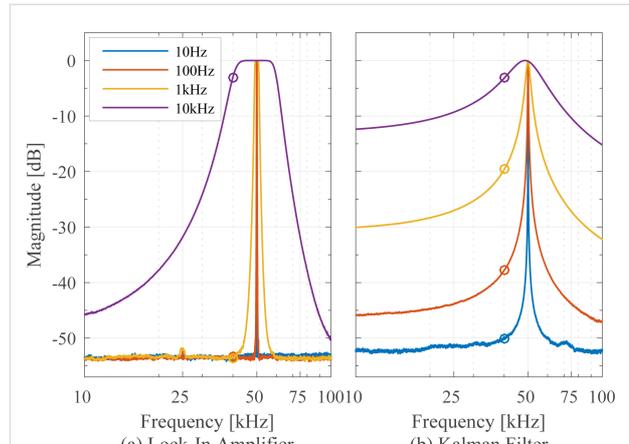


Figure 13: Off-mode rejection of (a) fourth-order lock-in amplifier and (b) first-order Kalman filter for a carrier frequency of $f_c = 50$ kHz at 10 Hz, 100 Hz, 1 kHz, and 10 kHz tracking bandwidth.

Noise evaluation

In order to determine the noise performance, the RMS noise of the amplitude estimate is evaluated as a function of the tracking bandwidth. The responses are compared against the theoretical and experimental response of an “ideal demodulator” represented by a low-pass filtered white noise process. A schematic block diagram of the reference experiment is shown in Figure 14. The band-limited white noise process can be described by a constant power spectral density within the bandwidth, i.e. [67],

$$S_{\text{blwn}}(f) = \begin{cases} A, & |f| \leq f_{\text{wn}} \\ 0, & |f| > f_{\text{wn}} \end{cases}, \quad (18)$$

where f_{wn} is the white noise bandwidth in Hz and A is the power spectral density in V^2/Hz . The RMS noise value σ can be obtained by calculating the total integrated noise (TIN) of the output of a system G driven by a white noise input which is given by [67]

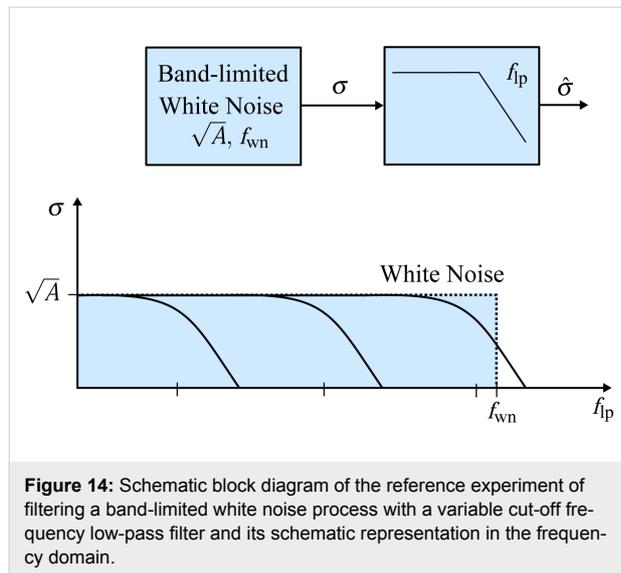
$$\sigma(G) = \sqrt{\int_0^{f_{\text{wn}}} A |G(j\omega)|^2 df}. \quad (19)$$

For the perfect band-limited system (Equation 18), $G = 1$ and Equation 19 simplifies to

$$\sigma = \sqrt{A} \sqrt{f_{\text{wn}}}. \quad (20)$$

The reference curves obtained from this experiment (see Appendix D for details) can be considered as “ideal demodula-

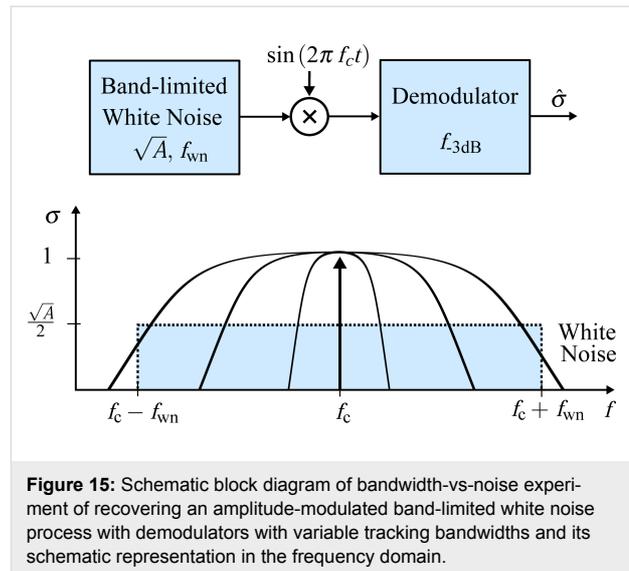
tors” and are compared to the modulated white noise experiment that is schematically shown in Figure 14.



In this experiment, a laboratory function generator (Agilent 33521A Waveform Generator) providing a 40 kHz bandwidth-limited white noise signal with an amplitude of 10 mV_{rms} is first measured directly through a second-order variable cut-off frequency low-pass filter (Stanford Research SR560 Low Noise Preamplifier). The acquisition front end of a micro system analyzer (Polytec MSA-050-3D) is used to capture the time-domain data sampled at $f_s = 2.56$ MHz for $T = 13.11$ s. The TIN is obtained by integrating the noise density estimate from DC to $f_s/2$ using Welch’s method with 16 averages. Subsequently, each demodulator is subjected to amplitude-modulated white noise as shown in Figure 15 with a carrier frequency of 50 kHz and the demodulated amplitude is recorded for several tracking bandwidths in the same manner.

For each demodulator, the results are shown in Figure 16. It can be seen that the lock-in amplifier follows the trend of the reference filtered white noise process for low tracking bandwidths but exhibits an exponentially growing TIN when the tracking bandwidth approaches the carrier frequency. This fact is due to the increasing $2f_c$ component in the amplitude estimate due to inadequate filtering of the mixing product. On the contrary, the high-bandwidth lock-in amplifier does not show this increase owing to the phase cancellation employed. However, the addition of the phase-shifted mixing products increases the noise for lower tracking bandwidths.

The Kalman filter and the Lyapunov filter show an equal trend without any noise increase until the bandwidth reaches the carrier frequency. The Kalman filter never crosses this point. In

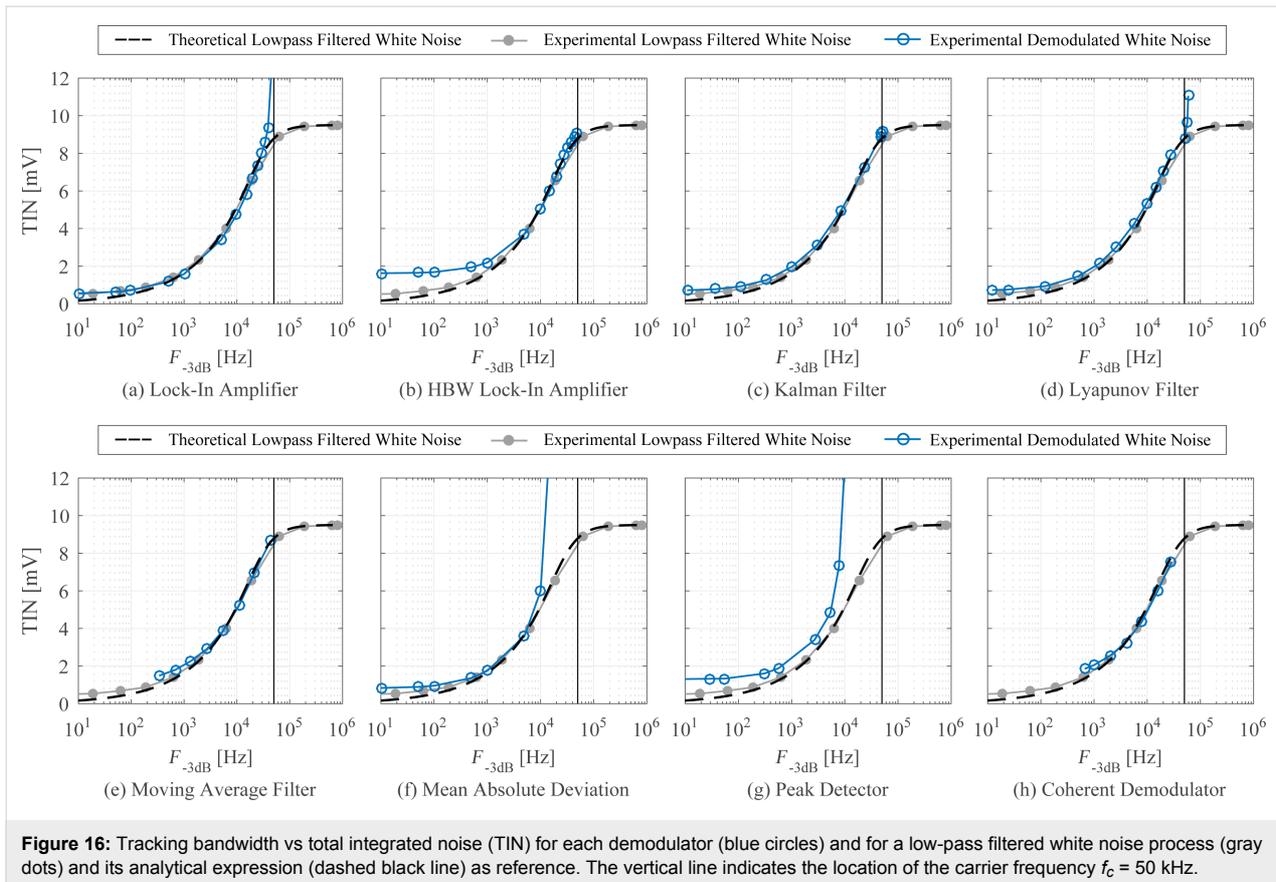


contrast, the succeeding nonlinearity displayed by the Lyapunov filter effectively reduces its useful bandwidth to that of the Kalman filter. The ability of the moving average filter and the coherent demodulator to reach the lower tracking bandwidth frequency range is dictated by the highest-order FIR filter that can be implemented on the FPGA. For the moving average filter this is $n = 384$ and for the coherent demodulator with post-integration filters this limit is $n = 144$ for the LabView hardware used in this work. Lastly, the mean absolute deviation method and the peak detector are constrained by inadequate filtering of the mixing products arising from the absolute value operation, which significantly limits their practical bandwidth.

At frequencies approaching DC, all methods, including the low-pass filtered white noise process, approach a constant value due to digital noise, residual DC-offsets and $1/f$ noise in the signal. This experiment highlights that amplitude noise needs to be taken into account when stating the maximum tracking bandwidth of demodulation methods.

Summary

Table 2 summarizes the results of the amplitude estimation techniques evaluated in this section. The results show that several demodulation methods are able to obtain amplitude estimates in a single cycle, corresponding to a maximum tracking bandwidth $f_{-3dB} = f_c$. However, this figure of merit needs to be assessed with caution as it does not reflect the noise present in the amplitude estimate due to insufficient filtering of mixing products. For instance, the lock-in amplifier can only be used up to 38 kHz (compared to 70 kHz as stated in Table 2) before the harmonic distortion makes this demodulator unusable. While the high-bandwidth lock-in amplifier eliminates this problem, the addition of orthogonal sinusoids increases the noise for low



bandwidths. In contrast, the Kalman filter and Lyapunov filter, despite being of low order, show excellent noise performance over the entire bandwidth of interest. The feasible tracking bandwidth range for each demodulator can be read from Figure 16. The sensitivity to other frequency components is assessed by the off-mode rejection experiment, which measures the ability to reject white noise or other deterministic frequency components. Higher off-mode rejection is achieved by increasing the equivalent order of the demodulator.

AFM imaging

In order to demonstrate the effect of insufficient demodulator bandwidth, a high-speed tapping-mode AFM experiment is conducted with a NT-MDT NTEGRA AFM equipped with a Bruker DMASP piezoelectrically actuated cantilever. Imaging was performed in constant-height mode to circumvent the common z -axis actuator bandwidth limitation. Since the z -axis controller bandwidth is reduced to the point where the sample features entirely appear in the amplitude error image, any

Table 2: Maximum free-running sample frequency f_s , maximum tracking bandwidth f_{-3dB} , equivalent order and off-mode rejection (OMR) at 1 kHz tracking bandwidth of each demodulation technique.

| method | max. f_s [kHz] | max. f_{-3dB} [kHz] | order | OMR [dB] |
|-------------------------|------------------|-----------------------|-------|----------|
| lock-in amplifier | 431 | 70.0 | 4 | 53 |
| HBW lock-in amplifier | 417 | 52.6 | 4 | 53 |
| Kalman filter | 300 | 50.5 | 1 | 20 |
| Lyapunov filter | 327 | 58.7 | 1 | 20 |
| moving average | 580 | 43.3 | 1 | 0 |
| mean absolute deviation | 977 | 70.9 | 4 | 0 |
| peak hold | 580 | 21.7 | 4 | 0 |
| peak detector | 800 | 58.5 | 4 | 0 |
| coherent demodulator | 362 | 28.6 | 2 | 45 |

imaging artifacts are either due to insufficient demodulator or cantilever bandwidth.

In order to render the demodulator the bottleneck, the fundamental resonance at $f_1 \approx 50$ kHz of the DMASP cantilever is heavily damped with model-based quality factor control [68]. The frequency responses from the cantilever actuation to tip displacement for various quality factor controller gains are shown in Figure 17 along side the corresponding tracking bandwidths obtained from drive amplitude modulation. Due to the integrated actuation of the cantilever, the control method achieved a quality factor as low as $Q_1 = 8$, resulting in a tracking bandwidth of 3.3 kHz, adequately matching the first-order approximation $f_1/(2Q_1)$ [36].

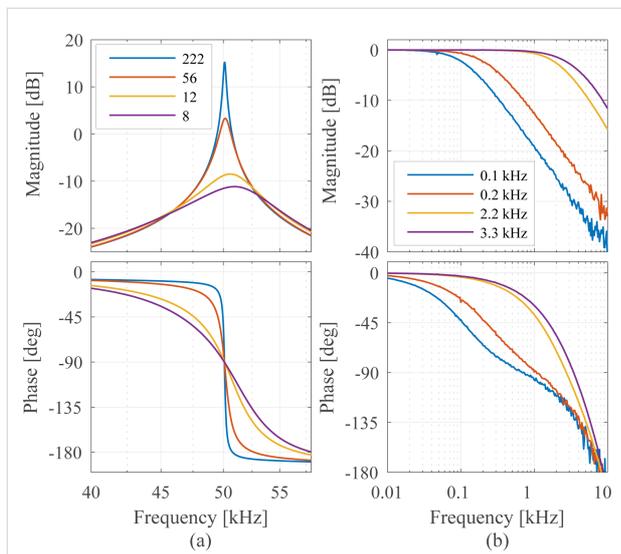


Figure 17: (a) Frequency response of the DMASP cantilever in open-loop (blue) and for various quality factor controller gains to reduce the quality factor as stated in the legend. (b) Tracking bandwidths of the DMASP cantilever determined via drive amplitude modulation for the frequency responses shown in (a) and color-coded accordingly.

AFM images of a calibration grating (NT-MDT TGZ3) with periodic features of height $h \approx 500$ nm were obtained at a speed of 627 $\mu\text{m/s}$ and 1.25 mm/s. Because the scanner rate of the AFM is limited to 31.37 Hz, the scan areas for the two different speeds are, respectively, 10 $\mu\text{m} \times 10 \mu\text{m}$ and 20 $\mu\text{m} \times 20 \mu\text{m}$. The image areas have been cropped to approximately the same region for a better comparison.

The high-speed constant height imaging with the lock-in amplifier, Lyapunov filter and Kalman filter are presented in Figure 18 and Figure 19. Each row corresponds to the same demodulator bandwidth. It can be seen that for small demodulator bandwidths, the sample features are not accurately tracked (first two rows of each figure). By setting a larger demodulator

bandwidth, the sharp sample features are properly tracked, which is clearly evident in the cross-section plots.

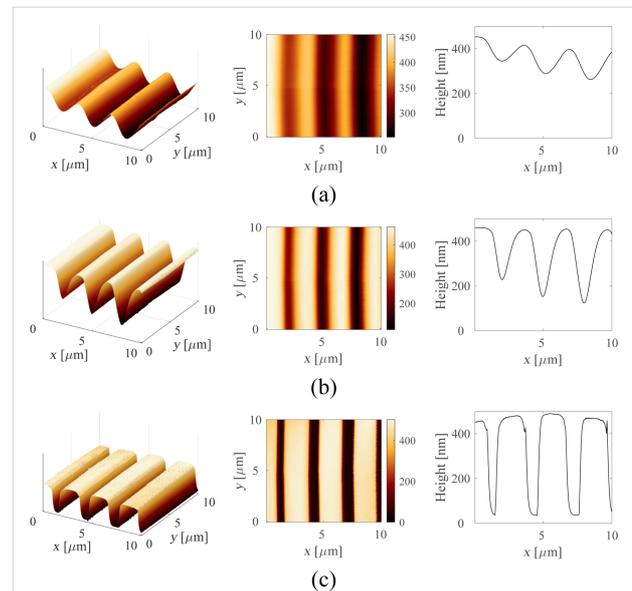


Figure 18: 3D image, 2D image and cross section of amplitude estimates obtained from (a) lock-in amplifier with $f_{lp} = 100$ Hz, (b) lock-in amplifier with $f_{lp} = 200$ Hz, and (c) Lyapunov filter with $\gamma = 60000$ at an imaging speed of 627.45 $\mu\text{m/s}$. The scanning direction is along the positive x- and y-axes.

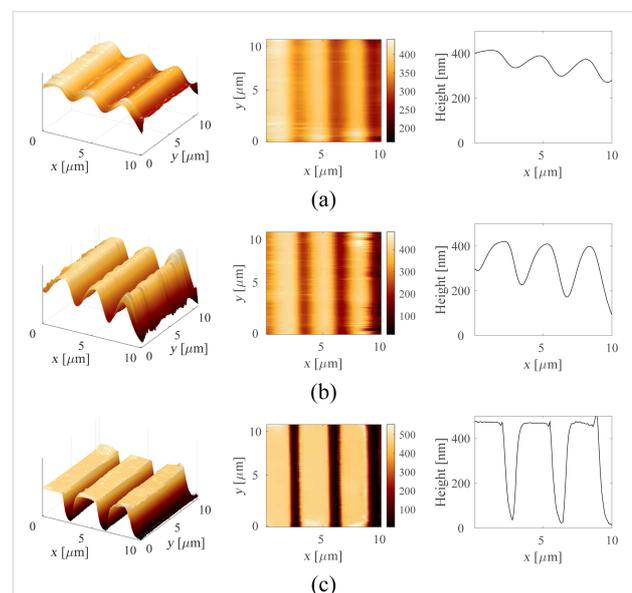


Figure 19: 3D image, 2D image and cross section of amplitude estimates obtained from (a) lock-in amplifier with $f_{lp} = 200$ Hz, (b) lock-in amplifier with $f_{lp} = 300$ Hz, and (c) Kalman filter with $Q = 0.004$, $R = 2$ at an imaging speed of 1.25 mm/s. The scanning direction is along the positive x- and y-axes.

Note that the purpose of the AFM images is to emphasize the need for a fast demodulator bandwidth when all other band-

width limiting components in the AFM loop are eliminated. Therefore, the authors perform imaging in constant-height mode, which entirely removes the z -axis controller and actuator limitation. Due to the 3.3 kHz cantilever bandwidth, the full potential of the fastest methods cannot be utilized and hence the AFM images themselves cannot serve as a means of differentiating between these methods.

Conclusion

This article provides an experimental comparison of the performance of conventional and novel digital demodulation techniques over their entire tracking bandwidth. The techniques include mixing-based methods namely the lock-in amplifier, high-bandwidth lock-in amplifier, coherent demodulator, Kalman filter, and Lyapunov filter, as well as rectification-based methods in the form of a moving average filter, mean absolute deviation computation and peak detection. The performance metrics considered were the tracking bandwidth, implementation complexity, sensitivity to other frequency components and tracking bandwidth vs noise performance.

The $2f_c$ component naturally arises in demodulation schemes employing mixing, which will distort the output if not adequately filtered. The lock-in amplifier relies on general low-pass filters to attenuate these mixing products, limiting the maximum achievable tracking bandwidth. While the high-bandwidth lock-in amplifier eliminates the $2f_c$ component via phase cancellation, it introduces additional noise at low frequencies due to the summation of the phase-shifted signals. The coherent demodulator, being an all-digital lock-in amplifier implementation, eliminates the mixing products by performing a precise numerical integration over a fixed-length time window. While this approach is able to achieve a high tracking bandwidth with minimal latency for short integration windows, a high sample to carrier frequency ratio is crucial for a high-performance implementation. The Kalman filter and the Lyapunov filter on the other hand employ internal feedback of the estimated states to reject the mixing products, which allows them to maximize the tracking bandwidth without introducing additional noise in the amplitude estimate. If maximum suppression of any signal away from the carrier frequency is the priority, the lock-in amplifier can still be regarded as the method of choice as it shows large off-mode rejection and the lowest noise at low tracking bandwidths.

Among the rectification-based methods, the RMS-to-DC conversion methods (mean absolute deviation and moving average filter) have the lowest implementation complexity. Due to their inability to reject unwanted frequency components they can only be used at small tracking bandwidths. Ando's peak hold method requires accurate timing within the digital imple-

mentation and a high sample to carrier frequency ratio to detect the zero-crossing accurately. A modified peak hold method (peak detector) alleviates the sample rate requirement, but insufficient filtering of the absolute value distortion requires low tracking bandwidths.

The above discussion highlights that there exist multiple trade-offs. Although there are many possible application goals, three of the most common are listed below along with the recommended demodulator.

1. Maximum bandwidth: The Kalman filter provides maximum tracking bandwidth without introducing excess noise or distortion. However, the Lyapunov filter performs comparably but is significantly simpler to implement.
2. Maximum noise suppression: The lock-in amplifier provides maximum off-mode rejection when the tracking bandwidth is low enough to avoid ripple.
3. Minimum implementation complexity: The RMS-to-DC conversion methods are simplest to implement but are very sensitive to other frequency components as they do not provide any off-mode rejection.

Appendix

A Kalman filter equations

The recursive implementation of the Kalman filter equations follows [62,63] by iterating between the prediction step

$$\begin{aligned}\hat{\mathbf{x}}_k^- &= \hat{\mathbf{x}}_{k-1}^- \\ \mathbf{P}_k^- &= \mathbf{P}_{k-1}^- + \mathbf{Q}\end{aligned}\quad (21)$$

and the measurement update step by calculating the Kalman gain \mathbf{k}_k

$$\mathbf{k}_k = \mathbf{P}_k^- \mathbf{c}_k^T \left(\mathbf{c}_k \mathbf{P}_k^- \mathbf{c}_k^T + R \right)^{-1} \quad (22)$$

The estimated states must then be corrected

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{k}_k \left(y_k - \mathbf{c}_k \hat{\mathbf{x}}_k^- \right), \quad (23)$$

and the covariance matrix can be updated with

$$\mathbf{P}_k = \left(\mathbf{I} - \mathbf{k}_k \mathbf{c}_k \right) \mathbf{P}_k^- \left(\mathbf{I} - \mathbf{k}_k \mathbf{c}_k \right)^T + \mathbf{k}_k R \mathbf{k}_k^T. \quad (24)$$

The main computations in Equations 22–24 are graphically represented by the block diagram shown in Figure 20. Due to

the time-varying system representation, the calculations in the prediction steps (Equation 21) are heavily simplified, benefiting a high-bandwidth FPGA implementation that can be realized with scalar operations [48].

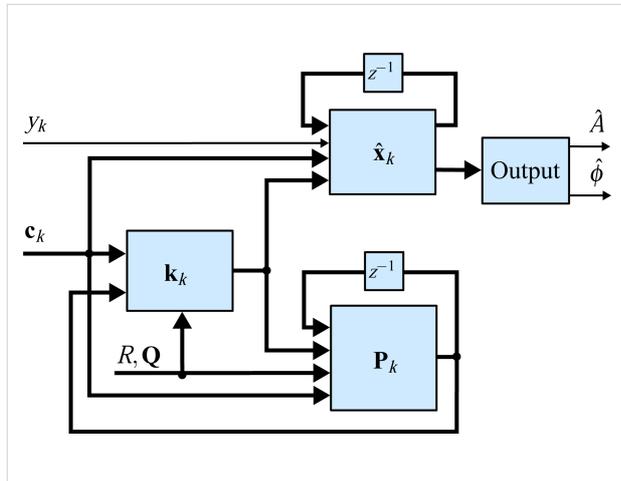


Figure 20: Functional block diagram of the Kalman filter implementation. Thick lines indicate vector-valued signal paths and thin lines indicate scalar signal paths.

B Comparison of coherent demodulator methods

Figure 21 shows a direct comparison of the coherent demodulator and half-period coherent demodulator using a single FIR integration filter (Single FIR) and with an additional post-integration filter (Double FIR). A higher attenuation at integer multiples of the carrier frequency due to the sinc² frequency response of the latter is responsible for the reduction of harmonics in the output of the demodulator. Notice, that this approach naturally comes at the expense of tracking bandwidth as is visible in the magnitude response in Figure 21a. For the half-period coherent demodulator ($n = 3$), the single FIR integration filter approach yields a -3 dB tracking bandwidth of 39.0 kHz, the addition of the post-integration filter reduces this bandwidth to 28.6 kHz. For the full-period coherent demodulator ($n = 6$), the single FIR integration filter approach yields a -3 dB tracking bandwidth of 21.6 kHz, the addition of the post-integration filter reduces this bandwidth to 15.6 kHz. The increased latency is also clearly visible from the phase responses in Figure 21a,b. On the other hand, for a fixed tracking bandwidth of 1 kHz, the addition of the post-integration FIR filter improves the off-mode rejection drastically as is visible in Figure 21c because of the faster roll-off of the equivalent sinc² filter.

C Demodulator tuning

By plotting the tuning parameter against the experimentally determined tracking bandwidth in Figure 22, the region of linear

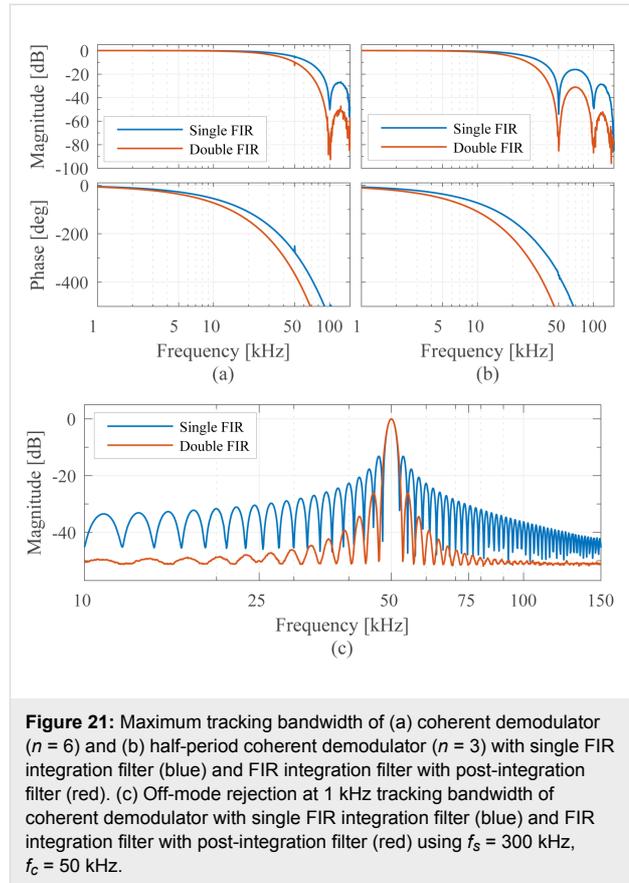


Figure 21: Maximum tracking bandwidth of (a) coherent demodulator ($n = 6$) and (b) half-period coherent demodulator ($n = 3$) with single FIR integration filter (blue) and FIR integration filter with post-integration filter (red). (c) Off-mode rejection at 1 kHz tracking bandwidth of coherent demodulator with single FIR integration filter (blue) and FIR integration filter with post-integration filter (red) using $f_s = 300$ kHz, $f_c = 50$ kHz.

relationship is determined. For the lock-in amplifier, high-bandwidth lock-in amplifier, mean absolute deviation method, and peak detector the tuning variable is the low-pass filter (LPF) cut-off frequency f_{lp} . With a known measurement noise covariance R , the Kalman filter is tuned based on the assumed covariance Q and the Lyapunov filter can be tuned by setting the integrator gain γ . The moving average filter and the coherent demodulator tuning is achieved by setting the amount of samples per integration window n .

The methods achieve a near perfect linear relationship across the entire bandwidths tested on a double logarithmic scale. At the upper range of the tracking bandwidths, the relationship function for the Kalman filter flattens out revealing the proximity to the bandwidth limitation. Also noticeable is a slight deviation from the linear trend for the Lyapunov filter associated with the peaking at the carrier frequency.

Note, that this specific plot is only accurate for the particular hardware and sample frequency chosen in this work. However, it is useful in determining a particular tuning setting necessary for a given tracking bandwidth for each demodulator. For example, a 1 kHz tracking bandwidth for the Kalman filter is achieved by $Q = \text{diag}(10^{-3}, 10^{-3})$.

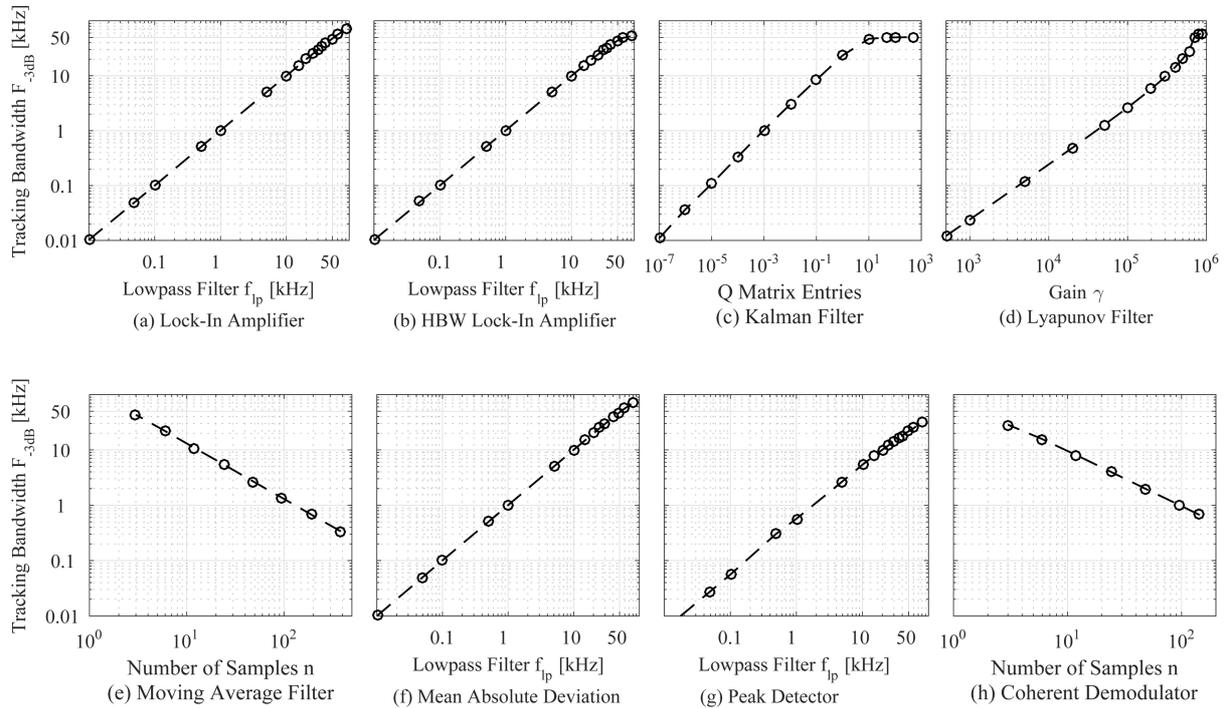


Figure 22: Relationship between demodulator tuning variable and achievable tracking bandwidth.

D Low-pass filtered white noise

Substituting a first-order low-pass filter system with cut-off frequency f_{lp}

$$G_1(s) = \frac{2\pi f_{lp}}{s + 2\pi f_{lp}} \quad (25)$$

into the general expression for low-pass filtered white noise (Equation 19) yields

$$\begin{aligned} \sigma(G_1) &= \sqrt{\int_0^{f_{wn}} A \frac{f_{lp}^2}{f_{lp}^2 + f^2} df} \\ &= \sqrt{A} \sqrt{f_{lp} \tan^{-1}\left(\frac{f_{wn}}{f_{lp}}\right)}. \end{aligned} \quad (26)$$

If the white noise bandwidth is much larger than the cut-off frequency $f_{wn} \gg f_{lp}$, Equation 26 can be simplified to

$$\sigma(G_1) = \sqrt{A} \sqrt{1.57 f_{lp}}. \quad (27)$$

Similarly, for a second-order low-pass filter system of the form

$$G_2(s) = \frac{(2\pi f_{lp})^2}{(s + 2\pi f_{lp})^2}, \quad (28)$$

the total integrated noise evaluates to

$$\begin{aligned} \sigma(G_2) &= \sqrt{\int_0^{f_{wn}} A \frac{f_{lp}^4}{(f_{lp}^2 - f^2)^2 + (2f f_{lp})^2} df} \\ &= \sqrt{A} \sqrt{0.5 f_{lp} \tan^{-1}\left(\frac{f_{wn}}{f_{lp}}\right) + 0.5 \frac{f_{lp}^2 f_{wn}}{f_{lp}^2 + f_{wn}^2}}. \end{aligned} \quad (29)$$

If the white noise bandwidth is much larger than the cut-off frequency $f_{wn} \gg f_{lp}$, Equation 29 simplifies to

$$\sigma(G_2) = \sqrt{A} \sqrt{0.5 \cdot 1.57 f_{lp}}. \quad (30)$$

In Figure 23, the analytical Equations 26 and 29 are plotted against experimental low-pass filtered white noise processes. For this purpose, a function generator (Agilent 33500B Wave-

form Generator) was used to generate 40 kHz bandwidth limited white noise, which was subsequently low-pass filtered with a low-noise voltage preamplifier with variable cut-off frequency (Stanford Research SR560). The output of the filter was captured in the time-domain, sampled at $f_s = 2.56$ MHz for $T = 13.11$ s, with the acquisition front end of a micro system analyzer (Polytec MSA-050-3D). The total integrated noise, is obtained by integrating the noise density from 0 to $f_s/2$ using Welch's method. It can be seen that for both systems, theory and experiment match very well. In theory the total integrated noise approaches zero as the filter bandwidth goes to zero. This discrepancy between theory and experiment can be attributed to digital noise, residual DC-offsets and $1/f$ noise in the signal.

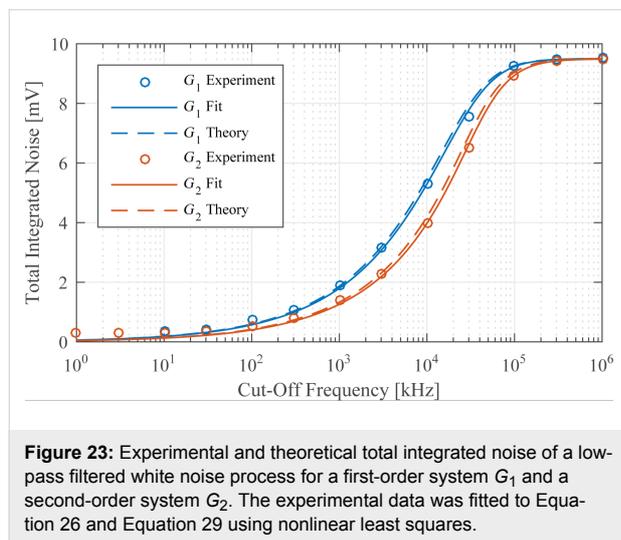


Figure 23: Experimental and theoretical total integrated noise of a low-pass filtered white noise process for a first-order system G_1 and a second-order system G_2 . The experimental data was fitted to Equation 26 and Equation 29 using nonlinear least squares.

Acknowledgements

This research was performed at The University of Newcastle, Callaghan, NSW, Australia.

References

- Haykin, S. *Communication systems*; John Wiley & Sons, 2008.
- Nyce, D. S. *Linear position sensors: theory and application*; John Wiley & Sons, 2004.
- Baxter, L. K. *Capacitive sensors: Design and Applications*; Wiley-IEEE Press, 1996.
- Fleming, A. J. *Sens. Actuators, A* **2013**, *190*, 106–126. doi:10.1016/j.sna.2012.10.016
- Tsui, D. C.; Stormer, H. L.; Gossard, A. C. *Phys. Rev. Lett.* **1982**, *48*, 1559–1562. doi:10.1103/PhysRevLett.48.1559
- Rugar, D.; Grütter, P. *Phys. Rev. Lett.* **1991**, *67*, 699–702. doi:10.1103/PhysRevLett.67.699
- van Trees, H. L. *Detection, Estimation, and Modulation Theory, Part IV, Optimum Array Processing*; John Wiley & Sons, 2004.
- Binnig, G.; Quate, C. F.; Gerber, C. *Phys. Rev. Lett.* **1986**, *56*, 930–933. doi:10.1103/PhysRevLett.56.930
- García, R.; Pérez, R. *Surf. Sci. Rep.* **2002**, *47*, 197–301. doi:10.1016/S0167-5729(02)00077-8
- García, R.; Magerle, R.; Perez, R. *Nat. Mater.* **2007**, *6*, 405–411. doi:10.1038/nmat1925
- Martin, Y.; Williams, C. C.; Wickramasinghe, H. K. *J. Appl. Phys.* **1987**, *61*, 4723–4729. doi:10.1063/1.338807
- Minne, S. C.; Soh, H. T.; Flueckiger, P.; Quate, C. F. *Appl. Phys. Lett.* **1995**, *66*, 703–705. doi:10.1063/1.114105
- Albrecht, T.; Grütter, P.; Home, D.; Rugar, D. *J. Appl. Phys.* **1991**, *69*, 668–673. doi:10.1063/1.347347
- Fukuma, T.; Kilpatrick, J. I.; Jarvis, S. P. *Rev. Sci. Instrum.* **2006**, *77*, 123703. doi:10.1063/1.2405361
- Kodera, N.; Yamamoto, D.; Ishikawa, R.; Ando, T. *Nature* **2010**, *468*, 72–76. doi:10.1038/nature09450
- Fantner, G. E.; Barbero, R. J.; Gray, D. S.; Belcher, A. M. *Nat. Nanotechnol.* **2010**, *5*, 280–285. doi:10.1038/nnano.2010.29
- Uchihashi, T.; Iino, R.; Ando, T.; Noji, H. *Science* **2011**, *333*, 755–758. doi:10.1126/science.1205510
- Casuso, I.; Khao, J.; Chami, M.; Paul-Gilloteaux, P.; Husain, M.; Duneau, J.-P.; Stahlberg, H.; Sturgis, J. N.; Scheuring, S. *Nat. Nanotechnol.* **2012**, *7*, 525–529. doi:10.1038/nnano.2012.109
- García, R.; Herruzo, E. T. *Nat. Nanotechnol.* **2012**, *7*, 217–226. doi:10.1038/nnano.2012.38
- Martínez, N. F.; Lozano, J. R.; Herruzo, E. T.; Garcia, F.; Richter, C.; Sulzbach, T.; Garcia, R. *Nanotechnology* **2008**, *19*, 384011. doi:10.1088/0957-4484/19/38/384011
- Herruzo, E. T.; Perrino, A. P.; Garcia, R. *Nat. Commun.* **2014**, *5*, 3126. doi:10.1038/ncomms4126
- Solares, S. D.; An, S.; Long, C. J. *Beilstein J. Nanotechnol.* **2014**, *5*, 1637–1648. doi:10.3762/bjnano.5.175
- Stark, R. W.; Heckl, W. M. *Rev. Sci. Instrum.* **2003**, *74*, 5111–5114. doi:10.1063/1.1626008
- Sahin, O.; Magonov, S.; Su, C.; Quate, C. F.; Solgaard, O. *Nat. Nanotechnol.* **2007**, *2*, 507–514. doi:10.1038/nnano.2007.226
- Raman, A.; Trigueros, S.; Cartagena, A.; Stevenson, A. P. Z.; Susilo, M.; Nauman, E.; Antoranz Contera, S. *Nat. Nanotechnol.* **2011**, *6*, 809–814. doi:10.1038/nnano.2011.186
- Platz, D.; Tholén, E. A.; Pesen, D.; Haviland, D. B. *Appl. Phys. Lett.* **2008**, *92*, 153106. doi:10.1063/1.2909569
- Dick, A. J.; Solares, S. D. *J. Comput. Nonlinear Dyn.* **2010**, *6*, 031005. doi:10.1115/1.4002341
- Forchheimer, D.; Forchheimer, R.; Haviland, D. B. *Nat. Commun.* **2015**, *6*, 6270. doi:10.1038/ncomms7270
- Kitchin, C.; Counts, L. *RMS to DC Conversion Application Guide*; Analog Devices, Inc., 1986.
- Cosens, C. R. *Proc. Phys. Soc., London* **1934**, *46*, 818. doi:10.1088/0959-5309/46/6/310
- Michels, W. C.; Curtis, N. L. *Rev. Sci. Instrum.* **1941**, *12*, 444–447. doi:10.1063/1.1769919
- Morris, E. D.; Johnston, H. S. *Rev. Sci. Instrum.* **1968**, *39*, 620–621. doi:10.1063/1.1683462
- Razavi, B. Architectures and circuits for RF CMOS receivers. In *Proceedings of the IEEE, Custom Integrated Circuits Conference*, 1998, May 14, 1998; IEEE, 1998; pp 393–400. doi:10.1109/CICC.1998.695005
- Karvinen, K. S.; Moheimani, S. O. R. *Mechatronics* **2014**, *24*, 661–671. doi:10.1016/j.mechatronics.2013.11.011
- Karvinen, K. S.; Moheimani, S. O. R. *Rev. Sci. Instrum.* **2014**, *85*, 023707. doi:10.1063/1.4865841
- Ando, T.; Uchihashi, T.; Fukuma, T. *Prog. Surf. Sci.* **2008**, *83*, 337–437. doi:10.1016/j.progsurf.2008.09.001

37. Ando, T. *Nanotechnology* **2012**, *23*, 062001. doi:10.1088/0957-4484/23/6/062001
38. Ando, T.; Koderá, N.; Takai, E.; Maruyama, D.; Saito, K.; Toda, A. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 12468–12472. doi:10.1073/pnas.211400898
39. Uchihashi, T.; Ando, T.; Yamashita, H. *Appl. Phys. Lett.* **2006**, *89*, 213112. doi:10.1063/1.2387963
40. Kokavecz, J.; Tóth, Z.; Horváth, Z.; Heszler, P.; Mechler, Á. *Nanotechnology* **2006**, *17*, S173. doi:10.1088/0957-4484/17/7/S12
41. Abramovitch, D. Y. Coherent Demodulation with Reduced Latency Adapted for use in Scanning Probe Microscopes. US7843627B2, Nov 30, 2010.
42. Abramovitch, D. Y. Low latency demodulation for Atomic Force Microscopes, Part I efficient real-time integration. In *American Control Conference (ACC)*, June 29–July 1, 2011; IEEE, 2011; pp 2252–2257. doi:10.1109/ACC.2011.5991144
43. Abramovitch, D. Y. Low latency demodulation for atomic force microscopes, Part II: Efficient calculation of magnitude and phase. In *Proceedings of the 18th IFAC World Congress*, 2011; pp 12721–12726. doi:10.3182/20110828-6-it-1002.00869
44. Ruppert, M. G.; Karvinen, K. S.; Wiggins, S. L.; Reza Moheimani, S. O. *IEEE Trans. Control Syst. Technol.* **2016**, *24*, 276–284. doi:10.1109/TCST.2015.2435654
45. Ragazzon, M. R. P.; Gravidahl, J. T.; Fleming, A. J. On Amplitude Estimation for High-Speed Atomic Force Microscopy. In *American Control Conference (ACC)*, July 6–8, 2016; IEEE, 2016; pp 2635–2642. doi:10.1109/ACC.2016.7525314
46. Ragazzon, M. R. P.; Ruppert, M. G.; Harcombe, D. M.; Fleming, A. J.; Gravidahl, J. T. *IEEE Trans. Control Syst. Technol.* **2017**, *PP*, 1–8. doi:10.1109/TCST.2017.2692721
47. Ruppert, M. G.; Harcombe, D. M.; Moheimani, S. O. R. State estimation for high-speed multifrequency atomic force microscopy. In *American Control Conference (ACC)*, July 6–8, 2016; IEEE, 2016; pp 2617–2622. doi:10.1109/acc.2016.7525311
48. Ruppert, M. G.; Harcombe, D. M.; Moheimani, S. O. R. *IEEE/ASME Trans. Mechatronics* **2016**, *21*, 2705–2715. doi:10.1109/TMECH.2016.2574640
49. Harcombe, D. M.; Ruppert, M. G.; Fleming, A. J. Higher-harmonic AFM Imaging with a High-Bandwidth Multifrequency Lyapunov Filter. In *Advanced Intelligent Mechatronics*, IEEE International Conference on Advanced Intelligent Mechatronics (AIM); 2017.
50. Sahoo, D. R.; Sebastian, A.; Salapaka, M. V. *Appl. Phys. Lett.* **2003**, *83*, 5521–5523. doi:10.1063/1.1633963
51. Sebastian, A.; Sahoo, D. R.; Salapaka, M. V. An observer based sample detection scheme for atomic force microscopy. In *Proceedings of the IEEE Conference on Decision and Control*, 2003; pp 2132–2137.
52. Sahoo, D. R.; Sebastian, A.; Salapaka, M. V. *Int. J. Robust Nonlinear Control* **2005**, *15*, 805–820. doi:10.1002/rnc.1025
53. Sahoo, D. R.; Agarwal, P.; Salapaka, M. V. Transient Force Atomic Force Microscopy: A New Nano-Interrogation Method. In *American Control Conference (ACC)*, 2007, 2007; pp 2135–2140.
54. Karvinen, K. S.; Ruppert, M. G.; Mahata, K.; Moheimani, S. O. R. *IEEE Trans. Nanotechnol.* **2014**, *13*, 1257–1265. doi:10.1109/TNANO.2014.2360878
55. Volder, J. E. *IRE Trans. Electron. Comput.* **1959**, *EC-8*, 330–334. doi:10.1109/TEC.1959.5222693
56. Lau, K.; Quevedo, D.; Vautier, B.; Goodwin, G.; Moheimani, S. *Control Eng. Practice* **2007**, *15*, 377–388. doi:10.1016/j.conengprac.2005.09.004
57. Oppenheim, A. V.; Schaffer, R. W. *Discrete-time signal processing*; Prentice-Hall, 1989.
58. Kalman, R. E. *Trans. ASME* **1960**, *82*, 35–45.
59. Anderson, B. D. O.; Moore, J. B. *Optimal filtering*; Dover Publications, 1979.
60. Zarchan, P.; Musoff, H. *Fundamentals of Kalman Filtering: A Practical Approach*; American Institute of Aeronautics and Astronautics, Inc., 2000.
61. Girgis, A. A.; Hwang, T. L. *IEEE Trans. Power Appar. Syst.* **1984**, *PAS-103*, 2943–2951. doi:10.1109/TPAS.1984.318297
62. Brown, R. G.; Hwang, P. Y. C. *Introduction to Random Signals and Applied Kalman Filterings*; John Wiley & Sons, 1997.
63. Simon, D. *Optimal State Estimation: Kalman, H-Infinity, And Nonlinear Approaches*; John Wiley & Sons, 2006.
64. Ioannou, P. A.; Sun, J. *Robust adaptive control*; Dover Publications, Inc., 2012.
65. Pawłowski, S.; Dobiński, G.; Szmaja, W.; Majcher, A.; Smolny, M. *Sens. Actuators, A* **2015**, *228*, 125–132. doi:10.1016/j.sna.2015.03.015
66. Monmasson, E.; Cirstea, M. N. *IEEE Trans. Ind. Electron.* **2007**, *54*, 1824–1842. doi:10.1109/TIE.2007.898281
67. Fleming, A. J.; Leang, K. K. *Design, Modeling and Control of Nanopositioning Systems*; Springer, 2014.
68. Ruppert, M. G.; Moheimani, S. O. R. *IEEE Trans. Control Syst. Technol.* **2016**, *24*, 1149–1159. doi:10.1109/TCST.2015.2478077

License and Terms

This is an Open Access article under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The license is subject to the *Beilstein Journal of Nanotechnology* terms and conditions: (<http://www.beilstein-journals.org/bjnano>)

The definitive version of this article is the electronic one which can be found at: [doi:10.3762/bjnano.8.142](https://doi.org/10.3762/bjnano.8.142)