NanoScience and Technology

Bert Voigtländer

Atomic Force Microscopy

Second Edition

Springer

NanoScience and Technology

The series NanoScience and Technology is focused on the fascinating nano-world, mesoscopic physics, analysis with atomic resolution, nano and quantum-effect devices, nanomechanics and atomic-scale processes. All the basic aspects and technology-oriented developments in this emerging discipline are covered by comprehensive and timely books. The series constitutes a survey of the relevant special topics, which are presented by leading experts in the field. These books will appeal to researchers, engineers, and advanced students.

More information about this series at http://www.springer.com/series/3705

Bert Voigtländer

Atomic Force Microscopy

Second Edition

Springer

Bert Voigtländer
PGI-3
Forschungszentrum Jülich
Jülich, Germany

RWTH Aachen
Lehrstuhl für Experimentalphysik IV A
Aachen, Germany

ISSN 1434-4904 ISSN 2197-7127 (electronic)
NanoScience and Technology
ISBN 978-3-030-13653-6 ISBN 978-3-030-13654-3 (eBook)
https://doi.org/10.1007/978-3-030-13654-3

The first edition (2015, ISBN 978-3-662-45240-0) was published under the title, "Scanning Probe Microscopy".

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

For Ortraud

Preface

The aim of this textbook is to introduce *Atomic Force Microscopy* to graduate students and others wishing to learn about the subject from fundamental principles. The original literature is fascinating but hard going for a newcomer; I had a hard time trying to understand it myself. Therefore, this textbook was written in an attempt to save other people's time by explaining the topics in a more easily digestible manner.

The first chapter of this book covers instrumental aspects and summarizes some basics like the harmonic oscillator and electronics. When discussing atomic force microscopy in the subsequent chapters, the book concentrates on the principles of the methods. This book arose from my previous book on scanning probe microscopy and represents a substantial extension and revision of the part on AFM of the prior book.

This book developed from a series of lectures which I gave at RWTH Aachen University. To this end, it is mainly written with graduate students in mind. However, since the treatment in the book goes more into greater depth than is possible in a lecture, it is my hope that it will also be useful for professionals in the field and may serve as a reference book in AFM laboratories.

This textbook is not a historical survey of the field, and no content in this book is originally from me. I learned everything from the primary and secondary literature and then reformulated it continuously in the course of teaching the subject. I was largely able to resist including my own research in this book, so it does not include any studies of epitaxy using the scanning tunneling microscope which I performed over the past years, and no charge transport measurements at the nanoscale using multi-tip scanning tunneling microscopy, which is my current research topic.

First of all, I would like to thank Vasily Cherepanov for his careful preparation of most of the figures. Moreover, he was regularly my “sparring partner” when discussing issues which were not clear to me. These discussions helped me a lot in furthering my understanding. I would like to thank Gerhard Meyer, who introduced me to scanning probe microscopy in 1990 and has helped me since then in various circumstances. Also many thanks to Josef Myslivecek for explaining the lock-in technique to me so clearly that I included it here in exactly the way he explained it to me. Irek Morawski introduced me to the FM-AFM technique and the quartz sensors.

I would like to thank my former students Anna Strozecka, Stefan Korte, Martin Scheufens, Martin Lanius, Marcus Blab, Sven Just, Richard Spiegelberg, Felix Lüpke, and Arthur Leis for intense discussions on various topics and for supplying material from their work. I am grateful to Helmut Stollwerk and Peter Coenen for their continuous support over the years.

I would also like to thank my son Felix for his help in typesetting some of the equations in LaTeX. My son Paul helped me to solve some equations using a computer algebra system.

I would like to stay in contact with readers via the webpage www.mprobes.com/AFMbook. On this page, supplementary material as well as errata will be posted.

Finally, it is my hope that this book will enable the reader to operate an atomic force microscope successfully and understand the data obtained with the microscope.

Jülich/Aachen, Germany Bert Voigtländer

Contents

Chapter 1
Introduction

In many areas of science and technology there is a trend toward the nanoscale or even the atomic level. For instance, electronics is already undergoing a transition from microelectronics to nanoelectronics. As transistors with critical dimensions in the the single digit nanometer range are now in production, consumer electronics products contain now real nanoelectronic devices. Also in many other areas the progress toward the nanoscale is under way.

An additional reason for the trend toward the nano/atomic scale is that material properties are ultimately determined by the atomic structure. In order to understand material properties fundamentally it is necessary to go down to the nano or atomic scale. However, since the atoms are very small 60 years ago most people thought that it will probably never be possible to have direct access to materials on this scale (Fig. 1.1).

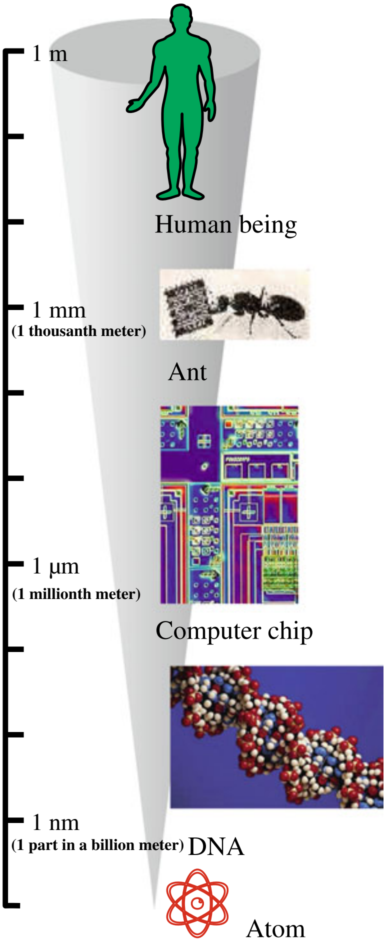
The founding father of nanoscience and nanotechnology was R.P. Feynman. In a visionary talk in 1959 he postulated the possibility of nanotechnology down to the very atoms. In his talk entitled "There is Plenty of Room at the Bottom" he did not use the word "nanotechnolgy" [1] since it had not been coined but he had the idea. This was very visionary in 1959 and he was not really certain so he phrased his vision in rhetorical questions and added some conditions. He reassured himself with his words [2]:

> But I am not afraid to consider the final question as to whether, ultimately – in the great future – we can arrange the atoms the way we want; the very atoms, all the way down!
>
> … when we have some control of the arrangement of things on the small scale we will get an enormously greater range of possible properties that substances can have, and of different things that we can do.
>
> What could we do with layered structures with just the right layers? What would the properties of materials be if we could really arrange the atoms the way we want them?

Feynman saw the potential of nanotechnology already in 1959 before anyone else did. Now 60 years later it is interesting to see how many of his predictions have been

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3_1

Fig. 1.1 Logarithmic size scale from the human to the atom



realized. In some cases things have been realized in a much simpler fashion than he envisaged. To position things on the nanoscale he envisaged a cascade of machines of decreasing size, each driving the next smallest one. As was discovered in 1990, it is possible to go all the way down to the nanoscale and build structures out of atoms in just one step from the macroscale to the atomic scale using a scanning tunneling microscope (STM) [3–6]. As an example Fig. 1.2 shows the word NANO built from single C_{60} molecules using an STM.

Feynman envisaged that nanotechnolgy is possible in principle and would be very useful, but at that time the technology for imaging and controlling matter at the nanoscale had not been invented. With improvements in electron microscopy, it became possible to image matter on the nanoscale. After the invention of scanning probe microscopes those microscopes became quickly another important method for nanoscale imaging. In scanning probe microscopy, a small probe is used to detect

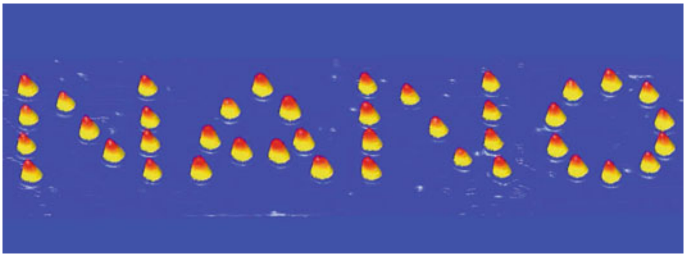


Fig. 1.2 The word NANO assembled from single C_{60} molecules by lateral motion with a scanning tunneling microscope (letter size: $15 \times 15\,\text{nm}^2$)

the local properties at a surface or interface down to nanometer/atomic resolution. By scanning a grid of points on the surface, the detected properties can be mapped and are usually represented as an image. Most frequently in image of the topography of the sample is generated, however, also images of several other properties can be acquired. Because of the scanning mechanism, all these techniques are summarized as scanning probe microscopes (SPM). If the interaction between the probe and the substrate is strong enough the substrate can be modified on the nanoscale.

The most striking property of this kind of microscope is that it provides resolution down to the atomic scale in real space. Here is an analogy which shows the precision of an SPM working with atomic resolution. Such instruments are about 10 cm in size and can image with a resolution of about 1 Å, corresponding to a precision of about 10^{-9} of its size. Scaling this precision of 10^{-9} up to macrosize dimensions would correspond to using a pencil 1,000 km in length to write letters from Cologne (Germany) in a notebook in Rome (Italy) with 1 mm resolution!

One important figure of merit in microscopy is the resolution. Figure 1.3 compares the resolution (right end of the boxes in Fig. 1.3) and the imaging ranges of different types of microscopy. The resolution of the human eye reaches down to one tenth of a millimeter. Ordinary optical microscopy (not any kind of super-resolution microscopy) reaches to slightly better than one micrometer due to the limitations set by the wavelength of visible light. Scanning electron microscopy (SEM) reaches to about one nanometer [7]. Transmission electron microscopy (TEM) [8] is capable of a resolution in the atomic range as are the various types of scanning probe microscopy.

While the resolution limit is important in microscopy also other characteristics are essential. For instance, the time to obtain an image, the contrast mechanisms (topography, chemical contrast ...), the surface sensitivity, the working environment (ambient, vacuum, liquid ...), and last but not least the price of the microscope. Each microscopy technique has its advantages and disadvantages for a particular application. For instance, if information on the 3D topography of a surface is required SPM with its excellent surface sensitivity is the method of choice. If, however,

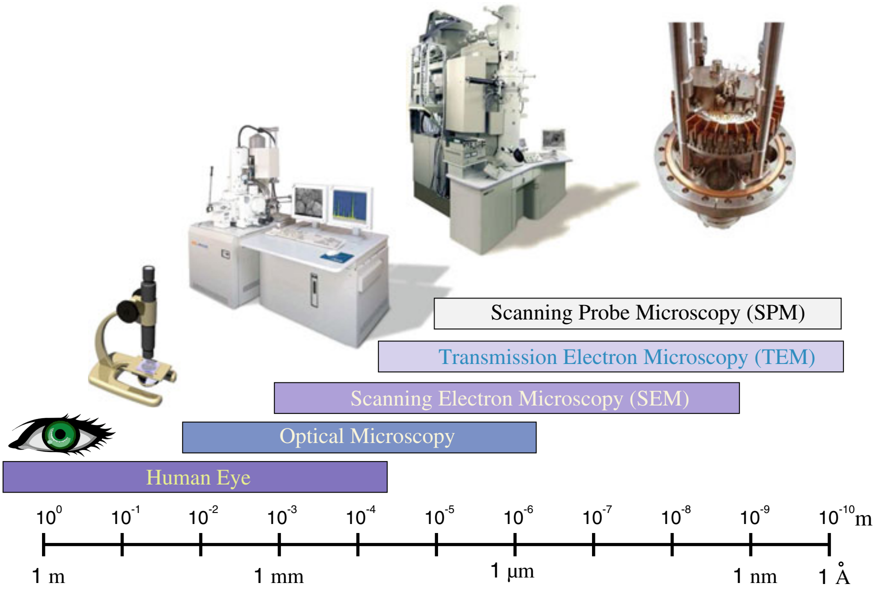

Scanning Probe Microscopy (SPM)
Transmission Electron Microscopy (TEM)
Scanning Electron Microscopy (SEM)
Optical Microscopy
Human Eye
10^0
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-7
10^-8
10^-9
10^-10 m
1 m
1 mm
1 µm
1 nm
1 Å

Fig. 1.3 Imaging ranges for different microscopy techniques in comparison

features below the surface, like e.g. dislocations, are to be imaged then TEM is more useful. If quick imaging within a few minutes down to the nanoscale is required then SEM should be used.

1.1 Scanning Tunneling Microscopy (STM)

The first kind of scanning probe microscope, the scanning tunneling microscope, (STM) was invented in 1981/1982 by Binnig and Rohrer [9–11] who received the Nobel prize in physics 1986 for this invention [12]. As STM is covered in detail in another book [13], we introduce this method only briefly in this introduction.

A schematic of an STM, with fine metal tip used as a probe, is shown in Fig. 1.4a. A voltage is applied between the tip and the (conducting) sample. The tip is approached toward the sample surface until a current flows. A current (the tunneling current) can be detected shortly before tip and sample come into direct contact. This happens at distances between tip and sample of the order of 0.5 nm. The tunneling current increases monotonously with decreasing tip-sample distance. Thus, a certain measured tunneling current corresponds to a specific tip-sample distance. Since the tunneling current varies strongly (exponentially) with the tip-sample distance this quantity can be used to control the tip-sample distance very precisely. We will see later that a 20% change in the tunneling current corresponds to a change in the tip-

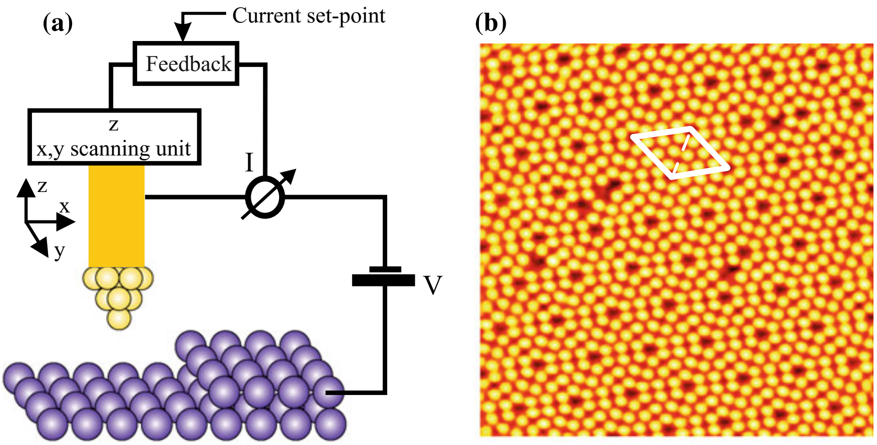


Fig. 1.4 **a** Schematic of a scanning tunneling microscope (STM). **b** STM image of the Si(111) surface. Individual atoms are observed as *yellow dots*. The rhombic unit cell is indicated by *white lines*. Besides the periodic arrangement of the atoms also defects such as single missing atoms can be observed

sample distance of only 0.1 Å. The tip is positioned with such high accuracy using piezoelectric actuator elements. The mechanical extension of this actuator elements is proportional to the voltage applied to their electrodes. In this way, the tip can be moved in x, y and z directions with sub-ångström precision.

While the tip is scanned along the surface in x and y directions, a feedback mechanism constantly adjusts the tip-sample distance by approaching or retracting the tip or the sample to a distance at which the tunneling current remains constant at a preset setpoint value. If there is an atomic step at the surface, as shown in Fig. 1.4a, and the tip approaches this step edge laterally during scanning, the tunneling current will rise beyond the setpoint value due to the smaller distance between tip and sample. As a reaction to this the feedback loop will generate a signal used to retract the tip in order to maintain the constant tunneling current at its setpoint value, corresponding to a specific tip-sample distance. The tip retraction is accomplished by applying the electric feedback signal to the piezoelement which changes the tip-sample distance. Recording the feedback signal (corresponding to the tip-sample distance) as a function of the lateral position results in a map (or image) of the tip height, which often corresponds to the surface topography of the sample surface. While the feedback mechanism was explained here for the STM, the principle is the same for all types of scanning probe microscopes. The tunneling current has just to be replaced by the actual quantity sensed, e.g. the tip-sample force in atomic force microscopy (AFM).

The interpretation of the tip height for constant tunneling current as the topography of the surface is a first approximation. So-called electronic effects can change this interpretation. A simplified example of this are atoms on a surface which have the same height (of their nuclei) but their electronic properties are different in the sense that one atom has a "higher electrical conductivity" than the other. The atom with the "higher conductivity" will appear higher (same tunneling current at a larger tip-sample distance) while for the case of the "less conducting atom" the tip has to approach closer to maintain the same tunneling current.

Figure 1.4b shows an atomically resolved image of a Si(111) surface [11]. Single silicon atoms are observed as yellow dots. The operation of an STM can be visualized experimentally by combining a scanning electron microscope (SEM) with an STM. The SEM can be used to image the motion of the STM tip during scanning. A movie of a scanning STM tip which is imaged while scanning with another microscope, namely an SEM, can be accessed at http://www.fz-juelich.de/pgi/pgi-3/microscope.

The tunneling effect is a quantum mechanical effect. The tunneling junction (sample-gap-tip) can be treated in different approximations. Here, we consider a simple one-dimensional approximation in order to grasp the very important exponential dependence of the tunneling current on the tip-sample distance.

In quantum mechanics, electrons in a solid are described by a wave function $\psi(\mathbf{r})$. In the free electron approximation the wave function of an electron of energy E and mass m is an oscillating complex function. The one-dimensional Schrödinger equation [14] in the presence of a constant potential V is solved by the (not normalized) wave function

$$\psi(z) \propto e^{\pm ikz}, \quad k = \sqrt{\frac{2m}{\hbar^2}(E - V)}. \tag{1.1}$$

Inside the solid ($z < 0$ in Fig. 1.5) the potential is constant and usually considered to vanish ($V = 0$) and the wave function is an oscillating function. When drawing this wave function, it should be remembered that this quantum mechanical wave function is a complex function, which is difficult to draw. Therefore, usually only the real or imaginary part is drawn, as in Fig. 1.5. The sinusoidal appearance of the real or imaginary part of the wave function should not make us forget that the absolute value $|\psi(z)|^2$ of such a wave function e^{ikz} has the constant value of one for all z.

In the following, we consider the electrons in a solid with the highest energy (at the Fermi level E_F) and call this energy the particle energy $E = E_F = E_{\text{particle}}$. The energy of these electrons at the Fermi level is lower than the energy of free electrons (the vacuum energy). This energy difference is roughly the bonding energy of the electrons inside the solid. If the Fermi energy were larger than the vacuum energy, the electrons would leak out of the solid toward the vacuum. The minimum energy needed to remove an electron from a solid is called the work function Φ, which is shown graphically in Fig. 1.5a.

Thus, at a surface there is a barrier (work function) preventing the electrons from leaving the solid to the vacuum level E_{vac}. In classical mechanics, particles cannot penetrate into a barrier which is higher than their energy. In quantum mechanics, particles can penetrate into a region with a barrier higher than their energy. In the vacuum region the term $E - V$ has the value $-\Phi$. Inserting this into (1.1) results (after pulling $\sqrt{-1} = i$ in front of the square root) in the following solution of the Schrödinger equation in the region of the potential barrier

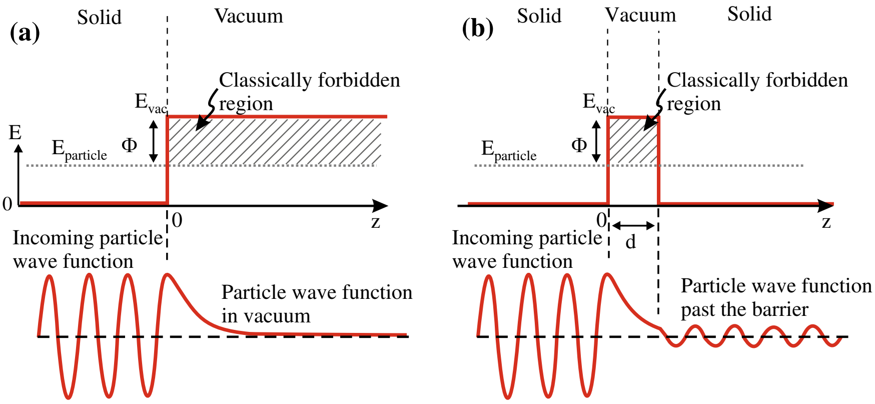

(a)
Solid
Vacuum
Classically forbidden region
E
E_{vac}
Φ
$E_{particle}$
0
0
z
Incoming particle wave function
Particle wave function in vacuum
(b)
Solid
Vacuum
Solid
Classically forbidden region
E_{vac}
Φ
$E_{particle}$
0
d
z
Incoming particle wave function
Particle wave function past the barrier

Fig. 1.5 **a** The *top graph* shows a potential barrier of height E_{vac} for $z > 0$ and the energy of an electron $E = E_{\text{particle}} = E_{\text{F}}$. The *lower graph* shows the real part of the electron wave function with an exponential decay of the wave function in the vacuum region. **b** The *top graph* shows the potential energy for a solid-vacuum-solid configuration. The *lower graph* shows the electron wave function (real part) oscillating in front of the barrier, exponentially decaying inside the barrier and again oscillating past the barrier

$$\psi(z) = \psi(0)e^{\pm ii\kappa z} = \psi(0)e^{\mp\kappa z}, \quad \kappa = \sqrt{\frac{2m}{\hbar^2}\Phi}. \tag{1.2}$$

This corresponds to an exponentially decaying real wave function inside the barrier (vacuum region), as shown in Fig. 1.5a. The exponentially rising solution in (1.2) is discarded, as it grows to infinity for $z > 0$.

If after some distance d the vacuum is replaced by another solid this configuration is already a one-dimensional model of the tunneling junction (electrode-gap-electrode). A potential diagram for such a tunneling barrier is shown in Fig. 1.5b. Since also inside the second solid the vacuum barrier is not present, the solution for the wave function inside the second solid is again an oscillating wave. This means that in quantum mechanics the electron has a finite probability in both metals. In the square barrier model a barrier, of height $\Phi = E_{\text{vac}} - E_{\text{F}}$ and width d is considered. In the course of the solution of the square barrier problem, the transmission coefficient for the wave function behind the barrier can be calculated. In the lowest order, neglecting all reflections of the wave function at the barrier, the probability of an electron being observed on the right side of the barrier is proportional to the absolute square of the wave function at the end of the barrier $|\psi(d)|^2$, which results according to (1.2) as

$$|\psi(d)|^2 = |\psi(0)|^2 e^{-2\kappa d}, \quad \kappa = \sqrt{\frac{2m}{\hbar^2}\Phi}. \tag{1.3}$$

A transmission coefficient T can be defined in the lowest order as

$$T \approx \frac{|\psi(d)|^2}{|\psi(0)|^2} = e^{-2\kappa d}. \tag{1.4}$$

The exact expression for the transmission coefficient can be found in [13, 14]. The main characteristics are: The transmission coefficient decays exponentially with the tip-sample distance d and decreases exponentially with the square root of the work function. If we use the right electrode as the tip, the tip probes the probability density of the electron states of the sample at distance d from the surface. It can be shown [13] that the tunneling current is proportional to the transmission coefficient.

Evaluating (1.4) using the free electron mass for m and a typical value for the work function of a metal ($\Phi \approx 4.5\,\text{eV}$), 2κ is about $20\,\text{nm}^{-1}$. Thus, a variation of the barrier thickness of 0.1 nm (relative to 0.5 nm) results in a difference in the transmission factor of an order of magnitude (~7.4). Hence the tunneling current increases by about an order of magnitude if the tip approaches by one Å to the sample. This sensitivity in the tip-sample distance is the reason for the extremely high vertical resolution of the STM which can reach the picometer regime. Atoms on the tip which protrude only 2.5 Å (about one atomic distance) less toward the sample carry only a factor of 150 less current. This means that the majority of the tunneling current is carried by the "last atom", which also explains the very high (ultimately atomic) lateral resolution of the STM. This is already all we will say here about STM. More details can be found in [13].

1.2 Introduction to Atomic Force Microscopy

One disadvantage of STM is that it can be used only for conducting samples since the tunneling *current* is the measured quantity. An atomic force microscope can also be used on insulating samples. The atomic force microscope (AFM) is alternatively known as the scanning force microscope (SFM). However, here we will use the more common name atomic force microscope. Instead of the tunneling current, which is the measured quantity in STM, in atomic force microscopy the force between the tip and sample is measured. In Fig. 1.6, a qualitative sketch of the force between tip and sample is given as function of the tip-sample distance. Three different regimes can be distinguished.

(a) If the tip is far away from the surface the force between tip and sample is negligible. (b) For closer distances an attractive (negative) force between tip and sample occurs. (c) For very small distances a strong repulsive force between tip and sample occurs. One problem with this behavior is that the tip-sample force which is used as measured signal depends non-monotonously on the tip-sample distance, i.e. for one value of the measured force in the attractive regime there are two tip-sample distances, point 1 and point 2 on the force-distance curve in Fig. 1.6. Care has to

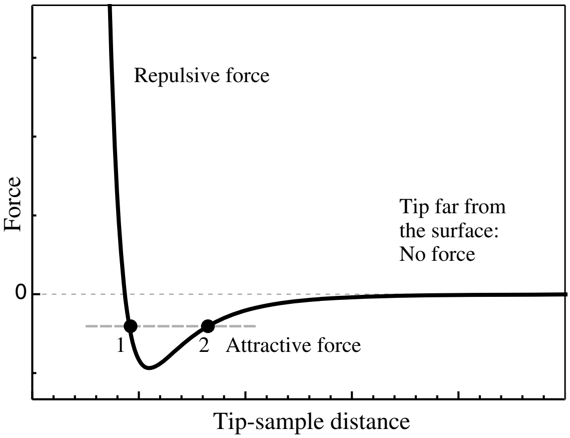


Fig. 1.6 Qualitative behavior of the force between tip and sample as function of tip-sample distance

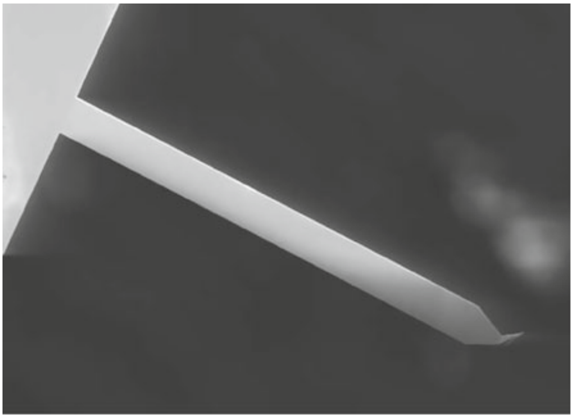


Fig. 1.7 SEM image of a silicon cantilever used in atomic force microscopy with a length of 450 μm

be taken to work only on one of the branches left or right of the minimum in the force-distance curve on which a monotonous force distance relation holds.

The force between tip and sample can be measured in a static mode using the deflection of a cuboid shaped flat spring (called cantilever) featuring a tip at its end. The cantilever acts as a spring and its deflection is proportional to the tip-sample force. If the stiffness of the cantilever spring k (spring constant) is known, the force between tip and sample can be determined by measuring the bending of the cantilever. Hooke's law gives $F_{\text{spring}} = -kz$, where F_{spring} is the spring force and z is the distance the cantilever spring is bent relative to its equilibrium position without the sample present. Figure 1.7 shows a typical silicon cantilever used as a force sensor in atomic force microscopy with a sharp tip (probe) at its end. The deflection of the lever is measured for instance using a laser beam reflected from the back of the cantilever into a split photodiode as shown in Fig. 1.8.

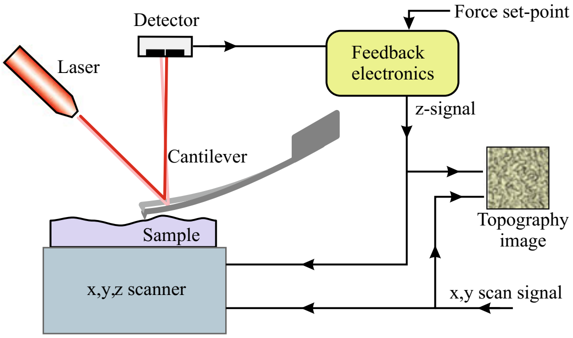


Fig. 1.8 Schematics of atomic force microscopy operation (static mode)

In the static mode of operation, the surface contour is mapped while scanning by changing the z-position of the tip or sample in such a way that the tip-sample force and, correspondingly, the tip-sample distance are kept constant. The z-signal maintaining a constant tip-sample distance is recorded as topography signal. In other words: the feedback loop maintains a constant force between the tip and the sample i.e. constant bending of the cantilever. The corresponding changes in the z-position required to maintain a constant tip-sample distance (i.e. constant force) correspond to the topography of the sample. If the measurements are performed in the repulsive regime of the force-distance curve the operating mode is called contact mode. In this case the last atoms of the tip are in direct contact with the surface atoms.

The atomic force microscope can also be operated in a mode known as the dynamic mode with an oscillating cantilever. This dynamic mode can also be operated in the attractive part of the tip-sample interaction. This mode of operation is called the non-contact mode. This is important when imaging soft samples (for instance polymers or biological samples), which would be destroyed by a strong repulsive tip-sample interaction. In the dynamic mode, the cantilever is excited to vibrate close to its free resonance frequency. When the atomic force microscope tip approaches the surface, the interaction between tip and sample changes the resonance frequency of the cantilever. The tip-sample force can be represented approximately by a second spring acting in addition to the cantilever spring. This additional spring leads to a change of the resonance frequency of the cantilever and correspondingly to a change of the cantilever amplitude. This change in amplitude can be used as a detection signal and can serve as the feedback signal for regulating the tip-sample distance. The distance regulation will be such that a constant amplitude and therefore a constant force (actually force gradient, as we will see later) is provided.

The idea of scanning probe methods can be considered more generally. A local probe is scanned over the surface which can detect physical or chemical properties with high spatial resolution. These techniques are often called SXM techniques where “X” stands for some specific interaction between tip and sample. Examples are for instance scanning capacitance microscopy (SCM) [15], Kelvin probe force microscopy (KPFM) [16], magnetic force microscopy (MFM) [17], and near-field scanning optical microscopy (NSOM/SNOM) [18].

1.3 A Short History of Scanning Probe Microscopy

It is a strange fact in the history of science that the scanning probe microscopy was invented so late. Nobody was brave enough to dare to think so simple: Use the blindman's stick principle all the way down to the atomic scale! The principle is so simple that there are several projects in which already pupils have built an STM. All the technical ingredients for an SPM were invented long before 1981. The piezoelectric effect was discovered at the end of the 19th century. The electronics for the STM is also simple; just a function generator to scan and a feedback controller. From 1930 on it would have been possible to build an STM as the scanning electron microscope was invented around this time. But no one dared to do so. This may be also an encouragement for your scientific carrier: be brave and visionary! Some important and nevertheless simple things may not have been discovered yet.

Here is a short history of scanning probe microscopy:

- 1972 Development of the Topografiner by Young, Ward, and Scire (precursor of the STM) [19].
- 1981 Construction of the first STM by Binnig et al. [9, 10].
- 1982 First image of the atomic structure of the Si(111)-(7 × 7) surface by Binnig et al. [11].
- 1985 Invention of the atomic force microscope (AFM) by Binnig et al. [20].
- 1986 Nobel prize in physics for the invention of the STM awarded to Binnig and Rohrer [12].
- 1987 Element-sensitive imaging of GaAs with the STM by Feenstra [21].
- 1989 AFM frequency modulation (FM) detection introduced by Albrecht, Grütter et al. [22].
- 1990 Optical beam deflection method introduced by Meyer and Amer [23].
- 1990 First positioning of single atoms on a surface with a low temperature STM by Eigler and Schweizer [3].
- 1993 Tapping mode in AFM introduced by Zhong et al. [24].
- 1995 First atomic resolution with an AFM by Giessibl [25].
- 1998 First vibrational spectroscopy with the STM by Stipe, Rezaei, and Ho [26].

More details on the early history of scanning probe microscopy can be found in [27]. Today scanning probe microscopes are standard tools in materials science, physics, chemistry, biology and engineering. Many thousands of these microscopes are in operation worldwide, and they are as common and as popular as the scanning electron microscopes.

1.4 Summary

- In scanning probe microscopy (SPM) a sharp probe tip is scanned over a surface and properties of the surface are sensed at the nano scale or atomic scale.

- Different kinds of microscopes are used for nanoscale imaging (scanning and transmission electron microscopes as well as scanning probe microscopes) and all have their advantages and disadvantages in terms of resolution, working environment, contrast mechanisms, time to obtain an image, and price.
- The atomic resolution in scanning tunneling microscopy (STM) results from the exponential dependence of the tunneling current on the tip-sample distance.
- Atomic force microscopy (AFM) can be also applied to insulating samples. The deflection of a small cantilever spring senses the force between tip and sample.
- In SPM, during scanning the height of the tip is adjusted by a feedback loop (and recorded as the topography signal) such that the measured signal (i.e. tunneling current or tip-sample force) and correspondingly the tip-sample distance is kept constant.
- The optical beam deflection method is used to measure the cantilever deflection in AFM. A laser beam is reflected from the back of the cantilever and a signal measuring the cantilever deflection is detected by a split photodiode.
- In the dynamic operation mode of AFM, the cantilever oscillates and the resonance frequency and subsequently the amplitude change due to the force between tip and sample.

References

1. N. Taniguchi, *On the basic concept of 'Nano-technology*, in *Proceedings of the International Conference on Production Engineering* (Tokyo, Part II, Japan Society of Precision Engineering (1974)
2. http://calteches.library.caltech.edu/1976/1/1960Bottom.pdf
3. D.M. Eigler, E.K. Schweizer, Positioning single atoms with a scanning tunnelling microscope. Nature **344**, 524 (1990). https://doi.org/10.1038/344524a0
4. M.F. Crommie, C.P. Lutz, D.M. Eigler, E.J. Heller, Waves on a metal surface and quantum corrals. Surf. Rev. Lett. **2**, 127 (1995). https://doi.org/10.1142/S0218625X95000121
5. L. Bartels, G. Meyer, K.-H. Rieder, Basic steps of lateral manipulation of single atoms and diatomic clusters with a scanning tunneling microscope tip. Phys. Rev. Lett. **79**, 697 (1997). https://doi.org/10.1103/PhysRevLett.79.697
6. A. Strozecka, J. Myslivecek, B. Voigtländer, Scanning tunneling spectroscopy and manipulation of C_{60} on Cu(111). Appl. Phys. A **87**, 475 (2007). https://doi.org/10.1007/s00339-007-3914-z
7. L. Reimer, *Scanning Electron Microscopy, Physics of Image Formation and Microanalysis*, 2nd edn. (Springer, Berlin, Heidelberg, 1998). http://dx.doi.org/10.1007/978-3-540-38967-5
8. D.B. Williams, C.B. Carter, *Transmission Electron Microscopy a Textbook for Materials Science*, 2nd edn. (Springer, Berlin, Heidelberg, 2009) http://dx.doi.org/10.1007/978-0-387-76501-3
9. G. Binnig, H. Rohrer, Ch. Gerber, E. Weibel, Tunneling through a controllable vacuum gap. Appl. Phys. Lett. **40**, 178 (1982). https://doi.org/10.1063/1.92999
10. G. Binnig, H. Rohrer, Ch. Gerber, E. Weibel, Surface studies by scanning tunneling microscopy. Phys. Rev. Lett. **49**, 57 (1982). https://doi.org/10.1103/PhysRevLett.49.57
11. G. Binnig, H. Rohrer, Ch. Gerber, E. Weibel, 7 × 7 Reconstruction on Si(111)Resolved in real space. Phys. Rev. Lett. **50**, 120 (1983). https://doi.org/10.1103/PhysRevLett.50.120
12. https://www.nobelprize.org/prizes/physics/1986/summary

13. B. Voigtländer, *Scanning Probe Microscopy Atomic Force Microscopy and Scanning Tunneling Microscopy*, 1st edn. (Springer, Berlin, Heidelberg, 2015). https://doi.org/10.1007/978-3-662-45240-0
14. L.I. Schiff, *Quantum Mechanics*, 3rd edn. (McGraw-Hill, New York, 1968)
15. J.R. Matey, J. Blanc, Scanning capacitance microscopy. J. Appl. Phys. **57**, 1437 (1985). https://doi.org/10.1063/1.334506
16. S. Sadewasser, Th. Glatzel (eds.), *Kelvin Probe Force Microscopy - Measuring and Compensating Electrostatic Forces*, 1st edn. (Springer, Berlin, Heidelberg, 2012). https://doi.org/10.1007/978-3-642-22566-6
17. H. Hopster, H.P. Oepen (eds.), *Magnetic Microscopy of Nanostructures*, 1st edn. (Springer, Berlin, Heidelberg, 2005). https://doi.org/10.1007/b137837
18. U. Dürig, D.W. Pohl, F. Rohner, Near-field optical scanning microscopy. J. Appl. Phys. **59**, 3318 (1986). https://doi.org/10.1063/1.336848
19. R. Young, J. Ward, F. Scire, The Topografiner: an instrument for measuring surface microtopography. Rev. Sci. Instrum. **43**, 999 (1972). https://doi.org/10.1063/1.1685846
20. G. Binnig, C.F. Quate, Ch. Gerber, Atomic force microscope. Phys. Rev. Lett. **49**, 57 (1982). https://doi.org/10.1103/PhysRevLett.56.930
21. R.M. Feenstra, J.A. Stroscio, J. Tersoff, A.P. Fein, Atom-selective imaging of the GaAs(110) surface. Phys. Rev. Lett. **58**, 1192 (1987). https://doi.org/10.1103/PhysRevLett.58.1192
22. T.R. Albrecht, P. Grütter, D. Horne, D. Rugar, Frequency modulation detection using high-Q cantilevers for enhanced force microscope sensitivity. J. Appl. Phys. **69**, 668 (1989). https://doi.org/10.1063/1.347347
23. G. Meyer, N.M. Amer, Novel optical approach to atomic force microscopy. Appl. Phys. Lett. **53**, 1045 (1989). https://doi.org/10.1063/1.100061
24. Q. Zhong, D. Inniss, K. Kjoller, V.B. Elings, Fractured polymer/silica fiber surface studied by tapping mode atomic force microscopy. Surf. Sci. Lett. **290**, L688 (1993). https://doi.org/10.1016/0039-6028(93)90582-5
25. F.J. Giessibl, Atomic resolution of the silicon (111)-(7 × 7) surface by atomic force microscopy. Science **267**, 68 (1995). https://doi.org/10.1126/science.267.5194.68
26. B.C. Stipe, M.A. Rezaei, W. Ho, Localization of inelastic tunneling and the determination of atomic-scale structure with chemical specificity. Phys. Rev. Lett. **82**, 1724 (1999). https://doi.org/10.1103/PhysRevLett.82.1724
27. C.C.M. Moody, *Instrumental Community: Probe Microscopy and the Path to Nanotechnology*, 1st edn. (MIT Press, Cambridge, 2011). ISBN 9780262134941

Chapter 2
Harmonic Oscillator

In atomic force microscopy, vibrations play a central role in several areas. If, for instance, an atomic force microscope rests on a table you might wonder what this has to do with vibrations. However, floor vibrations with amplitudes of roughly one tenth of a micrometer (100 nm) have to be compared to an amplitude stability of less than 0.01 nm which is necessary for atomically resolved imaging in AFM. Thus, the vibrational noise amplitude is about 10,000 times larger than the signal to be measured. This means that knowledge about vibrations and vibration isolation is essential for scanning probe methods. Another area where oscillations are an important topic is dynamic atomic force microscopy. In the dynamic mode of atomic force microscopy, a cantilever vibrating close to (or at) its resonance frequency is used as a sensor. The simplest way to study vibrations is to study the harmonic oscillator. In this chapter we will study the mechanical harmonic oscillator.

2.1 Free Harmonic Oscillator

The simplest example of a harmonic oscillator is a mass on a spring (Fig. 2.1). The position to which gravity extends the spring in equilibrium is chosen as the point of zero extension. The displacement relative to this point is called z. The force exerted by the spring on the mass m during the oscillation is given by Hooke's law as

$$F = -kz, \tag{2.1}$$

with k being the spring constant. If the spring deflection has negative values ($z < 0$, longer spring extension), the direction of the force is positive and vice versa. Thus, the minus sign in (2.1) appears because the force exerted by the spring has a direction opposite to the deflection z. Newton's second law tells us that the equation of motion for the mass m is

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3_2

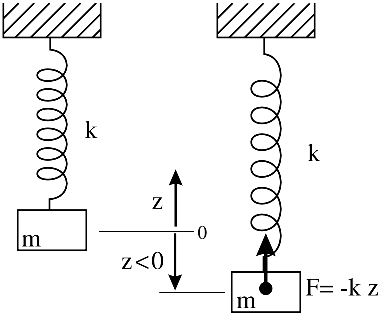


Fig. 2.1 The simplest example of a harmonic oscillator: a mass on a spring

$$ma = m\frac{\mathrm{d}^2z}{\mathrm{d}t^2} = m\ddot{z} = F = -kz. \tag{2.2}$$

An ansatz for the solution of the equation of motion (2.2) is $z = \cos(\omega_0 t)$ with ω_0 being a parameter which has to be determined.[1] We verify that this is a correct solution by differentiating z two times:

$$\frac{\mathrm{d}z}{\mathrm{d}t} = -\omega_0 \sin(\omega_0 t); \quad \frac{\mathrm{d}^2z}{\mathrm{d}t^2} = -\omega_0^2 \cos(\omega_0 t) = -\omega_0^2 z. \tag{2.3}$$

Formally (2.2) is solved if

$$\omega_0 = \sqrt{\frac{k}{m}}. \tag{2.4}$$

But what is the physical significance of ω_0? We know that the cosine function repeats itself if the argument is larger than 2π. Therefore, the mass makes (compared to $t = 0$) one complete cycle of oscillation if $\omega_0 t = 2\pi$. This time, we call the period of the oscillation T, and ω_0 is given by

$$\omega_0 = 2\pi/T. \tag{2.5}$$

The angular frequency ω_0 is the number of radians which the oscillation proceeds per time, while the frequency $f_0 = 1/T$ is the number of oscillations per time ($\omega_0 = 2\pi f_0$). Equation (2.4) tells us that if the mass is larger it takes a longer time for one oscillation and if the spring constant is stronger the mass will move more quickly. This frequency ω_0 at which the harmonic oscillator oscillates is also called the natural frequency of the oscillator, or also the resonance frequency of the oscillator, for reasons we will discuss later. Note that the period of oscillation (and also ω_0) does not depend on how far we stretch the spring at the beginning. Any solution multiplied by a constant factor is still a solution of (2.2).

[1] The argument of the cosine is named the phase. The phase increases linearly with time if ω_0 is constant.

We have found a solution to the equation of motion. But is this the only one or are there more solutions? Also the sine function provides a valid solution. The most general solution is a linear combination of a sine and a cosine function

$$z = A\cos(\omega_0 t) + B\sin(\omega_0 t). \tag{2.6}$$

There is a more intuitive way to find the general solution. When we used the cosine function as solution, the oscillation started with the maximum extension at time zero. However, alternatively also any other time during the oscillation could be chosen as the start of the oscillation. This shift of the time corresponds to a shift of the phase of the oscillation by a constant phase shift ϕ. Thus, all solutions are captured if the solution is shifted by a constant (but arbitrary) phase shift[2] ϕ, and the general solution results as

$$z = a\cos(\omega_0 t + \phi). \tag{2.7}$$

The two solutions given in (2.6) and (2.7) are in fact equivalent. Using the mathematical identity

$$\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta, \tag{2.8}$$

the following relations between A, B in (2.6) and a, ϕ in (2.7) are obtained

$$A = a\cos\phi, \quad B = -a\sin\phi. \tag{2.9}$$

Moreover, it can be shown that the solutions given in (2.6) and (2.7) are the general solution to the equation of motion. There are no other solutions.

In the general solution of the equation of motion, we introduced two more constants: A and B, or a and ϕ, respectively. How are these constants determined? They are determined by the initial conditions of the motion. For instance if we start the motion from a static extension $z_0 = a = A$, the values B and ϕ are zero. Now we determine these constants for the most general initial condition: z_0, v_0. The acceleration $a(t)$ cannot be specified as an initial condition. It is given by the spring constant, mass and $z(t)$ according to (2.2). We use the form for the general solution given in (2.6) and its derivative

$$v(t) = -\omega_0 A\sin(\omega_0 t) + \omega_0 B\cos(\omega_0 t). \tag{2.10}$$

These equations are valid for all times, but we know z and v at time $t = 0$. If we insert $t = 0$ we obtain

$$z_0 = A + B\cdot 0 = A \quad v_0 = -\omega_0 A\cdot 0 + \omega_0 B = \omega_0 B. \tag{2.11}$$

[2] Sometimes ϕ is called phase, as well as the whole argument of the cos function in (2.7). What is meant by the term phase should be clear from the context.

We therefore find that the constants A and B can be determined by the initial conditions as

$$A = z_0 \quad \text{and} \quad B = v_0/\omega_0. \tag{2.12}$$

2.2 Free Harmonic Oscillator with Damping

In in the equation of motion for the harmonic oscillator (2.2) no damping was included. Usually the viscous damping in a fluid (liquid or gas) is considered as proportional to the velocity $\dot{z}$, resulting in a frictional force in the direction opposite to the velocity as $F_{\mathrm{frict}} = -\beta\dot{z}$. Due to some conventions a different constant, the quality factor Q is introduced which is defined by the following equation $F_{\mathrm{frict}} = -m\omega_0/Q\dot{z}$. The physical significance of the quality factor Q will be discussed in detail later in this chapter, however, we see already here that for a decreasing damping force F_{frict} the quality factor Q increases. Adding the frictional force to the equation of motion (2.2) results in

$$m\ddot{z} = -kz - \frac{m\omega_0}{Q}\dot{z}, \tag{2.13}$$

or

$$\ddot{z} + \frac{\omega_0}{Q}\dot{z} + \omega_0^2 z = 0. \tag{2.14}$$

We choose the ansatz $z = A' \exp(\lambda t)$, because the exponential function is so easy to differentiate. Inserting this ansatz into (2.14) results in

$$\left(\lambda^2 + \frac{\omega_0}{Q}\lambda + \omega_0^2\right) z = 0. \tag{2.15}$$

Thus, the expression in the brackets has to vanish (quadratic equation), resulting in the following expression for lambda

$$\lambda_{1,2} = -\frac{\omega_0}{2Q} \pm \omega_0\sqrt{\frac{1}{4Q^2} - 1}. \tag{2.16}$$

In the cases we will consider here, the damping is small and only cases in which $Q > 1$ occur. In this case the expression under the square root becomes negative and we rewrite $\lambda_{1,2}$ as

$$\lambda_{1,2} = -\frac{\omega_0}{2Q} \pm i\omega_0\sqrt{1 - \frac{1}{4Q^2}} = -\frac{\omega_0}{2Q} \pm i\omega_{\mathrm{hom}}, \tag{2.17}$$

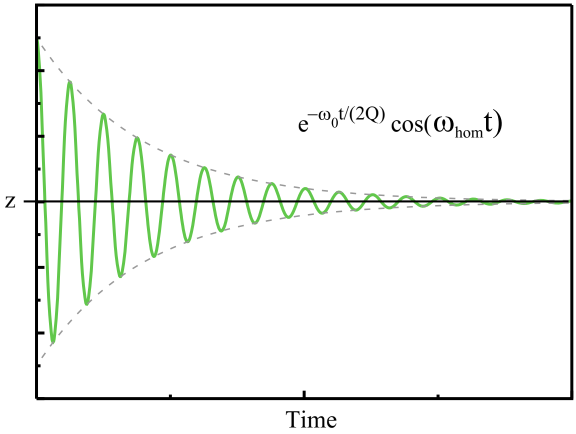


Fig. 2.2 Oscillation of a free harmonic oscillator with damping

with $\omega_{\text{hom}} = \omega_0\sqrt{1 - 1/(4Q^2)}$, since (2.14) with the zero on the right hand side is a called a *homogenous* differential equation. With two expressions for λ two solutions for $z = A' \exp(\lambda t)$ result. The general solution can be written as

$$z = e^{-\frac{\omega_0 t}{2Q}}(Be^{i\omega_{\text{hom}}t} + Ce^{-i\omega_{\text{hom}}t}). \tag{2.18}$$

It looks as if this solution is a complex solution, while the quantity z has to be real. However, a real solution results, if the two constants B and C are chosen as real and $B = C = A/2$. In this case the Euler equation leads to

$$z = Ae^{-\frac{\omega_0 t}{2Q}} \cos(\omega_{\text{hom}}t). \tag{2.19}$$

This solution is valid if the initial conditions are chosen as $z(t = 0) = A$ and $\dot{z}(t = 0) = 0$. It can be shown that for general initial conditions a phase shift ϕ has to be added in the cosine term [1]. This solution corresponds to an oscillation with a frequency ω_{hom} which is slightly lower than the frequency ω_0 of the undamped harmonic oscillator. This oscillation is damped by the exponential damping term $e^{-\frac{\omega_0 t}{2Q}}$, corresponding to the envelope of the damped oscillation, as shown in (Fig. 2.2).

2.3 Driven Harmonic Oscillator

In dynamic atomic force microscopy, we will consider a cantilever which is excited, driven or moved with a sinusoidal external excitation amplitude. The simplest model for this is a harmonic oscillator in which the upper fixing point of the spring is oscillated (excited) sinusoidally with $z_{\text{drive}}(t) = A_{\text{drive}} \cos(\omega_{\text{drive}}t)$ (Fig. 2.3). The resulting

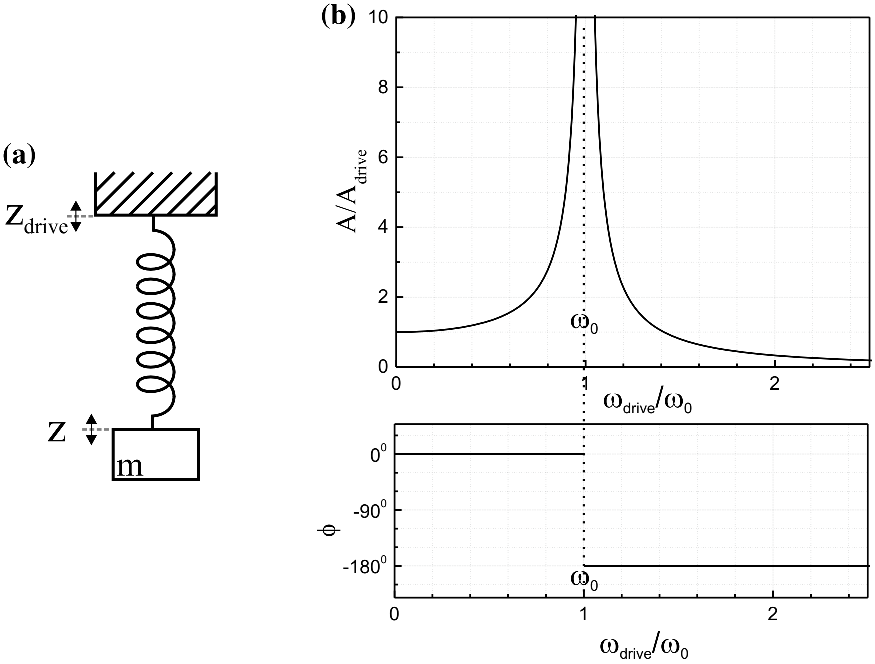


Fig. 2.3 **a** Sketch of a driven harmonic oscillator, driven with an oscillatory driving amplitude $z_{\text{drive}}(t) = A_{\text{drive}} \cos(\omega_{\text{drive}} t)$ **b** Amplitude and phase shift of an undamped driven harmonic oscillator as a function of ω_{drive} showing a resonance at ω_0

force of the spring on the mass m is then $F = -k(z - z_{\text{drive}})$. The equation of motion results as

$$ma = m\ddot{z} = -k(z - z_{\text{drive}}). \tag{2.20}$$

The driving frequency ω_{drive} can be different from the resonance frequency of the oscillator ω_0. The question arises at which frequency the driven harmonic oscillator will oscillate. At its resonance frequency ω_0, at the driving frequency ω_{drive}, or at some value in between? It turns out that the driven harmonic oscillator will oscillate in the steady-state at the driving frequency ω_{drive}. One special solution for the equation of motion is

$$z(t) = A\cos(\omega_{\text{drive}} t). \tag{2.21}$$

Inserting this ansatz into the equation of motion (2.20) results in

$$-m\omega_{\text{drive}}^2 A\cos(\omega_{\text{drive}} t) = -kA\cos(\omega_{\text{drive}} t) + kA_{\text{drive}}\cos(\omega_{\text{drive}} t). \tag{2.22}$$

We find that $z = A\cos(\omega_{\text{drive}} t)$ is a solution of the equation of motion if

$$A = \frac{\omega_0^2 A_{\text{drive}}}{\omega_0^2 - \omega_{\text{drive}}^2}. \tag{2.23}$$

The special solution (2.21) means that m oscillates at the driving frequency with an amplitude which depends on the driving frequency and also on the resonance frequency of the oscillator. If $\omega_{\text{drive}} < \omega_0$ then displacement and driving excitation are in the same direction. If $\omega_{\text{drive}} > \omega_0$ then A becomes negative. This is equivalent to a positive amplitude and a phase shift of $-180°$ of the oscillation $z(t)$ relative to the driving excitation. The amplitude and phase shift for an undamped driven harmonic oscillator are shown in (Fig. 2.3). If $\omega_{\text{drive}} \ll \omega_0$ the amplitude A approaches the excitation amplitude A_{drive}. If $\omega_{\text{drive}} \gg \omega_0$ the amplitude approaches zero because the mass can no longer follow the high frequency of the driving excitation.

As can be seen in Fig. 2.3 the amplitude A approaches infinity if ω_{drive} approaches ω_0. We will see in the next section that damping of the harmonic oscillator prevents this "resonance catastrophe".

2.4 Driven Harmonic Oscillator with Damping

Including damping to the driven harmonic oscillator is a more realistic case which we consider in the following. An additional friction term has to be included to the equation of motion (2.20). Usually the viscous damping in a fluid (liquid or gas) is considered as proportional to the velocity, resulting in a frictional force in the direction opposite to the velocity. In Fig. 2.4 the damping force is represented graphically by a viscous dashpot. This dashpot can be anchored in different ways: In Fig. 2.4a this dashpot is anchored to an external (not moving) reference frame, resulting in a frictional force proportional to $\dot{z}$, as $F_{\text{frict}} = -m\omega_0/Q\dot{z}$. This situation corresponds for instance to the situation in which an AFM cantilever oscillation is damped by the surrounding air. In Fig. 2.4b a situation is shown in which the viscous dashpot is anchored to the oscillating driving reference frame, resulting in a frictional force proportional to $\dot{z} - \dot{z}_{\text{drive}}$, as $F_{\text{frict}} = -m\omega_0/Q(\dot{z} - \dot{z}_{\text{drive}})$. This situation corresponds to the situation of vibration isolation, in which a scanning probe microscope has to be isolated from external vibrations, and is considered further in Sect. 3.6.1. Here we consider in the following the dashpot anchoring shown in Fig. 2.4a.

As driving excitation we consider an external exciting amplitude $z_{\text{drive}}(t) = A_{\text{drive}} \cos(\omega t)$. Here and in the following we replaced $\omega_{\text{drive}} \equiv \omega$, in order to have a simpler notation. The spring force acting on the oscillating mass is again proportional to the difference between the position of the mass z and the excitation amplitude z_{drive} as $F = -k(z - z_{\text{drive}})$. With this the equation of motion reads

$$m\ddot{z} = -m\frac{\omega_0}{Q}\dot{z} - k(z - z_{\text{drive}}). \tag{2.24}$$

After dividing by m and replacing $\omega_0^2 = k/m$ results in[3]

[3] We do not perform the replacement in all terms, because of a later use of the equations.

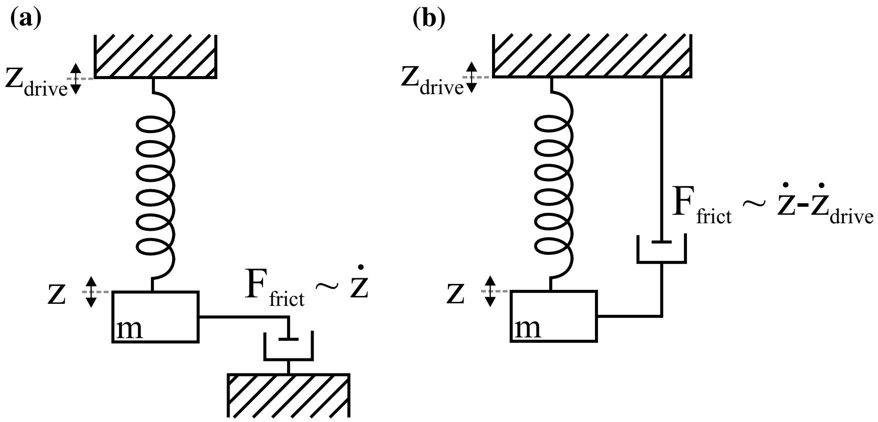


Fig. 2.4 Schematic of a driven damped harmonic oscillator for two different ways of anchoring the viscous dashpot representing the damping. **a** Dashpot anchored to an external fixed reference frame and **b** dashpot anchored to the driving reference frame. These two variants result in different expressions for the frictional force

$$\ddot{z} + \frac{\omega_0}{Q}\dot{z} + \frac{k}{m}z = \omega_0^2 z_{\text{drive}}. \tag{2.25}$$

Solving this equation would be quite difficult without the use of complex numbers. The trick here is to consider z and z_{drive} as complex numbers ($\tilde{z}$ and $\tilde{z}_{\text{drive}}$) and find the complex solution for the differential equation. Since the physical quantities are real and the differential equation is linear, at the end only the real part of $\tilde{z}$ is our solution. The deflections z and z_{drive} are regarded as complex numbers as

$$\tilde{z} = Ae^{i(\omega t+\phi)} = Ae^{i\phi}e^{i\omega t} = \hat{z}e^{i\omega t} \quad \text{and} \quad \tilde{z}_{\text{drive}} = A_{\text{drive}}e^{i\omega t}. \tag{2.26}$$

Without loss of generality we set the phase shift of the excitation amplitude z_{drive} to zero, i.e. A_{drive} is real, while $\hat{z}$ is regarded as a complex number with a (real) phase shift ϕ and (real) oscillation amplitude A as, $\hat{z} = Ae^{i\phi}$. The real part of $\tilde{z}$ will later be the real solution for the deflection z of the mass m. The nice thing about the complex notation is that differentiation of $\tilde{z}$ is now just multiplication with $i\omega$ ($\frac{\mathrm{d}\tilde{z}}{\mathrm{d}t} = \hat{z}i\omega e^{i\omega t} = i\omega\tilde{z}$). This means differentiation in (2.25) (with $z \rightarrow \tilde{z}$) can be easily executed and this differential equation converts to the simple algebraic equation

$$\left[(i\omega)^2\hat{z} + \frac{\omega_0}{Q}i\omega\hat{z} + \frac{k}{m}\hat{z}\right]e^{i\omega t} = \omega_0^2 A_{\text{drive}}e^{i\omega t}. \tag{2.27}$$

After dividing both sides by $e^{i\omega t}$, we obtain the complex solution

$$\hat{z} = \frac{\omega_0^2 A_{\text{drive}}}{\frac{k}{m} - \omega^2 + i\frac{\omega_0}{Q}\omega}. \tag{2.28}$$

Now the real z is the real part of the complex quantity $\tilde{z}$ as

$$z = \text{Re}(\tilde{z}) = \text{Re}(\hat{z}e^{i\omega t}) = \text{Re}(Ae^{i(\omega t+\phi)}). \tag{2.29}$$

Since A and ϕ are real, the resulting real position z reads

$$z = A\cos(\omega t + \phi), \tag{2.30}$$

with the amplitude A and phase shift ϕ between excitation amplitude and oscillation amplitude.

In order to calculate A we recall that $\hat{z} = Ae^{i\phi}$. Therefore, $\hat{z}\hat{z}^* = A^2$ and A^2 can be written as

$$A^2 = \frac{\omega_0^4 A_{\text{drive}}^2}{\left(\frac{k}{m} - \omega^2 + \frac{i\omega_0\omega}{Q}\right)\left(\frac{k}{m} - \omega^2 - \frac{i\omega_0\omega}{Q}\right)} = \frac{\omega_0^4 A_{\text{drive}}^2}{\left(\frac{k}{m} - \omega^2\right)^2 + \frac{\omega_0^2\omega^2}{Q^2}}. \tag{2.31}$$

Furthermore, the oscillation amplitude A can be written as a function of the normalized frequency ω/ω_0 and finally replacing k/m by ω_0^2 results in

$$A^2 = \frac{A_{\text{drive}}^2}{\left[1 - \left(\frac{\omega}{\omega_0}\right)^2\right]^2 + \frac{1}{Q^2}\left(\frac{\omega}{\omega_0}\right)^2}. \tag{2.32}$$

The phase shift ϕ of the oscillation relative to the excitation can be obtained as follows. In general the phase φ of a complex number $x = re^{i\varphi}$ can be obtained from the relation $\tan\varphi = \frac{\text{Im}(x)}{\text{Re}(x)}$. In order to calculate the phase shift ϕ, we recall that $\hat{z} = Ae^{i\phi}$. However, according to (2.28) the real and imaginary parts of $1/\hat{z}$ are much easier to find. Therefore, we write

$$\frac{1}{\hat{z}} = \frac{1}{Ae^{i\phi}} = \frac{1}{A}e^{-i\phi} = \frac{1}{\omega_0^2 A_{\text{drive}}}\left(\frac{k}{m} - \omega^2 + i\frac{\omega_0}{Q}\omega\right). \tag{2.33}$$

Using the fact that $\tan(-\phi) = -\tan\phi$, we see that

$$\tan\phi = \frac{-\omega_0\omega}{Q\left(\frac{k}{m} - \omega^2\right)}. \tag{2.34}$$

Also the phase shift ϕ can be written as function of the normalized frequency ω/ω_0, and replacing k/m by ω_0^2, as

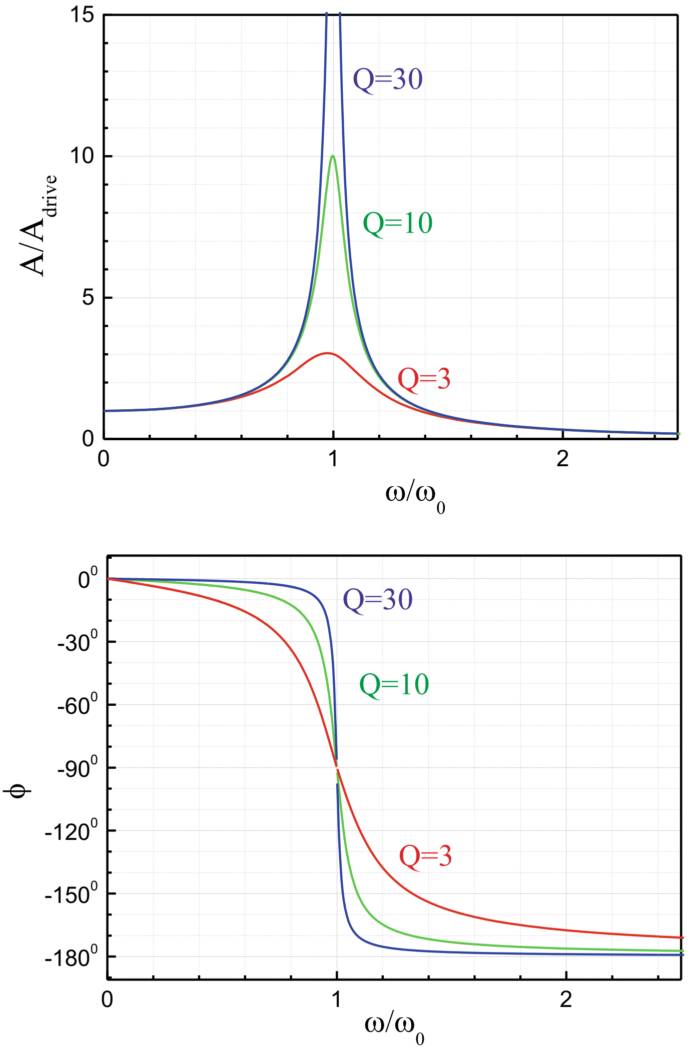


Fig. 2.5 Amplitude and phase shift of a damped driven harmonic oscillator as a function of $\omega \equiv \omega_{\text{drive}}$, for different values of damping, expressed by the quality factor Q

$$\tan\phi = \frac{-\frac{\omega}{\omega_0}}{Q\left[1-\left(\frac{\omega}{\omega_0}\right)^2\right]}. \tag{2.35}$$

With these results, the amplitude (2.32) and phase shift (2.35) in the solution (2.30) are calculated as a function of given variables. The resonance curve in Fig. 2.5 shows the amplitude and the phase shift of a driven damped harmonic oscillator for three different values of Q. For small driving frequencies $\omega \ll \omega_0$, the motion of the oscillator mass just follows the outer excitation with a phase shift approaching zero; i.e. the oscillation is in phase with the excitation. For larger driving frequencies the phase of the oscillation lags behind that of the excitation. With our convention regarding the sign of the phase shift (i.e. the positive sign $+\phi$ in (2.30)), the phase becomes negative for larger frequencies (phase lag), as shown in Fig. 2.5. Often the opposite convention for the phase shift is chosen (i.e. a negative sign $-\phi$ in (2.30)), which leads to positive phase shifts (phase lead). For frequencies $\omega \gg \omega_0$, the amplitude A approaches zero and the phase (shift) approaches $-180°$, i.e. the motion of the oscillator mass is always in opposite to the excitation.

If we take the limit $\omega \gg \omega_0$ in (2.32) we find that the amplitude is proportional to $1/\omega^2$ for small damping, i.e. $Q \gg 1$. As seen in Fig. 2.5, the smaller the damping, the higher the maximum amplitude is. For small damping the maximum of the resonance curve is very close to the resonance frequency of the free harmonic oscillator ω_0. At any driving frequency the phase shift is smaller than zero, which means that the oscillator displacement z always lags behind the driving excitation (Fig. 2.5). The phase shift at resonance ($\omega = \omega_0$) is $-90°$, while it approaches $-180°$ for large driving frequencies.

The amplitude at the resonance frequency $A(\omega_0)$ can be obtained using (2.32) as

$$A(\omega_0) = QA_{\text{drive}}, \tag{2.36}$$

i.e. the amplitude at resonance is Q times higher than the excitation amplitude. For the case of cantilevers in atomic force microscopy this resonance enhancement of the excitation amplitude can be quite high. Due to damping in air, Q-factors of 500 are usual for cantilevers in air. In vacuum, quality factors higher than 10,000 can be reached.

For the case that the oscillation frequency is very close to ω_0, i.e. $\omega \approx \omega_0$, the expression for the resonance curve (2.32) can be approximated as

$$A^2 = \frac{A^2_{\text{drive}}}{\left[\left(1+\frac{\omega}{\omega_0}\right)\left(1-\frac{\omega}{\omega_0}\right)\right]^2 + \frac{1}{Q^2}\frac{\omega^2}{\omega_0^2}} \approx \frac{A^2_{\text{drive}}}{4\left(1-\frac{\omega}{\omega_0}\right)^2 + \frac{1}{Q^2}}. \tag{2.37}$$

In order to obtain this we used the approximations $1 + \frac{\omega}{\omega_0} \approx 2$ and $\frac{\omega^2}{\omega_0^2} \approx 1$, which hold if $\omega \approx \omega_0$.

An important quantity is the width of the resonance curve. Therefore, we calculate in the following the frequency $\omega_{1/2}$ at which the amplitude of the oscillation decreases to $1/\sqrt{2}$ of its value[4] at ω_0. This condition for the amplitudes can be written as

$$A(\omega_{1/2}) = \frac{1}{\sqrt{2}}A(\omega_0) = \frac{1}{\sqrt{2}}QA_{\text{drive}}. \tag{2.38}$$

[4] We use the decrease of the amplitude to $1/\sqrt{2}$ instead of $1/2$, because in this case the energy in the harmonic oscillator, which is proportional to the square of the amplitude, decreases to one half of its maximum value.

If we insert $\omega = \omega_{1/2}$ in expression (2.37), the following relation results

$$A^2(\omega_{1/2}) \approx \frac{A^2_{\text{drive}}}{4\left(1 - \frac{\omega_{1/2}}{\omega_0}\right)^2 + \frac{1}{Q^2}} \approx \frac{1}{2} Q^2 A^2_{\text{drive}}. \tag{2.39}$$

Solving this expression for $\omega_{1/2} - \omega_0$ results in $\omega_{1/2} - \omega_0 \approx \frac{1}{2}\frac{\omega_0}{Q}$. Since the full width of the resonance curve is twice of this, we obtain

$$\delta\omega \approx \frac{\omega_0}{Q}. \tag{2.40}$$

This means the larger the Q-factor, the narrower the resonance is.

The maximum of the resonance amplitude, which we determine in the following, lies at a slightly lower frequency than ω_0. The maximum of the resonance curve occurs at the frequency at which the denominator in (2.32) becomes minimal. Differentiating the denominator of (2.32) with respect to ω/ω_0, and equating this derivative to zero results in the following expression for the frequency ω_{max} at which the resonance curve has its maximum

$$\omega_{\text{max}} = \omega_0 \left(\sqrt{1 - \frac{1}{2Q^2}} \right). \tag{2.41}$$

The corresponding shift of the resonance curve to lower frequencies results as

$$\omega_{\text{max}} - \omega_0 = \omega_0 \left(\sqrt{1 - \frac{1}{2Q^2}} - 1 \right). \tag{2.42}$$

For the case of an AFM cantilever considered as a harmonic oscillator we estimate some values for this frequency shift of the resonance curve due to the damping Q of the cantilever. For a resonance frequency of $f_0 = 300$ kHz and quality factors of $Q =$ 10,000 and $Q = 300$, a frequency shift of 0.8 mHz and 0.8 Hz results, respectively. These are very small values and correspondingly in most cases we will neglect this small shift and consider the maximum of the amplitude to be located at ω_0, unless the quality factor is very low.

2.5 Transients of Oscillations

The solution for the damped driven harmonic oscillator (2.30) is the so-called steady-state solution after transients due to the initial conditions have died out. An example for a transient is an oscillation which starts from rest. The amplitude is initially zero, builds up after the excitation starts, and reaches the steady-state amplitude in the limit of large times. The steady-state solution (2.30) does not contain such transients arising from specific initial conditions.

It can be shown that the general solution of the driven damped harmonic oscillator is the specific solution (2.30) of the inhomogeneous system (i.e. including the external driving) plus a solution of the corresponding homogeneous problem. The corresponding homogeneous problem is the damped harmonic oscillator without external driving, which was considered in Sect. 2.2 and resulted for small damping in an exponentially decaying oscillation $z_{\text{hom}} = A' \exp(-\omega_0 t/(2Q)) \cos(\omega_{\text{hom}} t + \phi')$ with the oscillation frequency ω_{hom} being slightly lower than the resonance frequency ω_0 of the free harmonic oscillator $\omega_{\text{hom}} = \omega_0 \sqrt{1 - 1/(4Q^2)}$ and with A' and ϕ' as coefficients determined by the initial conditions.

If we call the specific solution z in (2.30) z_s, the general solution for the driven, damped harmonic oscillator is given as $z_{\text{general}} = z_s + z_{\text{hom}}$. It is necessary to include the solution of the damped harmonic oscillator without external driving z_{hom} since it can describe the transients which are not described by z_s. All aspects of z_s are specified in terms of the driving frequency, the driving amplitude, and the phase shift. Yet we still need some way to impose the constraints given by the initial conditions $z(0)$ and $v(0)$ in the general solution. The two coefficients A' and ϕ' give the freedom to match the general solution to $z(0)$ and $v(0)$.

As an example we consider as initial condition that the oscillation starts from rest. In Fig. 2.6 the general solution for the initial condition: starting from rest, is shown to be composed of the specific solution of the inhomogeneous system (Fig. 2.6a) plus the solution for the homogeneous system (transient) z_{hom} (Fig. 2.6b). In Fig. 2.6c the sum of both is shown for the case that $\omega = \omega_{\text{hom}}$. The specific solution in Fig. 2.6a is approached within the decay time for the homogeneous solution Fig. 2.6b. The fact that the situation is not always that simple is shown in Fig. 2.6d. Here the driving frequency deviates from ω_{hom}, which leads to a beating behavior before a steady-state solution is reached.

If the driven damped oscillator is oscillating in steady-state (Fig. 2.6a) and the driving amplitude is stopped suddenly, the problem is converted to a homogeneous one and the oscillator will de-excite as shown in Fig. 2.6b. This is a sinusoidal oscillation with the envelope decreasing as $\exp(-\omega_0 t/(2Q))$. This means that after a time $\tau = 2Q/\omega_0 = TQ/\pi$ the amplitude has decreased by $1/e$. This characteristic time is called ring-down time and increases with smaller damping. The same time is needed to build up the steady-state oscillation amplitude after a start from rest. This means that the oscillation builds up (decays) within roughly Q oscillation cycles and Q can be expressed as

$$Q = \frac{1}{2}\tau\omega_0 = \pi\tau f_0. \tag{2.43}$$

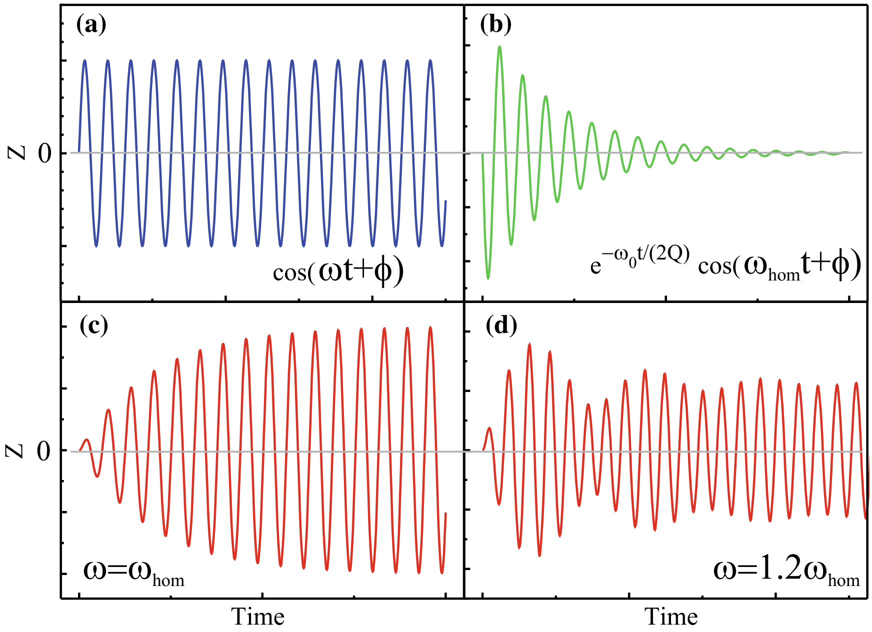


Fig. 2.6 The general solution for a damped driven harmonic oscillator is composed of the specific solution of the inhomogeneous driven system (steady-state solution), shown in (**a**) plus the solution of the homogeneous system without driving (transient), shown in (**b**). The initial conditions are chosen such that the general solution satisfies the given initial conditions (start from rest in this example). **c** and **d** show two examples of general solutions (for two different driving frequencies) starting from rest and approaching the steady-state solution for long times

2.6 Dissipation and Quality Factor of a Damped Driven Harmonic Oscillator

In the previous sections we have discussed that the Q-factor determines the height ($A(\omega_0) = QA_{\text{drive}}$) and the width ($\delta\omega = \omega_0/Q$) of the resonance curve of a driven damped oscillator. Now we will show that the quality factor of a driven damped harmonic oscillator can be also expressed as

$$Q = 2\pi \times \frac{\text{Energy stored in the oscillator}}{\text{Energy dissipated per cycle}}. \tag{2.44}$$

Let us first consider the energy dissipated per cycle. When the oscillator is initially at rest and an external oscillatory excitation is applied, energy is successively stored in the oscillator with the buildup of the oscillation (transient). If the oscillator is finally in a steady-state, the energy stored in the oscillator is constant and all the energy supplied by the external force ends (on average) up in the dissipative term. The instantaneous power dissipated is $F_{\text{frict}} \cdot v = m\omega_0/Qv^2$ and varies over one period, as v varies. The mean power consumed by the oscillator in steady-state can be written as[5]

$$\langle P\rangle = \langle F_{\text{frict}} \cdot v\rangle = \frac{m\omega_0}{Q}\langle v^2\rangle. \tag{2.45}$$

The brackets indicate an averaging over one oscillation period. Since $z = A\cos(\omega t + \phi)$, differentiation results in $v^2 = \omega^2 A^2 \sin^2(\omega t + \phi)$. If $\sin^2$ is averaged over one period a factor of one half results. Therefore, the average power results in

$$\langle P\rangle = \frac{m\omega_0}{Q}\langle v^2\rangle = \frac{m\omega_0}{2Q}\omega^2 A^2. \tag{2.46}$$

With this the energy dissipated per cycle is

$$\text{Energy dissipated per cycle} = \langle P\rangle T = \langle P\rangle 2\pi/\omega = \frac{\pi m\omega_0\omega A^2}{Q}. \tag{2.47}$$

Now the nominator of (2.44), i.e. the total energy stored in the oscillator will be evaluated. If we consider driving frequencies close to ω_0, the energy stored in the driven oscillator is approximately the energy of the free oscillator with the same amplitude A, as [2]

$$\text{Energy stored in the oscillator} \approx \frac{1}{2}kA^2 = \frac{1}{2}m\omega_0^2 A^2. \tag{2.48}$$

If we insert (2.47) and (2.48) into the right hand side of (2.44) and consider frequencies close to the resonance frequency $\omega \approx \omega_0$ the important expression for the quality factor (2.44) results.

2.7 Effective Mass of a Harmonic Oscillator

In this chapter, we always considered an idealized system consisting of a massless spring and a mass m at its end. However, in several cases of practical relevance this approximation is not fulfilled. For instance, in the case of a cantilever-type spring, often used in atomic force microscopy, the mass (of the cantilever) is distributed throughout the whole cantilever (Fig. 2.7b). It can not be expected that the mass distributed along the whole spring m_{spring} can simply replace the mass m at the end of a massless spring in the equations of motion. However, a properly defined effective mass m_{eff} can replace the point mass in the equations of motion.

Here we introduce the concept of the effective mass for the example of a coil spring (with mass m_{spring}) and assume that the mass is distributed homogeneously along its length and the oscillation amplitude is much smaller than the length of the

[5] This results, as the energy can be written as $E = \int F(z)\mathrm{d}z = \int F(z(t))\mathrm{d}z/\mathrm{d}t\mathrm{d}t$, and thus the power results as $P = \mathrm{d}E/\mathrm{d}t = F(t)v(t)$.

spring. In the following, we calculate the kinetic energy of the spring and we do not consider a mass M at the end of the spring.

When calculating the kinetic energy of the spring, we regard $v(z)$ as the velocity of a length element dz at the position z

$$dE_{\text{kin}} = \frac{1}{2} dm\, v^2(z) = \frac{1}{2} \frac{m_{\text{spring}}}{L} dz\, v^2(z). \tag{2.49}$$

According to Fig. 2.7a, the velocity distribution along the spring is linear with z and can be written as $v(z) = v_{\max} z/L$, with $v_{\max}$ being the maximum velocity at the end of the spring, i.e. $v(L)$. Integrating the kinetic energy along the spring results in

$$\begin{aligned} E_{\text{kin}} &= \frac{1}{2} \frac{m_{\text{spring}}}{L} \int_0^L v^2(z) \mathrm{d}z = \frac{1}{2} \frac{m_{\text{spring}}}{L} \int_0^L v_{\max}^2 \frac{z^2}{L^2} \mathrm{d}z \\ &= \frac{1}{2} \left(\frac{1}{3} m_{\text{spring}} \right) v_{\max}^2 = \frac{1}{2} m_{\text{eff}} v_{\max}^2. \end{aligned} \tag{2.50}$$

Thus, the kinetic energy of a mass-containing spring is equivalent to the one of a massless spring with an effective mass $m_{\text{eff}} = 1/3\ m_{\text{spring}}$ fixed to the end of the spring and the velocity $v_{\max}$ at the end of the spring, which is called just v in the case of a massless spring. From the calculated kinetic energy and the potential energy (in which the mass does not enter) the Lagrange function results as their difference.

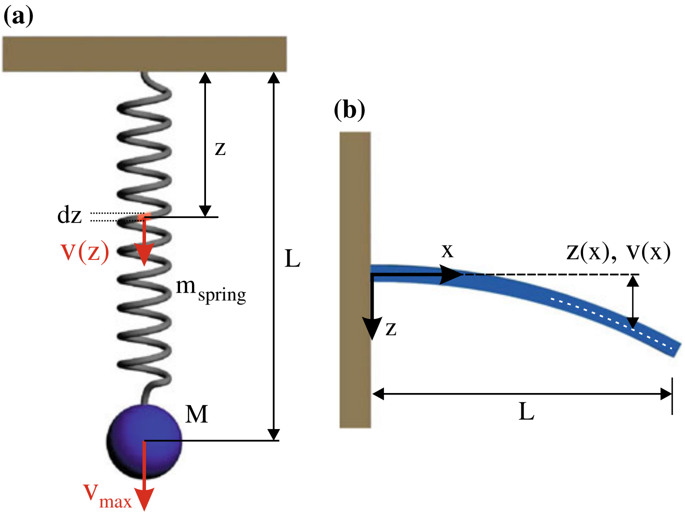


Fig. 2.7 **a** For a spring with mass m_{spring}, the velocity of a volume element depends on the position $v(z)$. The effective mass turns out to be $1/3$ of the spring mass. **b** For a cantilever beam the deflection and the velocity are non-linear as a function of x

The Euler-Lagrange equations can be used to obtain the equation of motion. Since the only difference with respect to the kinetic energy of the massless spring is the substitution of m_{eff} replacing the mass at the end of a massless spring, the Lagrange function and the resulting equation of motion are the same, as for the massless spring if the substitution for the mass is performed. As the equation of motion is the same, also the solution is the same when the effective mass is used. For instance calculating the resonance frequency of a harmonic oscillator in which the spring contains mass, the effective mass has to be used instead of the mass at the end of a massless spring as $\omega_0 = \sqrt{\frac{k}{m_{\text{eff}}}}$. If an additional mass M at the end of a spring is also considered, the effective mass becomes $m_{\text{eff}} = M + 1/3\, m_{\text{spring}}$.

For the situation of a cantilever beam the situation is more complicated, because the deflection z (in reaction to a force applied at the end of the cantilever) is not linear along the cantilever beam as shown in Fig. 2.7b. According to [3], the bending has the form $z(x) \propto -x^3 + 3x^2 L$. Since a harmonic oscillation is considered throughout the beam, the velocity distribution along the beam is proportional to the deflection $v(x) = cz(x)$. The constant of proportionality is determined by the condition $v(L) = v_{\max}$ as $c = v_{\max}/(2L^3)$. Thus, the velocity at position x along the beam results as

$$v(x) = \frac{v_{\max}}{2L^3}\left(-x^3 + 3x^2 L\right). \tag{2.51}$$

Using this expression for the velocity distribution along the beam, the (maximum) kinetic energy can be obtained by integration along the beam as

$$\begin{aligned} E_{\text{kin}} &= \frac{1}{2}\int_0^L \frac{m_{\text{cant}}}{L}\frac{v_{\max}^2}{4L^6}\left(-x^3 + 3x^2 L\right)^2 dx = \frac{1}{2}\left(\frac{33}{140} m_{\text{spring}}\right) v_{\max}^2 \\ &= \frac{1}{2} m_{\text{eff}} v_{\max}^2 . \end{aligned} \tag{2.52}$$

Thus, the effective mass for a cantilever beam turns out to be $m_{\text{eff}} = 0.2357\, m_{\text{spring}}$, instead of $m_{\text{eff}} = 1/3\, m_{\text{spring}}$ for a coil spring.

In the case of a cantilever spring, an effective mass has to be used in the equation of motion and all subsequently derived expressions such as $\omega_0 = \sqrt{k/m_{\text{eff}}}$. Throughout this text we use the concept of the harmonic oscillator and denote the mass as m in order to keep the notation simple. It has to be kept in mind that in fact the appropriate effective mass has to be used.

2.8 Linear Differential Equations

At the end of this chapter, we consider some general properties of linear differential equations with constant coefficients. A homogeneous linear differential equation up to the second order can be written as

$$a_1 x + a_2 \dot{x} + a_3 \ddot{x} = 0. \tag{2.53}$$

The following propositions hold for the homogeneous equation.

- Homogeneity: If x is a solution of the linear differential equation, Cx is also a solution.
- Superposition: If x_1 and x_2 are solutions of the linear differential equation, $x_1 + x_2$ is also a solution.
- Combining the two, we see that all linear combinations of two solutions are also solutions.

The corresponding inhomogeneous equations including an external driving force $F(t)$ can be written as

$$a_1 x + a_2 \dot{x} + a_3 \ddot{x} = F(t). \tag{2.54}$$

If we have a (special) solution of the inhomogeneous equation x_1, we can add any solution x_2 of the homogenous (free) equation $F(t) = 0$ and the sum $x = x_1 + x_2$ will be also a solution of the inhomogeneous system as we see if we add the inhomogeneous equation and the homogeneous equation as

$$a_1(x_1 + x_2) + a_2(\dot{x}_1 + \dot{x}_2) + a_3(\ddot{x}_1 + \ddot{x}_2) = a_1 x + a_2 \dot{x} + a_3 \ddot{x} = F(t). \tag{2.55}$$

Finally, we come to another important property of linear differential equations. If we have a solution x_1 for an external force $F_1(t)$ and a second solution x_2 for another external force $F_2(t)$, then a solution for the problem with the force $F_1(t) + F_2(t)$ is $x_1 + x_2$. This superposition principle is remarkable and is the basis for decomposing a complicated (arbitrary) force into Fourier components and composing the solution of the problem with a complicated force as a superposition of the solutions obtained for simple harmonic forces. This is also a late justification for why we only considered an external excitation (force) of simple harmonic form for the harmonic oscillator.

2.9 Summary

- The free harmonic oscillator has the resonance frequency of $\omega_0 = \sqrt{\frac{k}{m}}$.
- The driven harmonic oscillator oscillates at the driving frequency ω with an amplitude depending on ω and ω_0.
- If $\omega = \omega_0$ the amplitude becomes very large (resonance).
- For the damped driven oscillator the amplitude at resonance is damped with increasing damping force $F_{\text{frict}} = -m\frac{\omega_0}{Q}\dot{z}$.
- The phase shift between driving excitation and oscillation is zero if $\omega \ll \omega_0$, it is $-90°$ if $\omega = \omega_0$, and $-180°$ if $\omega \gg \omega_0$.
- The quality factor of the oscillation Q is 2π times the ratio of the energy stored in the oscillator to the energy dissipated per cycle. The Q-factor determines the

height ($A(\omega_0) = QA_{\text{drive}}$) and the width ($\delta\omega = \omega_0/Q$) of the resonance curve of a driven damped oscillator.

- The build up or the decay of the steady-state amplitude takes about Q oscillations, i.e. the corresponding time constant for the decay to $1/e$ is $\tau = 2Q/\omega_0$.
- If a spring has a non-negligible mass, the effective mass has to be used in the equation for the resonance frequency of the harmonic oscillator.

References

1. W. Greiner, *Classical Mechanics–Point Particles and Relativity*, 1st edn. (McGraw-Hill, New York, 2004). https://doi.org/10.1007/b97649
2. R.P. Feynman, R.B. Leighton, M. Sands, *The Feynman Lectures on Physics - Definitive Edition*, vol. 1 (Pearson, Addison-Wesley, Melno Park, 2006). ISBN 0-8053-0946-4
3. W.D. Pilkey, *Formulas for Stress, Strain and Structural Matrices*, 2nd edn. (Wiley, New York, 2005). https://doi.org/10.1002/9780470172681

Chapter 3
Technical Aspects of Atomic Force Microscopy

In order to perform nanoscale motions in AFM (e.g. during scanning) very precise actuators are required. Piezoelectric actuators achieve the required precision. We describe the principles of operation of these actuators and present examples of specific actuators. In the following principles of vibration isolation are considered, because the amplitude of floor vibrations is much larger than the desired amplitude of the tip-sample vibrations.

3.1 Piezoelectric Effect

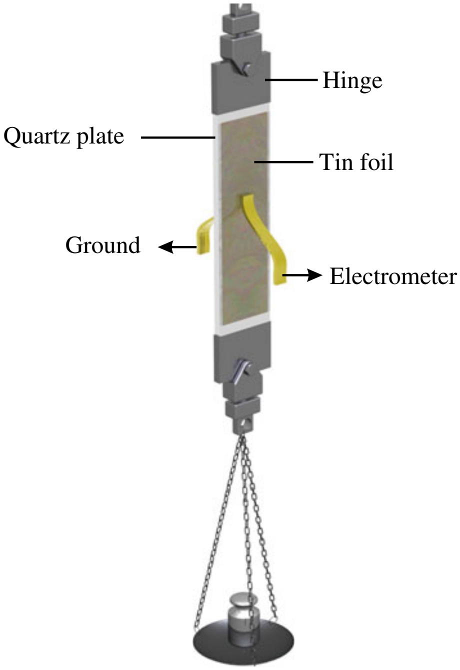
In order to position the probe tip or the sample, piezoelectric elements are used as actuators. The piezoelectric effect was discovered by the Curie brothers in 1880. A sketch of their experiment is shown in Fig. 3.1. Tin foils were attached as electrodes to two sides of a quartz plate. One tin foil was grounded and one connected to an electrometer. While a force was applied to generate vertical strain, an electrical charge was detected by the electrometer. The piezoelectric effect is used, for instance, to ignite pocket lighters (generating the voltage which generates the lightning spark) and many other technical applications such as sensor technology.

The converse effect occurs if a variable voltage is applied to the foils and a deformation of the crystal results. The converse piezoelectric effect is used in piezoelectric actuators. Since this deformation is very small and a continuous quantity, deformations much smaller than the diameter of an atom can be obtained for reasonably small voltages in the mV range.

In order to apply an external electric field inside the (electrically insulating) piezoelectric material, metallic electrodes at the surface are used. A voltage applied to the electrodes induces an electric field in the piezo material (as in a capacitor with a dielectric) and finally results in an extension of the piezo material. Vice versa, a strain of the piezo material leads to a surface charge and thus to a charge on the electrodes, and finally to a voltage between the electrodes.

B. Voigtländer, *Atomic Force Microscopy*, NanoScience and Technology, https://doi.org/10.1007/978-3-030-13654-3_3

Fig. 3.1 Curie brothers' experiment demonstrating the piezoelectric effect



The piezoelectric effect occurs only for crystals which are not centrosymmetric, i.e. do not have an inversion center. If an inversion center exists no net electric dipole moment can be induced inside the unit cell by straining the crystal. If a dipole moment is present at a position **r** inside the unit cell, the opposite dipole is also present at the position −**r** due to the inversion symmetry and the net dipole moment of the unit cell is zero. In a piezoelectric material however, a directional deformation leads to uncompensated microscopic dipoles inside the crystallographic unit cell. These microscopic dipoles lead to a charge at the surface of the crystal and a corresponding electric field inside the crystal. In the converse piezoelectric effect, the crystal unit cell is deformed by an external applied electrical field. An example of a piezoelectric material is crystalline quartz. Another example of a piezoelectric material used in piezoelectric actuators is PZT ceramics (lead zirconate titanate $\mathrm{Pb[Zr_xTi_{1-x}]O_3}$). PZT is piezoelectric and also ferroelectric, which means that there is a permanent net electric dipole even in the absence of any externally applied mechanical stress.

In the following, we explain the principle of the piezoelectric effect on the atomic scale using the example of a PZT unit cell. The unit cell, which is shown schematically in Fig. 3.2a, consists of Pb^{2+} at the corners of the unit cell, O^{2-} at face centered positions on the outer faces of the unit cell, forming an octahedron, and Ti^{4+} displaced from the center of the unit cell. In Fig. 3.2b, the unit cell is shown from the side with an arrow indicating the direction and size of the permanent electric dipole moment. The electric dipole inside the unit cell results in a net charge at the surfaces (xy-planes) of

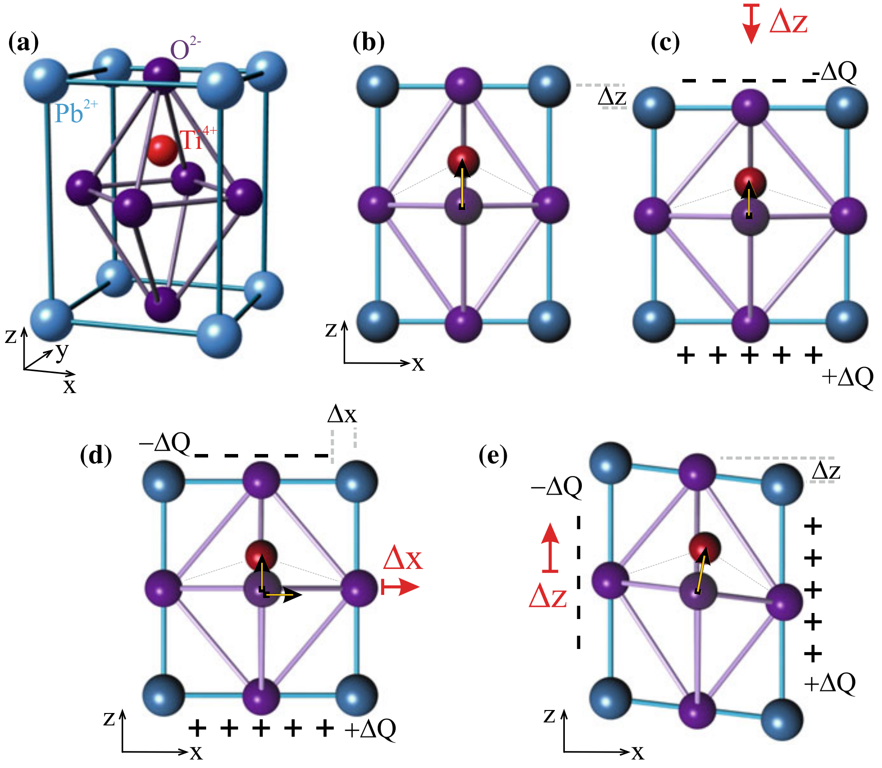


Fig. 3.2 **a** Schematic of the PZT unit cell. **b** *Side view* of the PZT unit cell with the dipole induced by the displaced Ti^{4+}. **c** Longitudinal piezoelectric effect: upon compression of the unit cell along the z-axis the magnitude of the dipole is reduced leading to a corresponding change of the surface charge. **d** Transverse piezoelectric effect: strain along the x-axis leads, due to the Poisson effect, to a change of the dipole along the z-direction and a corresponding change of the surface charge. **e** Shear piezo effect: a shear strain along the z-direction leads to a change of the x-component of the dipole and a corresponding change of the surface charge

the piezoelectric PZT material, as in the case of a capacitor with a dielectric material inside. The direction along which the permanent dipole moment points is taken as the z-direction and the material is said to be poled along the z-direction.

When the piezoelectric material is strained in the poling direction (e.g. compressed, as shown in Fig. 3.2c), the magnitude of the electric dipole moment decreases and correspondingly the electric field inside the material and the surface charge decrease. This case, where the strain is applied along the poling direction (z-direction) leading to a voltage between the two opposite xy-surface planes, is called the *longitudinal piezoelectric effect*.

The case in which the external strain is applied perpendicular to the poling direction (x-direction) is shown in Fig. 3.2d. In spite of the fact that the crystal is compressed in the x-direction, no dipole moment occurs in x-direction (nor in the

y-direction), because there is a mirror symmetry. For every atom there is an atom at the $-x$ position inside the unit cell canceling the net dipole moment along the x-direction. However, due to the Poisson effect any strain in x-direction also leads to a corresponding transverse strain in the z-direction (as well as in the y-direction). This strain in the z-direction will lead to a change of the dipole moment in z-direction and to a corresponding change of the surface charge on the xy-surface planes. This piezoelectric effect in which a strain along the x-direction results in a change of the dipole moment in z-direction is called the *transverse piezo effect*.

If a shear strain is applied along the z-direction, as shown in Fig. 3.2e, the dipole turns and induces a change of the component of the dipole moment in the x-direction and a corresponding build up of surface charge. This effect is called the *shear piezoelectric effect*. In the first order, the dipole moment in the z-direction does not change.

In the preceding we discussed the piezoelectric effect. However, the reverse reasoning also applies for the converse piezoelectric effect where a voltage applied to the outer metallic electrodes results in a strain of the piezoelectric material [1]. The charge on the outer metallic electrodes leads to a change of the dipole moment in the ferroelectric material. This corresponds to a capacitor with a dielectric, where an charge on the capacitor plates induces a polarization and a corresponding surface charge. In the case of a piezoelectric material the dielectric is already polarized without an outer electric field applied. The change of the dipole moment (change of the polarization) induces in piezoelectric materials a corresponding strain. This direction of the piezoelectric effect is relevant for piezoelectric actuators. In the following, we describe the strain produced in different types of piezoelectric actuators induced by a voltage applied to their electrodes.

3.2 Extensions of Piezoelectric Actuators

If a voltage ΔV is applied across a rectangular piece of piezoelectric material (Fig. 3.3a) of dimensions x, y, and z (poled in z-direction) the external applied electric field is, due to the plate capacitor configuration, $\mathscr{E}_3 = \Delta V/z$. In practical terms the field is applied to a piece of piezoelectric material via the metallic electrodes at the surfaces of the piezo element. Often the directions x, y, and z are labeled as 1, 2, and 3, respectively. The direction of the poling field is labeled as direction 3, or as the positive z-direction. As a result of the applied electric field, a strain is generated along the z-direction and also, via the transverse elongation of the material (Poisson effect), a transverse strain in the x-direction (as well as in the y-direction). If a piezo plate of thickness z (Fig. 3.3a) is strained in the z-direction by Δz, the corresponding strain is $S_3 = \Delta z/z$. The strain in x-direction is $S_1 = \Delta x/x$. The same also applies for the y-direction.

The mechanical strain developed in a piezoelectric material is known to be proportional to the applied electric field, with the *piezoelectric coefficients* as proportionality constants. The piezoelectric coefficients are material constants which depend, however, on the direction along which the electric field is applied and on the direction

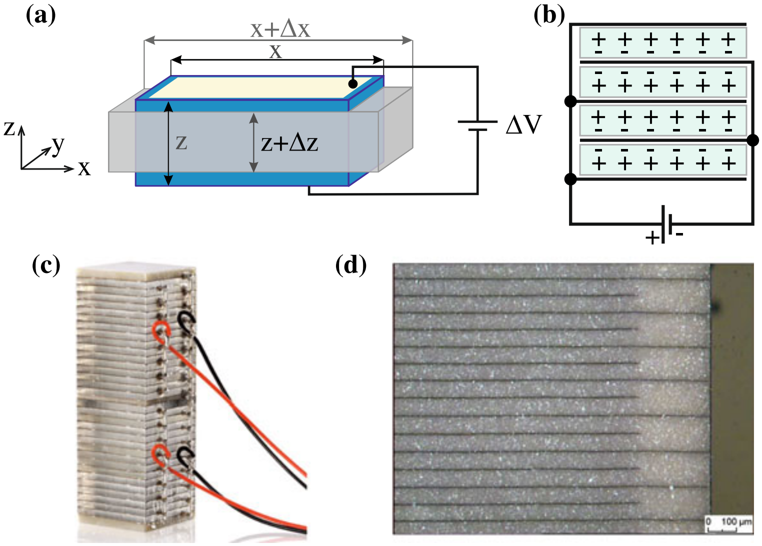


Fig. 3.3 **a** Sketch of a piezo plate (dimensions x, y, and z) poled in the z-direction. Considering the longitudinal piezo effect, an electric field in the z-direction induced by a voltage ΔV in z-direction induces a strain in z-direction, Δz. Considering the transverse piezoelectric effect a voltage in the z-direction also induces a strain in the x-direction, and also of course in y-direction. In this case, the piezo constant is proportional to the length x of the plate. **b** Since for the longitudinal piezo effect the piezo coefficient is independent of the plate thickness z, several plates have to be stacked on top of each other in order to tune (enhance) the piezo constant. **c** Photo of piezoelectric stack actuators made by gluing together single piezo plates. **d** Monolithic stack actuators with much smaller layer thickness of about 60 µm in this case (reproduced with permission from PI Ceramic [2])

along which the strain is considered. The piezoelectric coefficients are defined as the ratios of the strain components (in a certain direction) over the component of the applied electric field (in a certain direction), for example for the longitudinal piezo effect

$$d_{33} \equiv \frac{S_3}{\mathscr{E}_3}, \quad \text{while} \quad d_{31} \equiv \frac{S_1}{\mathscr{E}_3} \tag{3.1}$$

is the piezoelectric coefficient which applies in the case of the transverse piezoelectric effect. Because strain is a dimensionless quantity, the piezoelectric coefficients have dimensions of meter/volt. Their values are extremely small. For applications in scanning probe microscopy, a natural unit is Å/V. Since the voltage difference at the electrodes and the corresponding charge difference are related to the work ΔU which has to be supplied to put charge to the electrodes by $\Delta V = \frac{\Delta U}{\Delta Q}$, equivalent units for the piezoelectric coefficients are also coulomb/newton. This is also equivalent to the induced charge density (C/m^2) per applied stress (N/m^2).

While the piezoelectric coefficients are material properties the *piezo constant* is assigned to a specific actuator element with specific dimensions, and the electric field applied along a specific direction, and the strain considered in a specific direction. The piezo constant of an actuator is the ratio between the amount of motion in a certain direction and the voltage applied between the electrodes, e.g. $\Delta z/\Delta V$.

As a first example, a piezoelectric plate shown in Fig. 3.3a serves as our piezoelectric actuator, with the electric field applied along the z-direction (poling direction), and the strain considered in the z-direction as well. There is also strain present in the x-direction, which we will analyze later. The piezo constant $\Delta z/\Delta V$ can be calculated as follows

$$\frac{\Delta z}{\Delta V} = \frac{\Delta z/z}{\Delta V/z} = \frac{S_3}{\mathscr{E}_3} = d_{33}. \tag{3.2}$$

The piezo constant for motion of a piezo plate in the z-direction (induced by the longitudinal piezo effect) is not dependent on the thickness of the piezo plate z. The z-dependence in (3.2) is canceled out due to same dependence of both the electric field and the strain on z. This means the piezo coefficient of a plate cannot be tuned by changing its thickness (or, of course, also the diameter). The only way to tune or enhance the length extension per voltage is to stack several piezo plates on top of each other as shown schematically in Fig. 3.3b. With common electrodes in between the plates, neighboring plates have to have opposite poling and the electrical connections to the electrodes have to be as indicated in Fig. 3.3b. A photo of this type of piezo actuator known as a piezoelectric stack actuator, produced by the company PI, is shown in Fig. 3.3c. The net displacement is the sum of the displacements of the individual piezo plates. The dimensions of the piezoelectric stack actuators are very flexible. Typical dimensions are in the mm range for the thickness of a single plate and in the cm or even decimeter range for the height of the stack. Quite large piezo constants can be achieved in this way (corresponding to a displacement of 10 μm for a stack height of 10 mm).

There are actually two types of piezoelectric stack actuators. The first type consists of plates about half a mm in thickness, which are glued together to form a stack (Fig. 3.3c). Such stack actuators are characterized by high operating voltages of up to 1,000 V and low capacitances in the nF range. On the other hand, there are monolithic stack actuators which are characterized by a much smaller piezoelectric layer thickness (∼60 μm) as shown in Fig. 3.3d. These monolithic actuators are manufactured using a cofiring technology during sintering of the piezoelectric material together with the electrodes. This type of actuator has a lower operating voltage of about 120 V. The disadvantage of such a piezo actuator is its quite high capacity, in the μF range. If a quick extension of the actuator is required, quite high charging currents have to be supplied.

In a different kind of piezoelectric actuator, the elongation of a piezo plate in x-direction due to the transverse piezoelectric effect can be exploited (Fig. 3.3a). The piezo constant for the motion along the x-axis can be obtained as

$$\frac{\Delta x}{\Delta V} = \frac{\Delta x/x}{\Delta V/z}\frac{x}{z} = \frac{S_1}{\mathscr{E}_3}\frac{x}{z} = d_{31}\frac{x}{z}. \tag{3.3}$$

In this case, the piezo constant depends on the dimensions of the plate. The piezo constant is proportional to the length x of the piezo element and inversely proportional to its thickness z. Using the transverse piezo effect, the piezo constant of the actuator can be tuned by its dimensions. To obtain a large piezo constant a long piezo or a thin piezo element can be used. However, long, thin piezo elements lead to low resonance frequencies of the bending vibration, which is disadvantageous for stable AFM operation, as we will see later. Also for a small thickness, the electric field rises and may approach the allowed limits of the material. While we have considered a piezoelectric plate here, the most frequently used shape for a piezoelectric actuator based on the transverse piezo effect is the piezo tube, which we will consider in detail later. A piezo tube can be imagined as a plate which is rolled up to form a tube.

Of course, in a piezoelectric plate both piezoelectric effects (the longitudinal and the transverse) occur simultaneously. In both of the previous cases we focus on one effect and neglect the other due to the specific direction of the extension we are looking at. When discussing the longitudinal piezo effect of a plate we focus on the change of the thickness of the plate and do not consider the change in the width of the plate due to the transverse effect. On the other hand, when we focus on the transverse extension of a plate, we do not consider the change of the thickness of the plate.

In Fig. 3.4a a piezoelectric plate is shown which is poled in the z-direction (horizontal in this case) while the electric field (voltage) is applied along the x-direction, i.e. vertical. As we have seen in Fig. 3.2e, this configuration leads to a shear strain along the z-direction with $S_{\text{shear}} = \Delta z / x$. The piezo constant can be written as

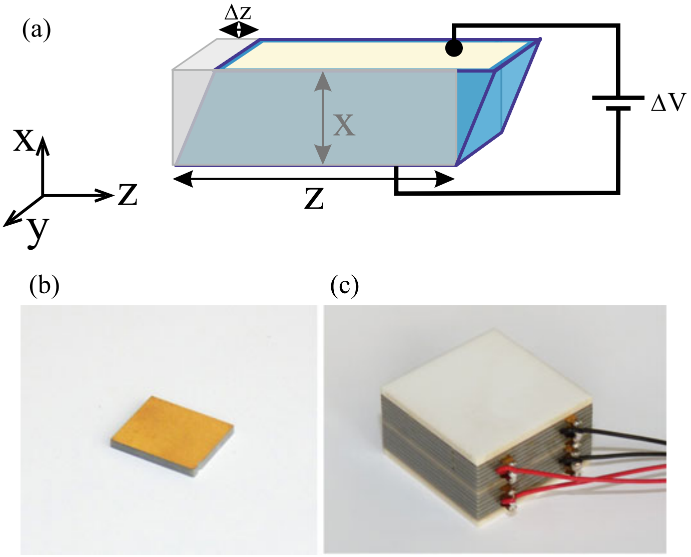


Fig. 3.4 **a** Sketch of a piezoelectric plate operated using the shear piezo effect. **b** Photo of a single shear piezo plate (6 mm × 7 mm). **c** Photo of a shear piezo stack (15 mm × 15 mm)

$$\frac{\Delta z}{\Delta V} = \frac{\Delta z/x}{\Delta V/x} = \frac{S_{\text{shear}}}{\mathscr{E}_1} \equiv d_{15}. \tag{3.4}$$

The corresponding piezo coefficient is called (due to some conventions) d_{15}. Thus, the piezo constant results as $\frac{\Delta z}{\Delta V} = d_{15}$. As in the case of the longitudinal effect, the piezo constant does not depend on the plate dimensions. Therefore, stacks of shear piezo elements are often used here as well. Shear piezos are attractive piezo elements as they induce a uniform lateral motion of their surface. As shown in Fig. 3.4b, shear piezos have a size of only a few millimeters. If shear piezo elements are stacked onto each other and rotated by 90°, motions in two orthogonal directions can be performed as shown in Fig. 3.4c.

3.3 Piezoelectric Materials

Initially, the piezoelectric effect was observed in crystalline materials, for instance in quartz. However, for use in piezoelectric actuators, single crystals are inconvenient. Today mostly lead zirconate titanate ceramics (PZT, $\mathrm{Pb[Zr_xTi_{1-x}]O_3}$) are used as materials for piezoelectric actuators because ceramics can be formed into various shapes and because of their large piezo coefficients. These materials are ferroelectric, which means they exhibit a permanent electric dipole even in the absence of an external electric field. The unit cell of PZT has an anisotropic structure below the Curie temperature, i.e. elongated in one direction as shown in Fig. 3.5a. Above the Curie temperature T_c, the crystal structure becomes cubic and the material loses its piezoelectric properties Fig. 3.5b.

Directly after sintering, piezoelectric ceramics does not exhibit a piezoelectric effect. This is due to two reasons: first the ceramic is a polycrystalline material with

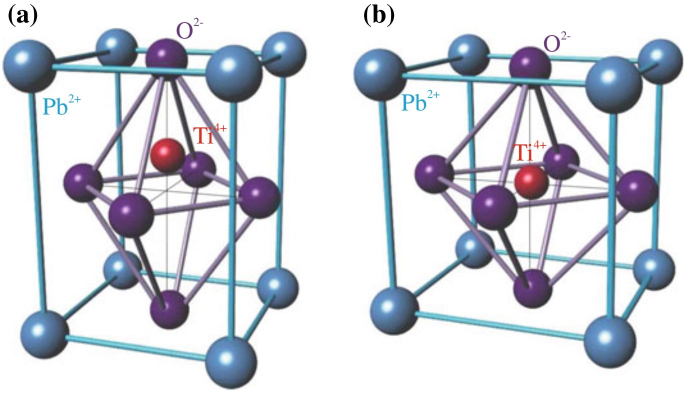


Fig. 3.5 Unit cell of the PZT crystal structure **a** below the Curie temperature **b** above the Curie temperature

Table 3.1 Some properties of piezoelectric materials

Material	PZT-5A	PZT-5H	PZT-8
d_{31} (Å/V)	−1.75	−2.50	−1.00
d_{33} (Å/V)	3.90	6.50	3.00
d_{15} (Å/V)	5.70	7.30	3.25
T_c (°C)	360	220	300
Density (g/cm^3)	7.7	7.7	7.6
Young's modulus (10^{10} N/m^2)	5.7	6.3	8.9
Q	90	100	1,200

randomly oriented crystallites and second also within a single crystallite there are different domains. Inside a domain the dipoles within the unit cell are oriented in parallel, while differently oriented domains exist in one crystallite as in the case of ferromagnetism. These domains are randomly oriented in the raw piezoelectric material when it is cooled below the Curie temperature after sintering. Ferroelectric ceramics become macroscopically piezoelectric when poled. This means an electric field (>2,000 V/mm) is applied to the piezoelectric ceramics at temperatures somewhat below the Curie temperature. Close to the Curie temperature the crystal structure is almost cubic. With a field applied, the electric dipoles can switch (by motion of the Ti atom) to one of the six possible directions (Fig. 3.5b) which lies closest to the applied electric field. During poling, the domains can reorient and the domain walls can also move. These domains stay roughly in alignment after cooling. The material now has a remanent alignment of the dipoles, which can be degraded by exceeding the mechanical, thermal and electrical limits of the material.

Some material properties of different piezoelectric materials are listed in Table 3.1. The PZT nomenclature for the materials in Table 3.1 is an industry standard to which several companies producing piezoelectric materials refer. However, the numbers should be considered only as rough estimate since the actual values vary from manufacturer to manufacturer. The *Curie temperature* T_c is the temperature above which the material loses its piezoelectric properties irreversibly (like a ferromagnetic material). Each material has a maximum operating temperature specified by the supplier, which is often well below the Curie temperature. The *mechanical quality factor Q* determines the sharpness of the mechanical resonance and the resonance amplitude of an actuator made from this material.

The material properties of the piezoelectric materials are also temperature-dependent. Most importantly the piezoelectric coefficients decrease for operation at low temperatures as shown in Fig. 3.6 [3] for the example of PZT-5A. As a rule of thumb, the piezo constants are for most piezo materials are roughly a factor of five lower at the temperature of liquid helium than at room temperature.

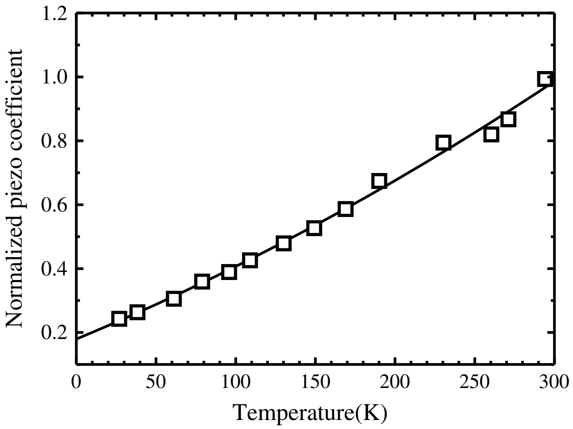


Fig. 3.6 Temperature dependence of the piezoelectric constants d_{31} for PZT-5A piezo ceramic material relative to the room temperature value (adapted from [3])

3.4 Tube Piezo Element

One central task in atomic force microscopy is to position the probe with an accuracy of less than one tenth of an ångström in all three dimensions. The tube piezo element (or tube scanner) is the most widely used actuator element to move the probe tip or the sample in order to scan a surface (fine motion). One single tube piezo element allows motions to be performed in three orthogonal directions. Further advantages are high piezo constants and high resonance frequencies. The tube scanner consists of a tube, made of piezoceramics (poled in radial direction), which is covered inside and outside with metal electrodes. The outer electrode is divided into four quadrants, as shown in Fig. 3.7. A motion in the z-direction (along the longitudinal axis) can be achieved by applying a voltage between the inner and all outer electrodes (Fig. 3.7b). A deflection in the xy-direction is induced by voltages of opposite polarity applied to the two opposite outer electrodes Fig. 3.7c. Due to the transverse piezoelectric effect, one segment of the tube extends along the tube axis, while the opposite segment shrinks, giving rise to a bending of the upper part of the tube, as shown in Fig. 3.7c. When a tube scanner is used to scan a tip, the tip (holder) is mounted axially on top of the tube scanner.

The vertical displacement $\Delta L = \Delta z$ of the top of the tube piezo element is calculated using (3.3) (exchanging the definitions of the directions x and z), leading to the following piezo constant

$$\frac{\Delta z}{\Delta V} = d_{31}\frac{L}{h}. \tag{3.5}$$

In order to obtain the lateral displacement Δx of the tube, we assume that the bending of the tube follows a circular arc as shown in Fig. 3.8. From this figure, we identify (due to the definition of the arc length) the bending angle as

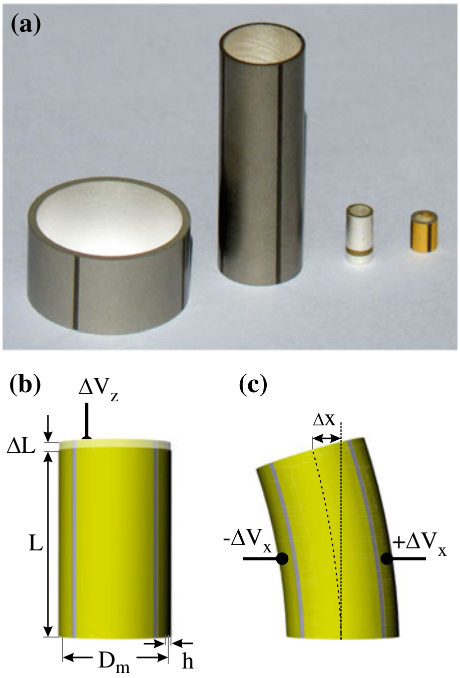


Fig. 3.7 **a** Photograph of several tube piezo elements. **b** Schematic *side view* of a tube scanner showing the vertical extension along z. **c** Schematic of the lateral movement in the x-direction

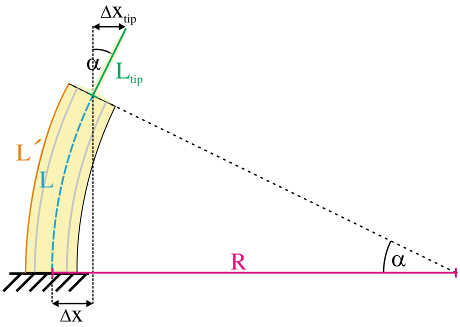


Fig. 3.8 Sketch of the geometry of a bent piezo tube with the relevant parameters

$$\alpha = \frac{L}{R}. \tag{3.6}$$

Further, we identify $L' = L + \Delta L$, which can also be written as

$$L' = \alpha \left(R + \frac{D_m}{2} \right) = L + \alpha \frac{D_m}{2}. \tag{3.7}$$

This results in

$$\alpha = 2\frac{\Delta L}{D_m}, \tag{3.8}$$

with D_m being the mean diameter of the tube. From Fig. 3.8 we also determine that the cosine of the bending angle can be written as

$$\frac{R - \Delta x}{R} = \cos\alpha \approx 1 - \frac{\alpha^2}{2}. \tag{3.9}$$

Thus, the x-deflection of the tube is given by

$$\Delta x = \frac{R\alpha^2}{2}. \tag{3.10}$$

Replacing R using (3.6) and (3.8) results in the following expression for the x-deflection of the tube

$$\Delta x = \frac{\Delta L L}{D_m}. \tag{3.11}$$

For the length extension ΔL of the piezo tube we can make the simplified assumption that it is the vertical length extension Δz according to (3.5). With this assumption the piezo constant for the x-deflection results as

$$\frac{\Delta x}{\Delta V} = \frac{d_{31}L^2}{D_m h}. \tag{3.12}$$

A better approximation for the length extension ΔL, which considers non uniform stress in the electrodes due to bending, is considered in Appendix A and results in the following expression for the piezo constant for horizontal bending

$$\frac{\Delta x}{\Delta V} = \frac{2\sqrt{2}}{\pi}\frac{d_{31}L^2}{D_m h}. \tag{3.13}$$

This equation corresponds to the bipolar operation of the tube where voltages $-\Delta V$ and $+\Delta V$ are applied to opposite electrodes.

If we consider as an example particular dimensions of a piezo tube (PZT-5A) as follows: length 25.4 mm, mean diameter 5.84 mm, wall thickness 0.51 mm, this results in a piezo coefficient for x and y directions of 725 Å/V and for the z-direction of 90 Å/V. The most effective design parameter to tune the piezo coefficient is the length of the tube, as the xy-piezo coefficient is quadratically dependent on the tube length.

What we have considered up to now is the deflection of the top of the piezo tube. However, if a tip is mounted on a scanner tube, it is usually mounted at a distance L_{tip} above the center of the piezo tube. In this case, an additional deflection Δx_{tip} results, which can be written according to Fig. 3.8 and using (3.6), (3.8), and (A.5) as [4]

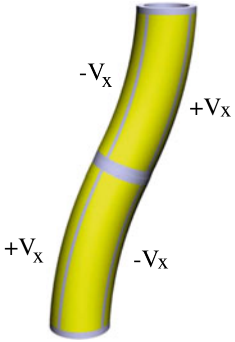


Fig. 3.9 Instead of an outside electrode divided into four segments the outer electrode has eight segments. The *upper part* of the piezo is bent in the opposite direction to prevent a displacement in the z-direction

$$\Delta x_{\text{tip}} = L_{\text{tip}} \sin\alpha \approx L_{\text{tip}}\alpha = L_{\text{tip}} \frac{2\Delta L}{D_m} = L_{\text{tip}} \frac{4\sqrt{2}}{\pi} \frac{d_{31} L \Delta V}{D_m h}. \tag{3.14}$$

Combining this with (3.13), the total piezo constant for the horizontal deflection results in

$$\frac{\Delta x_{\text{tot}}}{\Delta V} = \frac{\Delta x + \Delta x_{\text{tip}}}{\Delta V} = \frac{2\sqrt{2}}{\pi} \frac{d_{31} L_{\text{piezo}}}{D_m h} \left(L_{\text{piezo}} + 2L_{\text{tip}}\right), \tag{3.15}$$

denoting the length of the piezo tube as L_{piezo}.

One disadvantage of the tube scanner is the fact that x, y and z motions are not completely decoupled. The x, y motion acts approximately on a sphere. Therefore, every lateral motion also results in a slight motion in the z-direction and vice versa. This is because the tube scanner relies on bending and not on linear motion. There is a method to prevent this coupling [5]. As shown in Fig. 3.9, a z displacement can be prevented during an xy-motion by an opposite bending in the upper part of the piezo which now has eight electrodes on the outer side. With this trick, a coupling of the xy-displacement to the z-displacement is eliminated. The disadvantage of this type of scanner is that the scan range in x and y direction is reduced by a factor of two for a given piezo length. Also the electrode structure and the cabling are more complicated.

3.4.1 Resonance Frequencies of Piezo Tubes

Here we summarize equations for the resonance frequencies of tubes, and also of beams such as those used as cantilevers in atomic force microscopy, taken from [6]. These equations are obtained using the assumptions underlying the (classical) Euler-Bernoulli beam theory, which are the proportionality of stress and strain (small bending), as well as the condition that a plane cross section of the beam remains plane under bending, i.e. shear deformations are ignored. As a boundary condition it is assumed that one end of the tube (beam) is rigidly fixed to a rigid wall.

The frequency of the ith longitudinal (axial) vibrational stretching mode of a rod or tube with one end clamped and one end free is

$$f_{\text{stretch},i} = \frac{\lambda_i}{2\pi L}\sqrt{\frac{E}{\rho}}, \tag{3.16}$$

where L is the length of the beam, ρ is its volume density, and E Young's modulus.[1] The value of λ_i for the ith resonance is given by $\lambda_i = \pi/2 \cdot (2i - 1)$. For the lowest resonance ($i = 1$) the stretching frequency results as

$$f_{\text{stretch},1} = \frac{1}{4L}\sqrt{\frac{E}{\rho}} = \frac{c}{4L}, \tag{3.17}$$

where c is the longitudinal velocity of sound, which is given in long rods as $c = \sqrt{E/\rho}$. For a mass M at the end of the beam (tube) the following expression holds for the lowest axial resonance frequency

$$f_{\text{stretch},1} \approx \frac{1}{2\pi}\sqrt{\frac{AE}{ML}}, \tag{3.18}$$

with A being the cross sectional (material-containing) area of the beam (tube).

The resonance frequencies of the bending modes of a beam (perpendicular to the beam axis) clamped at one end and free at the other end are given by

$$f_{\text{bend},i} = \frac{\lambda_i^2}{2\pi L^2}\sqrt{\frac{EI}{\rho A}} = \frac{\lambda_i^2\kappa}{2\pi L^2}\sqrt{\frac{E}{\rho}}. \tag{3.19}$$

The values for λ_i are 1.875 and 4.694 for the first two modes, respectively. The dimensions of the beam enter into the area moment of inertia (also called second moment of inertia) $I = \int x^2 \mathrm{d}A$, where x is the direction of bending. The expression $\sqrt{I/A} = \kappa$ is called the radius of gyration and has the following expressions: for a circular rod $\kappa = D/4$, for a tube $\kappa = \sqrt{D^2 + d^2}/4$, with D being the outer diameter and d inner diameter. For a tube with negligible wall thickness $\kappa = D/(2\sqrt{2})$ results, and for a beam with rectangular cross section (with width w and thickness t) $\kappa = \frac{1}{12}wt^3$ results for bending in the direction of the thickness.

With an additional mass M at the end of the beam and the mass of the beam m, the first resonance frequency can be expressed as

$$f_{\text{bend}} = \frac{1}{2\pi}\sqrt{\frac{3EI}{L^3(M + 0.2357\text{m})}}. \tag{3.20}$$

[1] In tables sometimes also the elastic compliance S is used, which corresponds to the reciprocal of Young's modulus.

Simple numeric estimates for the resonance frequencies are obtained from these equations. As an example, we consider the lowest bending frequency of a tube. Following (3.19) the bending frequency results as

$$f_{\text{bend}}^{\text{tube}} = \frac{0.56\sqrt{D^2 + d^2}}{4L^2}\sqrt{\frac{E}{\rho}}. \tag{3.21}$$

For a PZT-5A tube with the dimensions length 12 mm, outer diameter 3.2 mm, and inner diameter 2.2 mm, the calculated resonance frequencies are 56 and 10.1 kHz for the stretching and the bending mode, respectively. These resonance frequencies can

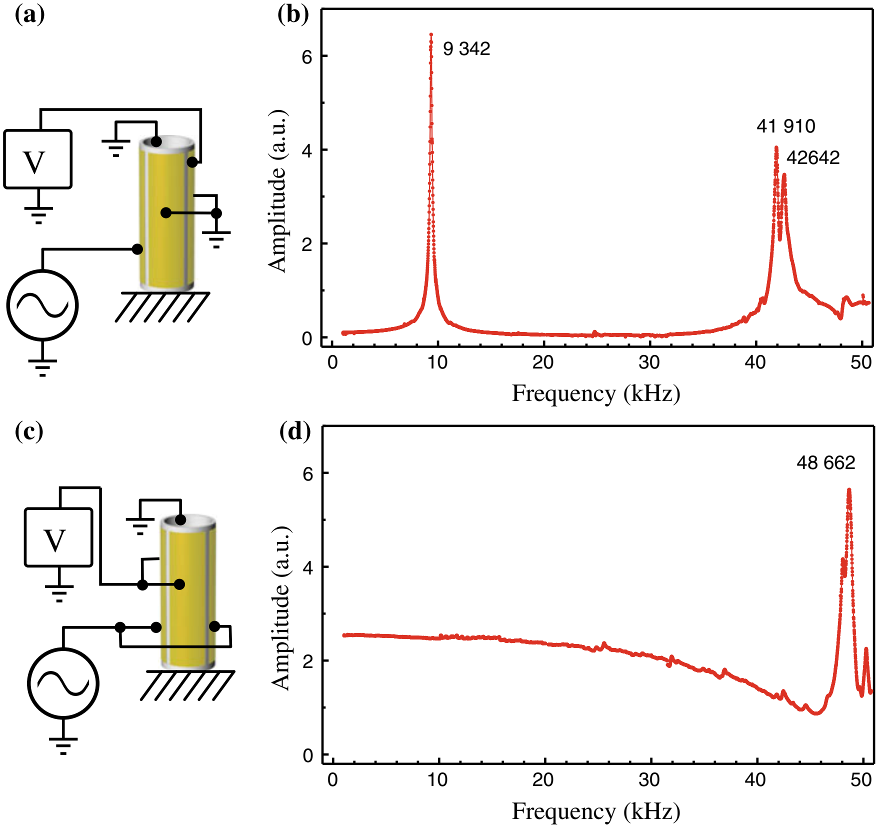


Fig. 3.10 **a** Schematic of the measurement setup with an electric excitation of the mechanic oscillation of a tube piezo element (bending mode). The amplitude of the mechanically excited oscillation is detected by the piezoelectric effect. **b** Amplitude of the mechanic oscillation. Resonances are observed at the first bending mode at 9.3 kHz and at the second bending mode around 42 kHz. **c** Schematic setup for the excitation of the stretching mode. **d** The first stretching resonance frequency is measured at 49 kHz

also be measured experimentally in a setup like the one shown in Fig. 3.10a. An AC voltage is applied to one of the four outer electrodes. Due to the piezoelectric effect the tube bends and a voltage is induced by the piezoelectric effect on the opposite electrode (the two other outer electrodes and the center electrode are grounded, as shown in Fig. 3.10a). This kind of excitation excites the bending modes. The first bending resonance is measured at 9.3 kHz (Fig. 3.10b), which corresponds roughly to the calculated value of 10.1 kHz. The higher frequencies around 42 kHz correspond to the second bending mode and do not correspond so well to the calculated value of 62 kHz. Figure 3.10c shows the configuration for the excitation of the stretching mode. The measured frequency of 49 kHz corresponds roughly to the calculated frequency of 56 kHz.

Generally, the bending resonance frequencies are overestimated by the equations for two reasons: the neglect of shear forces in the Euler-Bernoulli theory and the idealized boundary conditions. At one end, the tube (beam) is considered to be fixed rigidly to a stiff support. However, the support has some elasticity and, if the tube is glued to the support, also its elasticity enters into the considerations.

If tube piezos have been depolarized, e.g. by too high temperature, they can be repolarized by applying a DC voltage between the inner and outer electrodes (the polarity should be the same as during poling, which is different for different manufacturers). The necessary voltage depends on the wall thickness of the tube. An electric field of about twice the coercitive field (cf. Fig. 3.11) should be used for several hours at room temperature, or rather at elevated temperature but still below the Curie temperature.

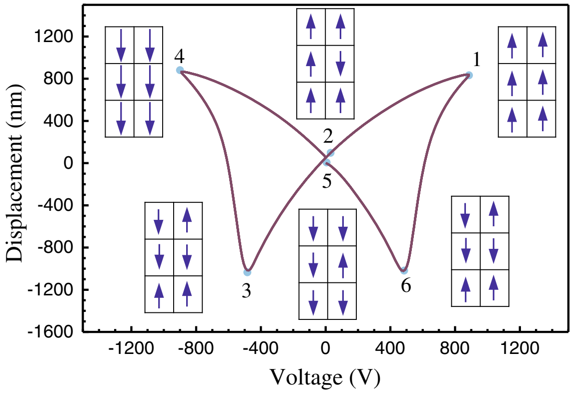


Fig. 3.11 The *butterfly curve* of the piezoelectric material PIC 151 [2] for the applied field and the displacement, both in 3-direction. The strain is shown in dependence of the applied electric field for large electric fields. The corresponding polarization of ferroelectric domains is also indicated in a simplified scheme. The *butterfly curve* shown here was kindly measured by aixACCT [7]

3.5 Non-linearities and Hysteresis Effects of Piezoelectric Actuators

The positioning performance of piezoelectric actuators is limited by the effects of non-linearities, hysteresis, and creep, which will be discussed in the following. The simplest non-ideal property of piezoelectric actuators is the non-linearity of the motion as function of the applied voltage, as any linear effect becomes non-linear at larger amplitudes. More complicated effects are hysteresis and creep, as they depend on the history of the system.

3.5.1 Hysteresis

There are mainly two contributions which lead to a strain of a piezoelectric ceramic in the presence of an outer electric field. The intrinsic effect results from the displacement of the ions inside the crystal lattice in the presence of an electric field, as shown in Fig. 3.2. This effect is approximately linear and non-hysteretic.

A second extrinsic contribution results from the reorientation of the ferroelectric domains present in the crystal lattice. A ferroelectric ceramic consists of sintered crystallites which have a random orientation of their crystalline lattice. Inside a crystallite, ferroelectric domains with different orientations exist as follows. As seen in Fig. 3.5, the Ti ion in the crystal lattice can move in six different directions, and domains with six different orientations (ferroelectric domains) can exist in the crystal lattice. The ferroelectric domains with their inner electric field in the up-direction have lowest energy and the domains with anti-parallel orientation have the highest energy. Thus, there is an energetic tendency for a reorientation of the domains parallel to the applied electric field. However, there is also an intrinsic energetic barrier which has to be overcome by the Ti atom when jumping from one of the six directions to another one.[2] With increasing and decreasing electric field the sizes of different domains change. Due to the barriers which have to be overcome to reach a low energy state, the inner state of the system (roughly the volume of each domain orientation) depends on the history of the system leading to the hysterietic behavior.

Hysteretic behavior in general means that the response of the system (extension of the piezo) does not only depend on the external conditions (applied electric field in our case), but also on the internal state of the system (i.e. its history and here specifically the state of the domain structure). The hysteresis behavior of a piezoelectric ceramic is usually shown in a butterfly curve, where the strain is plotted in dependence of the applied electric field (Fig. 3.11). This figure also shows a schematic sketch of the polarization in the domains. The domains are considered to be square and

[2] In this simplified consideration, we have left out the formation energy of domain walls which results in the formation of larger domains. Larger domains mean less domain wall energy. A further contribution in the energy balance is the build up of mechanical strain inside the domains when an external electric field is applied.

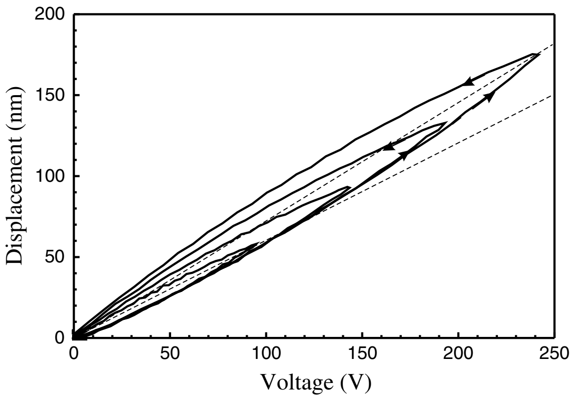


Fig. 3.12 The displacement induced by an applied voltage also shows hysteretic behavior in a range up to 200 V for the applied voltage and the displacement, both in 3-direction. The average piezo constant indicated by the *dashed lines* increases for increasing voltage amplitudes. Due to this the piezo constants and the corresponding displacements can vary by 10–25%. The curves shown here was kindly measured by aixACCT [7] on a PIC 151 ceramic [2]

aligned with respect to the applied field. Also only two of the six possible domain orientations are considered. Point 1 corresponds to saturation polarization where all domains are aligned and also corresponds to maximum strain. If the electric field is subsequently reduced to zero the point of remanent polarization is reached (point 2), where most of the dipoles are still oriented parallel to the outer field. This state corresponds to a certain remanent strain. Between point 1 and point 2 the strain is mainly induced by the intrinsic piezoelectric effect. When the electric field changes orientation the domains also begin to reverse their orientation and the strain is increasingly also induced by domain reorientation. Approaching point 3, the net polarization of the domains is zero. With an increased electric field in the opposite direction the domains begin to align to the opposite direction and correspondingly the strain increases again to its maximum value (point 4). When the electric field is subsequently reversed again, the strain follows a different curve from point 4 to point 5 to point 6 and to point 1. This means that the strain induced by domain reorientation is subject to hysteresis, i.e. depends not only on the external applied electric field but also on the history or the internal state of the system.

The butterfly curve shows the large signal response of piezoelectric ceramics. The working range of piezoelectric materials is between point 1 and point 2 for unipolar operation. For bipolar operation which is used to drive tube piezo elements in scanning probe microscopy, point 3 must not be reached because it corresponds to a depolarization of the piezo. Usually only electric fields substantially below the point of depolarization should be used.

In Fig. 3.12, smaller voltage signals which are used for scanning in AFM are shown together with the corresponding displacement. Also here a hysteresis is visible indicated by the lens-shaped curves which correspond to voltage sweeps form zero to a maximal voltage and back to zero (indicated by the arrows). Such a voltage sweep corresponds to scanning one line in an AFM image. Two effects are observed during these voltage sweeps: first the displacement is different for increasing and decreasing voltages and second this hysteresis increases for larger voltage amplitudes.

Due to this hysteretic behavior the piezo constant (displacement divided by voltage) is not constant anymore. The piezo "constant" depends on the applied voltage and also on the history of the system (which voltages were applied before). If we define the maximum displacement divided by the maximum voltage during one voltage sweep as average piezo constant for this voltage sweep, we see that this average piezo constant increases with the voltage amplitude. This effect results from the increasing contributions due to extrinsic domain reorientation at larger voltages. The average piezo constants are indicated by dashed lines in Fig. 3.12 for the two voltage sweeps with smallest and largest amplitudes. The average piezo constant for the smallest and the largest voltage sweeps in Fig. 3.12 differ by about 18% in this case. This means that due to the effect of hysteresis the piezo constant and correspondingly the piezo displacements vary by 10–25% for different voltages.

This variation (increase) of the piezo constant for larger voltages leads to significant image distortions at larger scan sizes, visible for instance when imaging defined gratings on the scale of several micrometers. The piezoelectric coefficients quoted by the manufacturers of piezo elements are those in the small voltage limit.

3.5.2 Creep

When considering hysteresis (i.e. the domain orientation in dependence of the applied electric field), always a very slow, quasi-static change of the electric field was considered. Since the domain reorientation is an energetically activated process, this process also depends on time. In the case of an instantaneous change of the electric field, the domain reorientation (domain wall motion) and the subsequent build-up of strain (extension of the piezo) do not happen instantaneously but take some time after the electric field has been established. As a result of a sudden jump in the voltage applied to the piezo electrodes the change in position is not instantaneous. A certain time dependence of the position, called creep, is observed. A measurement of creep (displacement as function of time) for short times after an instantaneous voltage jump is shown in Fig. 3.13. For an ideal piezo actuator without creep the displacement would occur only at the time of the voltage jump and not change afterward.

In AFM, the creep results in an effect at the turning points of the scanning movements of each scan line. A positive piezo extension still occurs due to creep, while the voltage change has already reversed its direction. In the vertical direction creep occurs after the (rapid) approach of the tip to the sample. During the approach process, large variations of the z-position are usual and after the approach to the surface a creep in z results.

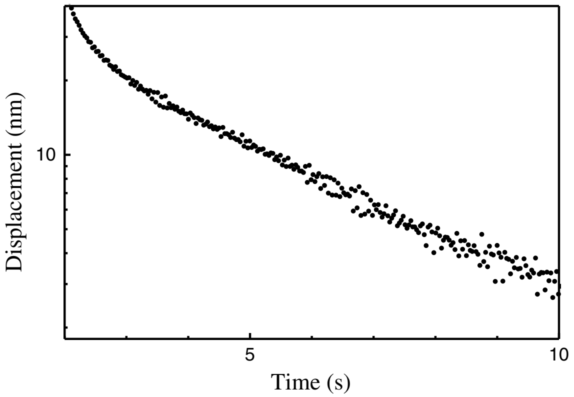


Fig. 3.13 Creep is the piezo displacement after an instantaneous voltage jump. The curve shown here was kindly measured by aixACCT [7] on a PIC 151 ceramic [2]

Creep and hysteresis are also the reason why in scanning probe methods two successive scan lines should not be scanned in opposite directions (first line: $+x$, second line $-x$, ...) but always in the same direction (first line: $+x$, second line $+x$, ...) (no data are acquired while scanning backwards in the $-x$-direction). For lines scanned in opposite directions, a mutual shift in the position of up to 20% would result due to creep and hysteresis.

3.5.3 Thermal Drift

Thermal drift of the mechanical setup leads to image distortions. This is a general effect on all mechanical components of the microscope, and is not limited to piezo elements; specifically, when the sample has been previously annealed (for instance in the process of sample cleaning). Usually it takes some time after approach before the thermal drift is reduced sufficiently for imaging. In low temperature experiments thermal drift is suppressed.

Due to all the above mentioned limitations of piezoelectric materials, piezoelectric actuators operated in open loop (i.e. without an independent measurement of the distance moved) are generally not suitable for an accurate positioning on the nanoscale. Therefore, in atomic force microscopy the piezoelectric actuators are often operated in closed loop. This means that the position of the actuator is measured independently (not relying on a proportionality between the applied voltage and the distance moved). In a feedback loop the actual (measured) position is controlled to the desired setpoint value. More details on the closed loop operation of piezoelectric nanopositioners are discussed in Sect. 4.2.2.

An independent calibration of the piezoelectric actuators in AFM is obtained using commercially available calibration grids for horizontal and vertical calibration [8]. In actual AFM scans of these structures of known height and width the piezoelectric actuators can be calibrated. If atomic resolution is achieved, the lateral calibration can be performed by taking atomically resolved images of a known surface structure. The vertical calibration can be performed at (single) monoatomic step edges.

3.6 Vibration Isolation

In order to keep the scanning probe stable with respect to the sample with an accuracy of less than 0.1 Å would (ambitiously) require a vibrational noise level of about a factor of ten lower than this for the relative motion between tip and sample, i.e. 1 pm. In this case, the usual amplitudes of building vibrations of $\sim 0.1\,\mu\text{m}$ have to be reduced by a factor of 10^5 in order to obtain a sufficiently stable tip-sample distance. As we will see in the following, to accomplish this task both good vibration isolation and a rigid microscope have to be combined.

We will perform the analysis of the vibration isolation in two steps. In the first step, we will consider the microscope as a rigid construction of mass m and ask: How can this mass be isolated from outside vibrations? In the second step, we also consider the microscope itself as a oscillating system where the tip oscillates against the sample and we ask: How can these tip-sample oscillations be reduced?

3.6.1 Isolation of the Microscope from Outer Vibrations

If the microscope is considered as a rigid mass, outside vibrations are transmitted from the ground. An effective vibration isolation can be obtained by a spring suspension (Fig. 3.14a). The microscope assembly (mass m) is fixed to a spring with spring constant k. This harmonic oscillator has a resonance frequency of $\omega_{\text{spring}} = \sqrt{k/m}$. The damping of the oscillating system is described by the quality factor Q_{spring}. An external (sinusoidal) vibration $z_1(t)$ with amplitude z_1^0 and frequency ω (vibration from of the building floor) is coupled into the system (Fig. 3.14a). As a reaction to this outside forced excitation, the mass m performs an oscillation $z_2(t)$ with amplitude z_2^0 at the driving frequency ω. We refer the motions z_1 and z_2 relative to a fixed (not oscillating) reference system. The elastic force on the mass (stretching of the spring) depends on the *difference* of the positions $(z_2 - z_1)$. Thus, the restoring force of the spring acting on the mass m is

$$F_{\text{spring}} = -k(z_2 - z_1), \tag{3.22}$$

The anchoring of the viscous dashpot (Fig. 3.14a) corresponds to the situation shown also in Fig. 2.4b and thus the frictional damping force depends on the *difference* of the velocities $(\dot{z}_2 - \dot{z}_1)$. Therefore, the damping force F_{frict} is

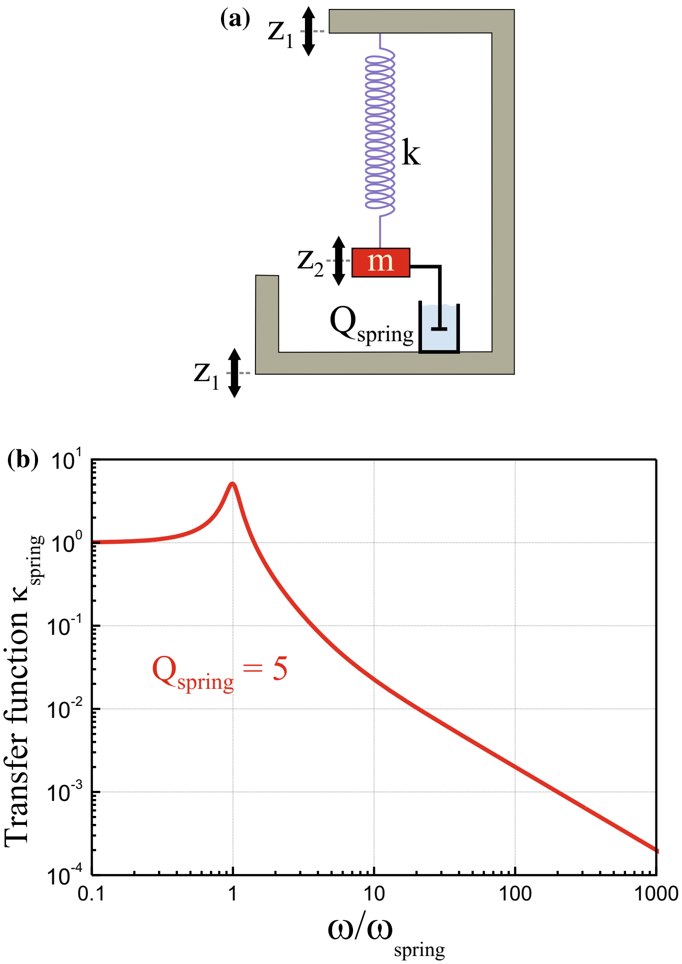


Fig. 3.14 **a** Vibration isolation of a microscope (represented by a mass m) against external vibrations z_1 using a spring suspension. **b** Transfer function of the vibration isolation system for $Q_{\text{spring}} = 5$

$$F_{\text{frict}} = -m\frac{\omega_{\text{spring}}}{Q_{\text{spring}}}(\dot{z}_2 - \dot{z}_1). \tag{3.23}$$

The equation of motion for the mass m reads now

$$m\ddot{z}_2 = -m\frac{\omega_{\text{spring}}}{Q_{\text{spring}}}(\dot{z}_2 - \dot{z}_1) - k(z_2 - z_1), \tag{3.24}$$

or using $\omega^2_{\text{spring}} = k/m$, and reordered slightly results in

$$\ddot{z}_2 + \frac{\omega_{\text{spring}}}{Q_{\text{spring}}}\dot{z}_2 + \omega^2_{\text{spring}} z_2 = \frac{\omega_{\text{spring}}}{Q_{\text{spring}}}\dot{z}_1 + \omega^2_{\text{spring}} z_1. \tag{3.25}$$

For a sinusoidal vibration of the frame z_1 can be written in the complex notation (skipping the tilde used for complex numbers in Sect. 2.4)

$$z_1(t) = z_1^0 e^{i\omega t}, \tag{3.26}$$

the steady-state solution for the motion of the mass m is

$$z_2(t) = z_2^0 e^{i\omega t}. \tag{3.27}$$

with z_1^0 and z_2^0 being complex amplitudes which include a relative phase shift between the two amplitudes.

Substituting (3.26) and (3.27) into (3.25) we obtain (again using the power of the complex method: differentiation is just multiplication by $i\omega$)

$$-\omega^2 z_2 + i\frac{\omega_{\text{spring}}}{Q_{\text{spring}}}\omega z_2 + \omega^2_{\text{spring}} z_2 = i\frac{\omega_{\text{spring}}}{Q_{\text{spring}}}\omega z_1 + \omega^2_{\text{spring}} z_1. \tag{3.28}$$

or

$$\left(-\omega^2 + i\frac{\omega_{\text{spring}}}{Q_{\text{spring}}}\omega + \omega^2_{\text{spring}}\right) z_2^0 e^{i\omega t} = \left(i\frac{\omega_{\text{spring}}}{Q_{\text{spring}}}\omega + \omega^2_{\text{spring}}\right) z_1^0 e^{i\omega t}. \tag{3.29}$$

Finally, we obtain

$$\frac{z_2^0}{z_1^0} = \frac{\omega^2_{\text{spring}} + i\frac{\omega_{\text{spring}}}{Q_{\text{spring}}}\omega}{\omega^2_{\text{spring}} - \omega^2 + i\frac{\omega_{\text{spring}}}{Q_{\text{spring}}}\omega}. \tag{3.30}$$

This ratio is still a complex number, since both amplitudes are complex quantities having a real amplitude and phase. The ratio of the absolute values of the amplitudes is called the transfer function of the vibration isolation system $\kappa_{\text{spring}}(\omega)$, which can be written as

$$\kappa_{\text{spring}}(\omega) = \frac{|z_2^0|}{|z_1^0|} = \sqrt{\frac{\omega^4_{\text{spring}} + \frac{\omega^2_{\text{spring}}}{Q^2_{\text{spring}}}\omega^2}{(\omega^2_{\text{spring}} - \omega^2)^2 + \frac{\omega^2_{\text{spring}}}{Q^2_{\text{spring}}}\omega^2}}. \tag{3.31}$$

When expressed in terms of the normalized frequency $\omega/\omega_{\text{spring}}$, the following expression for the transfer function results

$$\kappa_{\text{spring}}(\omega) = \frac{|z_2^0|}{|z_1^0|} = \sqrt{\frac{1 + \frac{1}{Q_{\text{spring}}^2}\left(\frac{\omega}{\omega_{\text{spring}}}\right)^2}{\left[1 - \left(\frac{\omega}{\omega_{\text{spring}}}\right)^2\right]^2 + \frac{1}{Q_{\text{spring}}^2}\left(\frac{\omega}{\omega_{\text{spring}}}\right)^2}}. \tag{3.32}$$

The response of the system to a driven oscillation $\kappa_{\text{spring}}(\omega)$ can be divided into three regimes (Fig. 3.14b). For $\omega \ll \omega_{\text{spring}}$ the outside excitation is transmitted with a transfer function of one, i.e. without any damping. For a frequency close to the resonance frequency of the system (in resonance), the outside excitation is even amplified, i.e. the vibrations are increased instead of damped. At $\omega = \omega_{\text{spring}}$ the transfer function becomes

$$\kappa_{\text{spring}}(\omega_{\text{spring}}) = \sqrt{1 + Q_{\text{spring}}^2}. \tag{3.33}$$

In the regime $\omega \gg \omega_{\text{spring}}$ and Q_{spring} very large, the transfer function (3.31) reduces to

$$\kappa_{\text{spring}}(\omega) \approx \left(\frac{\omega_{\text{spring}}}{\omega}\right)^2. \tag{3.34}$$

This shows that for excitation frequencies ω much larger than the resonance frequency ω_{spring} and for small damping, the external vibrations are suppressed $\sim 1/\omega^2$. We have seen that damping (small Q_{spring}-factor) avoids resonance excitation. However, on the other hand damping deteriorates vibration isolation at higher frequencies. The transfer function becomes $\sim 1/\omega$ for $Q_{\text{spring}} = 1$. In Fig. 3.15 the transfer function is

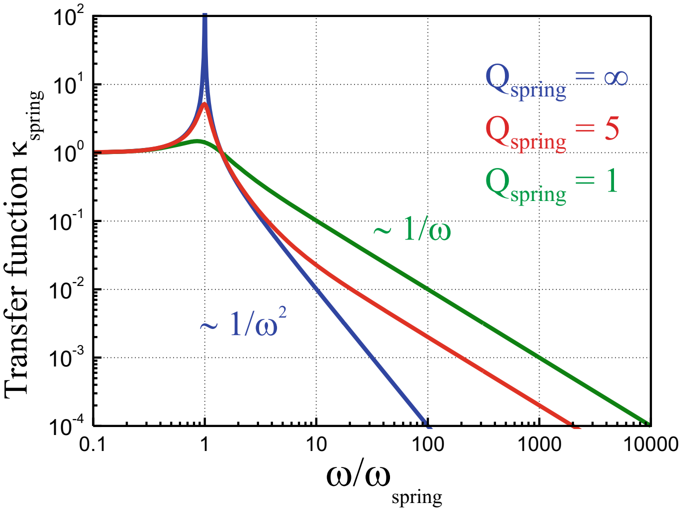


Fig. 3.15 Transfer function of a spring suspension system for different values of the quality factor Q_{spring}

shown for different values of Q_{spring}. In typical spring suspension systems, a compromise between good damping at high frequencies and large resonance enhancement is chosen for $Q_{\text{spring}} \approx 2 - 5$.

The vibration isolation (for instance from building vibrations) will be better the lower the resonance frequency of the spring system is. Therefore, the resonance frequency of the spring system is the prime parameter of a vibration isolation system. In the following, we will show that this parameter only depends on the extension length of the spring Δl.

Hooke's law results in $k\Delta l = mg$, with g being the gravitational acceleration. If we insert the result for m into the equation for the resonance frequency of the system $f_{\text{spring}} = \frac{1}{2\pi}\sqrt{k/m}$ the resonance frequency for the system can be written as

$$f_{\text{spring}} = \frac{1}{2\pi}\sqrt{\frac{k}{k\Delta l/g}} = \frac{1}{2\pi}\sqrt{\frac{g}{\Delta l}}. \tag{3.35}$$

To achieve a resonance frequency of 1 Hz the spring should be stretched by 25 cm. To achieve a resonance frequency of 0.5 Hz the spring has to be stretched by 1 m. This length is difficult to integrate in a system. Some reduction of the length of the springs can be achieved by using pretensioned springs. Such springs are available in principle, but, it is difficult to manufacture springs which simultaneously feature a high pretension force and a low spring constant.

Note that the mass and the spring constant do not enter explicitly into the expression for the resonance frequency. This equation is the same as for a simple pendulum with length Δl. Therefore, a spring suspension system acts as a isolation device for both vertical and horizontal environmental vibrations.

3.6.2 *The Microscope Considered as a Vibrating System*

In the second step of our analysis of the vibration isolation, we consider the microscope itself as a vibrating system. While it is wise to couple the sample most rigidly to the scanner/tip assembly, this (stiff) mechanical loop of the AFM (or generally the scanning probe microscope (SPM)) can also be characterized as a vibrating system with a (quite high) resonance frequency ω_{SPM} and a quality factor Q_{SPM} (Fig. 3.16a). The softest part in the mechanical loop is the piezo material with a typical quality factor of 100. Let z_2 describe the oscillation of the microscope support base (or sample in Fig. 3.16a), and z_3 the vibration of the scanner/tip assembly. Here one point is important (which makes life much easier): it is not the vibration amplitude of the tip z_3 that has to be reduced to a minimum but only the *difference* of the motion between tip and sample $z_3 - z_2$. Only the relative motion between tip and sample counts! The differential equation for the vibrating tip z_3 relative to an external fixed reference is

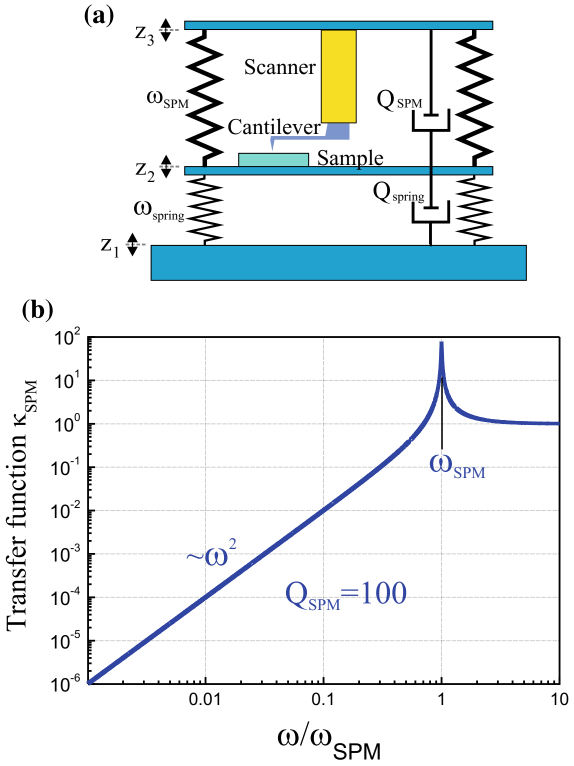


Fig. 3.16 **a** The scanning probe microscope (SPM) itself is considered as an oscillating system characterized by ω_{SPM} and Q_{SPM}. Tip and sample oscillate against each other. **b** Transfer function κ_{SPM} according to (3.40) for the microscope with resonance frequency ω_{SPM}

$$\ddot{z}_3 + \frac{\omega_{\mathrm{SPM}}}{Q_{\mathrm{SPM}}}(\dot{z}_3 - \dot{z}_2) + \omega_{\mathrm{SPM}}^2(z_3 - z_2) = 0. \tag{3.36}$$

The spring force is proportional to $z_3 - z_2$ and the frictional force is proportional to $\dot{z}_3 - \dot{z}_2$. Using the complex method to solve the differential equation results in

$$-\omega^2 z_3 + i\frac{\omega_{\mathrm{SPM}}}{Q_{\mathrm{SPM}}}\omega(z_3 - z_2) + \omega_{\mathrm{SPM}}^2(z_3 - z_2) = 0, \tag{3.37}$$

or

$$-\omega^2 z_2 - \omega^2(z_3 - z_2) + i\frac{\omega_{\mathrm{SPM}}}{Q_{\mathrm{SPM}}}\omega(z_3 - z_2) + \omega_{\mathrm{SPM}}^2(z_3 - z_2) = 0. \tag{3.38}$$

The (complex) ratio of the difference of the amplitudes $z_3^0 - z_2^0$ to the amplitude of the base of the microscope z_2^0 is obtained as

$$\frac{z_3^0 - z_2^0}{z_2^0} = \frac{\omega^2}{\omega_{\mathrm{SPM}}^2 - \omega^2 + i\frac{\omega_{\mathrm{SPM}}}{Q_{\mathrm{SPM}}}\omega}. \tag{3.39}$$

The transfer function results in

$$\kappa_{\mathrm{SPM}}(\omega) = \left| \frac{z_3^0 - z_2^0}{z_2^0} \right| = \sqrt{\frac{\omega^4}{(\omega_{\mathrm{SPM}}^2 - \omega^2)^2 + \frac{\omega_{\mathrm{SPM}}^2}{Q_{\mathrm{SPM}}^2}\omega^2}}. \tag{3.40}$$

The resulting transfer function is plotted in Fig. 3.16b and can be approximated by

$$\kappa_{\mathrm{SPM}}(\omega) \approx \left(\frac{\omega}{\omega_{\mathrm{SPM}}} \right)^2, \tag{3.41}$$

for $\omega \ll \omega_{\mathrm{SPM}}$, and small damping, with ω_{SPM} being the resonance frequency of the SPM (mechanical loop between tip and sample). When the excitation frequency ω is much lower than the resonance frequency of the microscope ω_{SPM}, good damping of the external vibrations is achieved. In Fig. 3.16b we use $Q_{\mathrm{SPM}} = 100$, since the material with the lowest Q-factor in the mechanical loop is the piezo ceramic, which has a typical mechanical quality factor of about 100.

3.6.3 Combining Vibration Isolation and a Microscope with High Resonance Frequency

The concept for an effective vibration isolation is to combine the two approaches and use a low resonance frequency for the vibration isolation system and a high resonance frequency for the mechanical loop of the microscope. According to (3.31), a vibration of the frame with amplitude $|z_1^0|$ is transmitted to the SPM base with amplitude $|z_2^0|$ as

$$z_2^0 = \kappa_{\mathrm{spring}} z_1^0. \tag{3.42}$$

(From now on, we consider the amplitudes as real and omit the absolute signs.) Furthermore the vibration amplitude of the SPM base z_2^0 induces (according to (3.40)) a relative amplitude between tip and sample of

$$z_3^0 - z_2^0 = \kappa_{\mathrm{SPM}} z_2^0. \tag{3.43}$$

In total, an outer vibration of amplitude z_1^0 induces a relative tip sample vibration of amplitude $z_3^0 - z_2^0$ as

$$z_3^0 - z_2^0 = \kappa_{\mathrm{SPM}} z_2^0 = \kappa_{\mathrm{SPM}} \kappa_{\mathrm{spring}} z_1^0. \tag{3.44}$$

Thus, the total transfer function can be written as

$$\kappa_{\mathrm{total}} = \frac{z_3^0 - z_2^0}{z_1^0} = \kappa_{\mathrm{SPM}} \kappa_{\mathrm{spring}}. \tag{3.45}$$

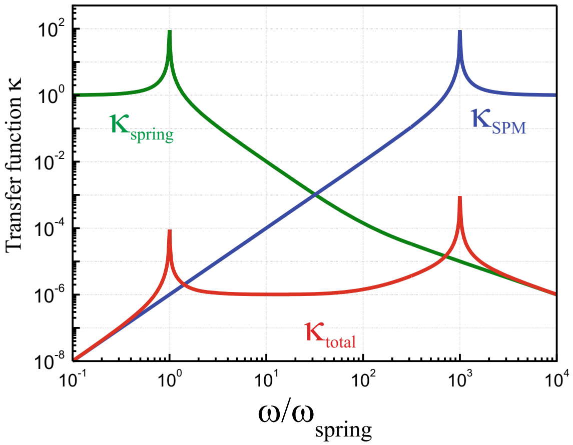


Fig. 3.17 Transfer function of the combined system κ_{total} given by the product of the individual transfer functions of the spring suspension system κ_{spring} and the SPM itself κ_{SPM} for the case of small damping, i.e. $Q_{\text{SPM}} = Q_{\text{spring}} = 100$

The transfer function of the combined system is the product of the transfer functions of the individual systems.

According to (3.34) and (3.41), the total transfer function can be approximated in the frequency range $\omega_{\text{spring}} < \omega < \omega_{\text{SPM}}$ as

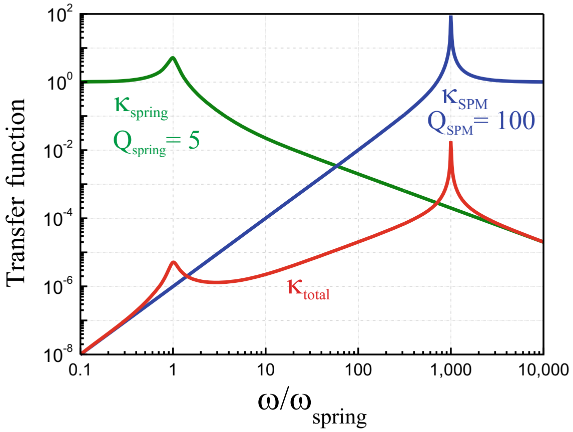
$$\kappa_{\text{total}} \approx \left(\frac{\omega_{\text{spring}}}{\omega}\right)^2 \left(\frac{\omega}{\omega_{\text{SPM}}}\right)^2 = \left(\frac{\omega_{\text{spring}}}{\omega_{\text{SPM}}}\right)^2 . \qquad (3.46)$$

This behavior of an approximately constant transfer function in between the resonance frequencies ω_{spring} and ω_{SPM} can be seen in Fig. 3.17 in which the transfer function is shown in the limit of negligible damping ($Q_{\text{SPM}} = Q_{\text{spring}} = 100$).

If, for example, the resonance frequency of the spring suspension system is 1 Hz and the resonance frequency of the SPM is 1 kHz, the overall transfer function for intermediate frequencies has a constant value of 10^{-6}, as shown in Fig. 3.17. If we would be able to raise the resonance frequency of the SPM to 10 kHz the total transfer function for the transmission of an external vibration to the tip-sample distance would go to 10^{-8}!

Next we consider more realistic transfer functions by including damping. For the spring suspension system we consider $Q_{\text{spring}} = 5$, while we assume $Q_{\text{SPM}} = 100$. When damping is included the total transfer function is not constant anymore in the range between ω_{spring} and ω_{SPM}. The total transfer function according to (3.31) and (3.40) is plotted in Fig. 3.18 together with the individual transfer functions of the spring suspension and the SPM. It is assumed that the SPM mechanical loop can be approximated by a single resonance frequency 1,000 times higher than the resonance frequency of the spring suspension. With this assumption, the transfer function stays below the initial desired value of 10^{-5} up to $\omega/\omega_{\text{spring}} < 40$. The quite high values of the transfer function for higher frequencies (which arises due to the relatively strong damping of the spring suspension) could be regarded as problematic. However, as

Fig. 3.18 Transfer function of the combined system κ_{total} which is the product of the individual transfer functions of the spring suspension system κ_{spring} and the SPM itself κ_{SPM}



we will see in the next section, the driving amplitude of the exciting floor vibrations decreases at larger frequencies.

In summary, the spring suspension acts as a low-pass for vibrations with frequencies smaller than the resonance frequencies of the spring ω_{spring}, while it damps the vibrations at larger frequencies. On the other hand, the SPM assembly acts as a high-pass for vibrations with a frequency larger than ω_{SPM}, while it damps the vibrations at lower frequencies. The total transfer function is the product of the transfer

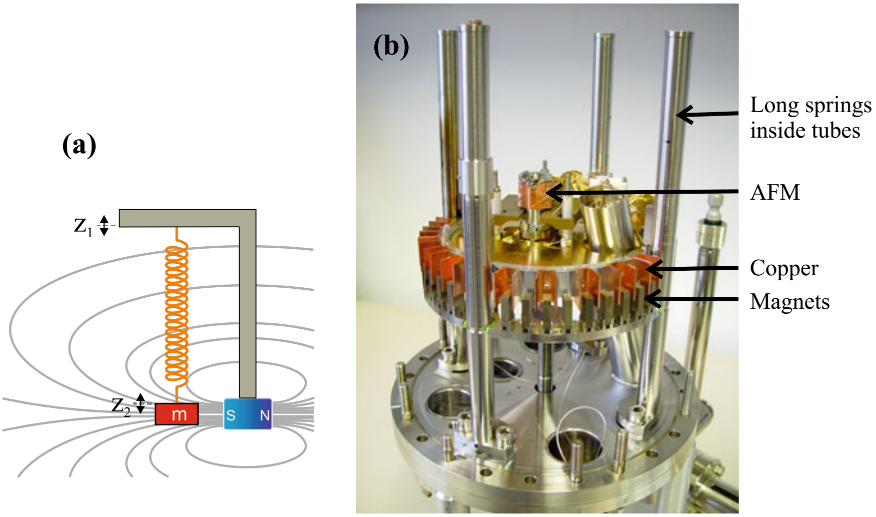


Fig. 3.19 **a** Principle of an eddy-current damping system with a magnet next to a conductor in which the energy of the vibrations is dissipated as eddy currents. **b** Photo of an eddy-current damping system with isolating an SPM from outer vibrations

functions of the spring suspension and SPM. In order to keep the total transfer function low at all frequencies, a low resonance frequency of the vibration isolation, as well as a high frequency of the microscope mechanical loop are required.

The necessary damping of a spring suspension system is often performed by eddy-current damping. When a conductor (usually copper) moves in a magnetic field, damping forces are generated by eddy currents inside the conductor, as shown in the schematic in Fig. 3.19a. An example of an eddy-current damping system is shown in Fig. 3.19b. The disadvantage of a spring suspension system is the large size. Another way of damping is to use a stack of metal plates separated by rubber (e.g. Viton®) pieces, which act as springs and dampers simultaneously. A further method of vibration isolation is to mount the SPM on pneumatic isolation legs (also used for optical tables). A typical resonance frequency of such a table is 1–2 Hz, and a transfer function of smaller than 0.01 can be achieved for frequencies larger than 10 Hz. A suppression of acoustic noise can be achieved by putting the AFM into an acoustic enclosure or acoustic hood.

3.7 Building Vibrations

When we were mentioning that the amplitudes of building vibrations are of the order of ~0.1 μm this value is an integral value obtained by integration over time or frequency. In the previous section we have considered the frequency dependence of the transfer of vibrations, however also the primary floor vibrations have a frequency dependence. This frequency dependence of vibrational noise, can be expressed by the noise spectral density (of the velocity v) $N_v(f)$, which is introduced in Sect. 5.5 and discussed in more detail in Appendix B.

Geophones (accelerometers) are typically used to measure the spectral density of the velocity noise $N_v(f)$ of the building vibrations. In Fig. 3.20, the spectral density

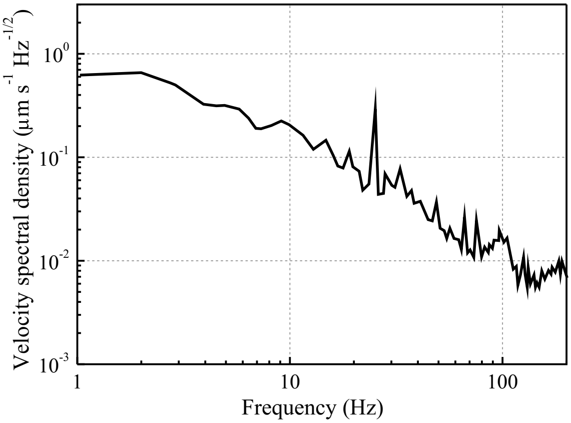


Fig. 3.20 Velocity spectral noise density N_v (RMS) of the building vibrations measured on the floor in a building at the Research Center in Jülich

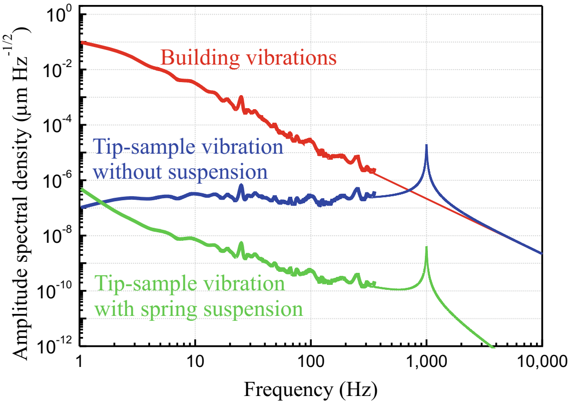


Fig. 3.21 Calculated tip-sample ($z_3^0 - z_2^0$) vibrational amplitude spectral density as a function of the frequency f, obtained using the measured building vibrations and the appropriate transfer function from Fig. 3.18. The amplitude spectral density of the building vibrations is shown as a *red line*. The data are taken from Fig. 3.20 and extrapolated for higher frequencies. The *green* and *blue curves* show the behavior with and without a spring suspension system, respectively

of the velocity noise of the building vibrations measured on a floor in a building in Research Center Jülich is plotted as function of vibration frequency. The general behavior is that the amplitude deceases with increasing frequency. The highest values are observed for low frequencies around 1–2 Hz with a value of $N_v \approx 0.7\,\mu\text{m}/(\text{s}\sqrt{\text{Hz}}\,)$

Building vibrations can be influenced by external conditions like nearby railway lines or motorways. Also inside a building the building vibrations are increased by compressors, large machines, and ventilation systems. As a general rule the intrinsic building vibrations are more pronounced in higher floors and correspondingly lowest in the basement of a building. For this reason, sensitive scanning probe microscopes can be often found in the basement.

In order to convert the measured data from the velocity to oscillation amplitude or acceleration, we recall that

$$z = z_0 \cos(\omega t), \tag{3.47}$$

$$v = \dot{z} = -z_0\omega \sin(\omega t) := -v_0 \sin(\omega t), \tag{3.48}$$

$$a = \ddot{z} = -z_0\omega^2 \cos(\omega t). \tag{3.49}$$

The same relations between amplitude, velocity and acceleration apply also for the corresponding noise densities.[3]

Now we include the measured building vibrations in the vibration analysis. The amplitude noise density $N_z(f)$ of the building vibrations $z_1^0(f)$ can be calculated

[3] Note that accelerometers often measure the root mean square (RMS) amplitude which is smaller than the peak amplitude by a factor of $1/\sqrt{2}$.

from the measured velocity noise density $N_v(f)$ (shown in Fig. 3.20) using (3.48). According to (3.44), the relevant tip-sample vibrational amplitude $z_3^0 - z_2^0$ can be expressed as a function of frequency as

$$z_3^0 - z_2^0 = \kappa_{\text{total}}(f) z_1^0(f). \tag{3.50}$$

If we multiply the total transfer function by the measured spectral density of the floor vibration amplitude (derived from Fig. 3.20), the expected tip-sample spectral density of the vibration amplitude arising due to the floor vibrations is shown in Fig. 3.21. The case where no spring suspension is invoked is shown as blue line, leading to a roughly constant tip-sample vibration amplitude density of $N_z(f) = 10^{-7}\,\text{nm}/\sqrt{\text{Hz}}$. This leads according to (5.14) to an RMS value of the tip-sample amplitude of $z_3^0 - z_2^0 = N_z\sqrt{1\,\text{kHz}} = 3.16\,\text{pm}$ for a bandwidth of 1 kHz.

The tip-sample vibrational amplitude is decreased further by invoking a spring suspension, as shown by the green curve in Fig. 3.21. The building vibrations are damped for frequencies larger than the resonance frequency, specifically also at the resonance frequency of the SPM ω_{SPM}. However, a spring suspension also leads to a resonance at the natural frequency of the spring suspension system ω_{spring}, which has to be suppressed by proper damping of the spring suspension system. The decrease of the tip-sample vibration amplitude above the resonance frequency of the spring suspension system can result in RMS values of the vibration amplitude substantially below one picometer for a bandwidth of 1 kHz.

3.8 Summary

- Due to the piezoelectric effect a voltage applied to the electrodes of a piezoelectric element leads to a strain, i.e. a motion of some part of the element.
- The piezo constant describes the sensitivity of a piezoelectric actuator in Å/V.
- The most frequently used piezoelectric actuator element in scanning probe microscopy is the tube piezo element. It allows x, y, and z-motion with one single element.
- Problems with piezoelectric actuators are the coupling of lateral and vertical motion, non-linearity, hysteresis, and creep.
- Sharp SPM tips can be fabricated by self-adjusting electrochemical etching.
- The resonance frequency of a spring suspension system depends only on the extension length Δl as $\omega_{\text{spring}} = \sqrt{\frac{g}{\Delta l}}$.
- It is not necessary to minimize the amplitude of the tip vibration and the sample vibration individually but only the *difference* between tip and sample position.
- For effective vibration isolation a low resonance frequency of the spring suspension system ω_{spring} is combined with a high resonance frequency of the SPM assembly ω_{SPM}, i.e. a stiff mechanical loop between tip and sample.

- The transfer function (i.e. the attenuation of external vibrations) is constant for small damping $\kappa_{\text{total}} \approx \left(\frac{\omega_{\text{spring}}}{\omega_{\text{SPM}}}\right)^2$ for $\omega_{\text{spring}} < \omega < \omega_{\text{SPM}}$.
- The expected tip-sample vibration amplitude can be calculated by multiplying the total transfer function by the (measured) building vibration amplitude.

References

1. W. Heywang, K. Lubitz, W. Wersing (eds.), *Piezoelectricity*, 1st edn. (Springer, Heidelberg, 2008). https://doi.org/10.1007/978-3-540-68683-5
2. PI Ceramic GmbH, Lindenstrasse, 07589 Lederhose, Germany. http://www.piceramic.com
3. K.G. Vandervoort, R.K. Zasadzinski, G.G. Galicia, G.W. Crabtree, Full temperature calibration from 4 to 300 K of the voltage response of piezoelectric tube scanner PZT-5A for use in scanning tunneling microscopes. Rev. Sci. Instrum. **64**, 896 (1994). https://doi.org/10.1063/1.1144139
4. C. Wei, A circular arc bending model of piezoelectric tube scanners. Rev. Sci. Instrum. **67**, 2286 (1998). https://doi.org/10.1063/1.1146934
5. M. Hannss, W. Naumann, R. Anton, Performance of a tiltcompensating tube scanner in atomic force microscopy. Scanning **20**, 501 (1998). https://doi.org/10.1002/sca.1998.4950200703
6. W.D. Pilkey, *Formulas for Stress, Strain and Structural Matrices*, 2nd edn. (Wiley, New York, 2005). https://doi.org/10.1002/9780470172681
7. aixACCT Systems GmbH, Talbotstr. 25, 52068 Aachen, Germany. www.aixacct.com
8. P. Eaton, P. West, *Atomic Force Microscopy* (Oxford Universiy Press, New York, 2010)

Chapter 4
Atomic Force Microscopy Designs

The design of an AFM has to enable two different tasks: First it has to allow for a xyz-motion during scanning (fine motion, or scan motion), for the acquisition of the surface topography. As the range of the piezo actuators performing this motion is limited to usually $<100\,\mu\mathrm{m}$, the second task of an AFM design is to bring the cantilever tip and the sample initially so close together, that their distance is within the range of the z-fine motion. This task is called the coarse approach. Both of these tasks have to be satisfied while simultaneously maintaining a stiff mechanical structure with high resonance frequencies allowing for good vibration isolation and small (thermal) drift of the tip relative to the sample. In this chapter we discuss several types of coarse positioners as well as scanners for the fine motion and introduce the principles of some particular AFM designs.

4.1 Coarse Positioners

A standard way to achieve coarse positioning between the cantilever tip and the sample in AFM is the use of a fine thread screw which is driven by a stepper motor. Since both components of such a coarse positioner are widespread industrial products, we do not discuss this type of coarse positioning here in more depth. Instead we discuss here the working principle of inertial sliders, which are used for coarse positioning purposes in atomic force microscopy as well.

4.1.1 Inertial Sliders

How an inertial slider works in principle can be easily grasped by the following experiment: Place a sheet of paper on a table and place a coin on the paper. Now you can move the coin without touching it by shaking the paper on the table with your hand in a saw-tooth pattern, i.e. quick in one direction and slow in the opposite

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3_4

direction. The coin will stay in frictional contact with the paper during the slow movement (small slope part of saw-tooth motion) and move together with the paper. However, during the steep slope part of the saw-tooth motion the frictional contact between the coin and the paper will disengage due to the inertia of the coin and the coin will not move (or move only slightly) relative to the table. This simple principle is the basis for many nanopositioners.

All these inertial sliders consist of two essential parts: a mover which is moved by a piezo actuator relative to a reference frame and an object to be moved called slider in the following [1]. This very general configuration of an inertial slider is shown in Fig. 4.1a. The term inertial slider is used because inertia is important for the function of these devices. Inertia is the "resistance" of a mass to change its state of motion. Newton's first law, which is also called the law of inertia, states that if no force acts on a mass this mass will not change its velocity due to its inertia. In the following we describe the motions from an external fixed inertial frame (support). We also assume that the friction forces do not depend on the velocity but that they are proportional to the normal force which the slider exerts on the mover.

The force accelerating the slider mass is transmitted from the mover via the frictional surface to the slider. The slider stays in frictional engagement with the mover if the static friction force $F_{\mathrm{frict}}^{\mathrm{stat}}$ is larger than the force on the slider due to its acceleration as

$$m\, a_{\mathrm{mover}} = m\, a_{\mathrm{slider}} < F_{\mathrm{frict}}^{\mathrm{stat}} = m\, a_{\mathrm{frict}}^{\mathrm{stat}} = \mu_{\mathrm{stat}}\, m\, g, \tag{4.1}$$

with μ_{stat} being the coefficient of static friction of the frictional surface, m the mass of the slider, and g the gravitational acceleration. Since μ_{stat} is of the order of one, the acceleration of the mover must be roughly smaller than g in order to remain in frictional engagement. In this phase of motion, called "riding phase", the slider moves together with the mover.

The frictional surface remains in static frictional contact if forces smaller than the threshold force $F_{\mathrm{frict}}^{\mathrm{stat}}$ are applied. If however, $m\, a_{\mathrm{mover}} > F_{\mathrm{frict}}^{\mathrm{stat}}$ the frictional contact disengages, transforms to a sliding frictional contact and the slider will not move together with the mover (Fig. 4.1b). The necessary accelerations larger than g can be easily reached by piezoelectric actuators with their resonance frequencies in the kHz range. If the frictional engagement at the friction surface is lost, only the smaller kinetic frictional coefficient μ_{kin} acts at the frictional surface and the force acting on the slider reduces to

$$m\, a_{\mathrm{slider}} = F_{\mathrm{frict}}^{\mathrm{kin}} = \mu_{\mathrm{kin}}\, m\, g. \tag{4.2}$$

The direction of this force due to the kinetic friction (positive/negative) corresponds to the sign of the relative velocity $v_{\mathrm{mover}} - v_{\mathrm{slider}}$.

In Fig. 4.2 the position, the velocity and the acceleration of the mover and slider relative to an external reference are shown during the "riding phase" and the "sliding phase". The saw-tooth signal of the mover is approximated by a small slope and a large slope segment. The sharp corners (which are rounded in reality) give rise to an

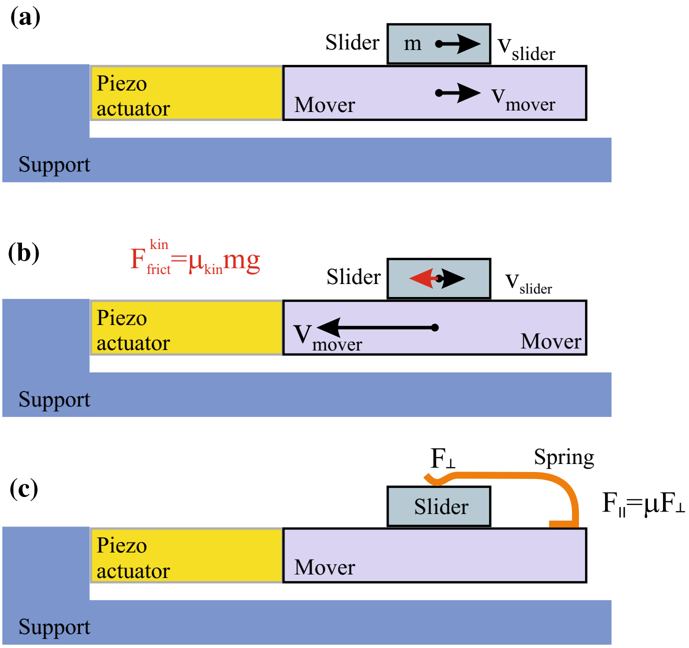


Fig. 4.1 Operating principle of an inertial slider. **a** Riding phase: $ma_{mover} \leq F^{stat}_{frict} = \mu_{stat}$ mg. **b** Sliding phase: $ma_{mover} > F^{stat}_{frict}$. **c** Inertial slider with spring

acceleration at these points. Due to the small slope of the position in the riding phase, the peak in the acceleration at time zero is smaller than the threshold acceleration a^{stat}_{frict}, and the slider stays in frictional engagement with the mover. During the riding phase, mover and slider are in static frictional contact and move with the same constant velocity. The acceleration is zero and the position changes linearly for both the mover and the slider. When the saw-tooth signal changes from the small slope ("riding phase") to the steep slope ("sliding phase"), the mover accelerates for a short time (negative spike in the acceleration, Fig. 4.2c) and the static frictional contact is lost. After this transient state, the mover acceleration is zero again and the mover now has a high (constant) velocity, the mover position changes linearly with a large slope. During the acceleration peak of the mover, which is (much) larger than the threshold acceleration a^{stat}_{frict}, the slider loses static frictional contact. Now a negative force due to the kinetic friction acts on the slider according to (4.2). This leads to a linearly decreasing velocity of the slider. During this deceleration due to the kinetic friction the position of the slider develops as shown in Fig. 4.2a.

When the velocity of the mover stops (at time 1 in Fig. 4.2) there is another sharp (this time positive) spike in the acceleration of the mover. The slider continues to stay in kinetic friction and decelerates from the velocity which it acquired during the riding phase until the slider stops. Now the slider engages with the mover again, i.e. the frictional surface transforms to static friction. After the completion of a sequence, the

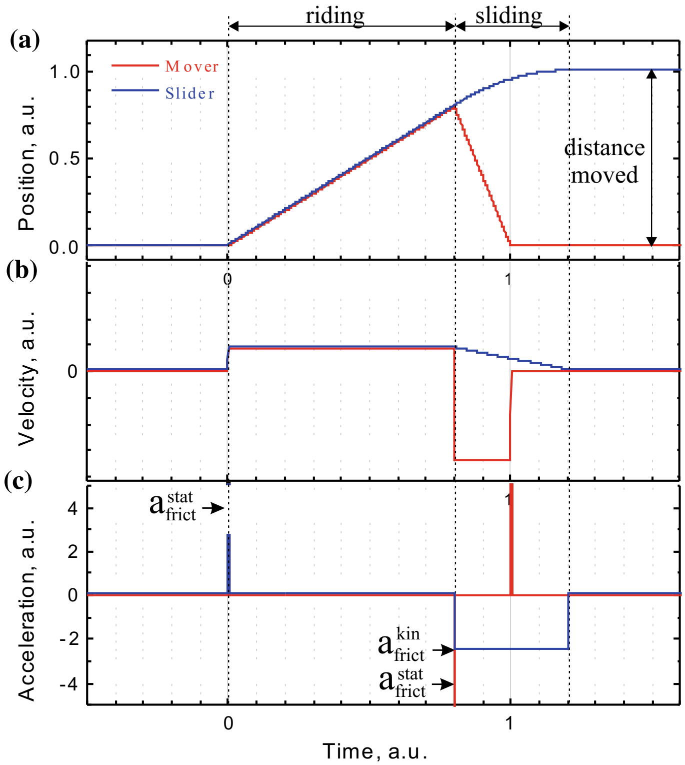


Fig. 4.2 Position, velocity and acceleration of the mover and the slider during inertial motion as a function of time relative to an external fixed support

slider has moved relative to the mover by a certain distance as indicated in Fig. 4.2a. In reality, the sliding phase occurs in a much shorter time relative to the riding phase than shown in Fig. 4.2. Also the transitions between the different regions are not sharp but rounded and the acceleration during the steep slope segment of the saw-tooth signal does not vanish.

Here we note two points resulting from the detailed analysis. First, the motion of the slider is not zero during the sliding phase, but it decelerates from the velocity during the riding phase to rest. This deceleration is induced by the kinetic friction force which acts during the sliding phase. The second point is that during the sliding phase no acceleration of the mover is required (apart from the initial transient, which leads to a transition from static friction to kinetic friction). Also with zero acceleration during the sliding phase the slider moves relative to the mover.

In most inertial sliders, the force normal to the frictional surface $F_\perp$ is not supplied by the gravitation (as assumed up to now), but by other means like magnets or springs as shown in Fig. 4.1c. This has the advantage that the inertial slider can work in any orientation if $F_\perp \gg m\,g$. In this case, the maximal static frictional force

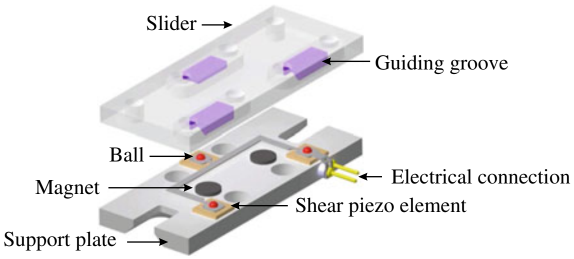
$F_{\text{frict}}^{\text{stat}} = \mu_{\text{stat}} F_{\perp}$ is independent of the mass of the object to be moved and frictional engagement is lost if

$$m\, a_{\text{mover}} > F_{\text{frict}}^{\text{stat}} = \mu_{\text{stat}}\, F_{\perp}. \tag{4.3}$$

In order to lose frictional contact (to go into sliding phase) $m\, a_{\text{mover}}$ has to be larger than the static friction force $F_{\text{frict}}^{\text{stat}}$. This means that either the mass m of the slider or the acceleration of the mover a_{mover} has to be large in order to fulfill the relation $m\, a_{\text{mover}} > \mu_{\text{stat}} F_{\perp}$. There are certain limits to the acceleration of the mover. The first fundamental limit is that the mover cannot be moved at frequencies higher than the resonance frequency of the mover (or rather the combined system of piezo actuator and mover). Another effect which limits the acceleration of the mover is the speed at which the power supply of the piezo actuator can pump charge to the piezo element. A piezoelectric actuator can be considered from the electrical viewpoint as a capacitor which is charged quickly during the steep slope segment of the saw-tooth signal. The slew rate is the maximal voltage change per time provided by the power supply for a certain piezo capacity. Assuming now a certain maximum limit for the acceleration of the mover (given by the resonance frequency or the slew rate of the power supply), the mass of the slider m is the free parameter which can be tuned (increased) in order to raise the force $m\, a_{\text{mover}}$ above the limit for sliding $\mu_{\text{stat}}\, F_{\perp}$. This means a certain (minimum) mass of the slider is needed for operation of the inertial slider. In practical applications for nanopositioning systems a high mass of the slider has several disadvantages. Ideally, the size of inertial sliders used for nanotechnology should be as small as possible. However, a certain mass (corresponding also to a certain size of the slider) is needed for operation of the inertial motion, as stated above. Another reason for a small mass of the slider is that a large mass also intrinsically leads to undesired low eigenfrequencies ($\omega_0 = \sqrt{k/m}$). Therefore, the high mass required for the operation of the inertial motion contradicts the requirement of a small mass for small devices with high eigenfrequencies and an appropriate compromise between these opposing demands has to be found. Later we will also introduce nanopositioners which do not rely on inertia.

A practical implementation of an inertial slider as nanopositioner is shown in Fig. 4.3 [2]. On a baseplate three shear piezo elements are mounted which provide motion up to about one micrometer in one direction. On top of the shear piezo elements, hemispherical balls are mounted, usually made of hard materials like ruby, sapphire, or stainless steel. These three balls correspond to the mover in the previous discussion. The slider is held by magnetic force on top of the three balls. Small magnets in the middle of the baseplate exert a force onto the magnetic slider, which rests firmly on the three balls. The motion of the slider is guided along one direction by a groove in the slider in which two of the three balls are resting. A saw-tooth pattern of motion is applied to the piezo elements and leads to a motion of the slider along one direction, as described above. The step size of an inertial slider can be chosen down to the nanometer range, but also larger step sizes (micrometer) are possible and allow quick positioning even in the millimeter range.

Fig. 4.3 Sketch of an inertial slider (length 35 mm)



4.2 AFM Scanners

After the cantilever tip and the sample have been brought together, using the coarse positioner, within the range of the fine positioners, the other task which has to be preformed by an AFM is the fine scanning motion in all three spatial directions. In several AFM designs the tube scanner is used for the fine scanning motion. We have discussed the tube scanner in detail in Sect. 3.4. Here we discuss the flexure-guided piezo actuator and present briefly different types of position sensors which are used for the closed loop operation of AFM scanners.

4.2.1 Flexure-Guided Piezo Nanopositioning Stages

A popular continuously moving nanopositioning system uses flexures to guide the motion. It relies on the elastic deformation of a spring-like solid metal structure which confines the motion in only one direction and is driven by a piezo element. The working principle can be seen in Fig. 4.4a. In a metal block, small trenches are cut by wire EDM (Electrical discharge machining). These trenches are shaped in a meandering way so that they act as hinges and allow a spring-like motion along one

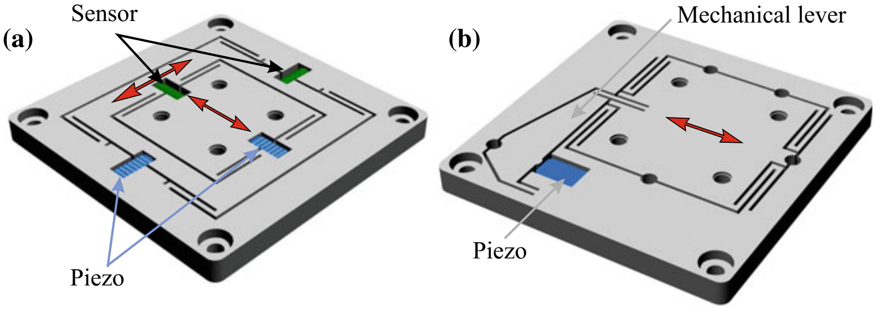


Fig. 4.4 **a** Flexure-guided piezo nanopositioning xy-stage with position sensors included. **b** Flexure-guided piezo stage with an integrated mechanical lever amplifying the motion

direction for the material inside, while being stiff along the orthogonal direction. A second set of trenches forms flexures to guide the motion along the orthogonal direction. Due to the relatively thick plate (∼10 mm), such a flexure structure is also stiff in the vertical direction. Stacks of piezo elements (blue in Fig. 4.4a) are used to move the flexures. As the mass of the inner part moved is smaller than the mass of the outer part moved, the inner part has a higher resonance frequency. Due to the higher resonance frequency the inner part is used for the motion along the fast scan direction. Often a mechanical lever is included in the flexures (Fig. 4.4b) in order to amplify the motion ranges up to about hundred micrometers.

In order to allow for a closed loop operation, position sensing detectors e.g. capacitive position detectors can be integrated in the flexure stage to allow a precise measurement of the motion (shown schematically in green in Fig. 4.4a. One disadvantage of the flexure-guided piezo nanopositioning stages is that they are relatively large and have a high mass compared to e.g. a tube scanner.

4.2.2 Closed Loop Operation of Piezoelectric Nanopositioners

A general problem of piezoelectric nanopositioners used as scanners in AFM is the non ideal behavior of piezoelectric materials. Non-linearity, hysteresis, and creep of the piezo elements lead to the effect that the applied voltage is not directly proportional to the elongation of the piezo element (as discussed in Sect. 3.5). While a simple non-linearity could be compensated by a proper non-linear calibration between the applied voltage and the distance moved, hysteresis and creep are much more difficult to compensate for, as they depend on the history of past voltages which have been applied to the piezo element.

In order to compensate for all non-ideal effects of piezoelectricity, a closed loop operation is required. This means that a position sensor measures the actual distance the piezoelectric nanopositoner moved and in a feedback loop, the voltage at the piezoelectric actuator is adjusted such that the measured displacement of the actuator reaches the desired displacement. This is the best way to eliminate all effects of piezo hysteresis and creep. However, the measurement of the piezo extension results in larger sizes of the piezoelectric actuator. Also an increased number of cables and additional control electronics are needed. Nowadays, closed loop operation is often used in atomic force microscopes.

Position sensors can be easily integrated in flexure guided nanopositioning stages, as schematically indicated in (Fig. 4.4a). In the following we list some types of position sensors; more information can be found in [3]. Generally, contactless sensors are distinguished from sensors which require contact to the nanopositioner.

Resistive strain sensors or strain gauges require contact to the nanopositioner. They consist of a thin layer of conducting foil laminated between two insulating layers. As the strain gauge is elongated, the resistance increases proportionally. Such

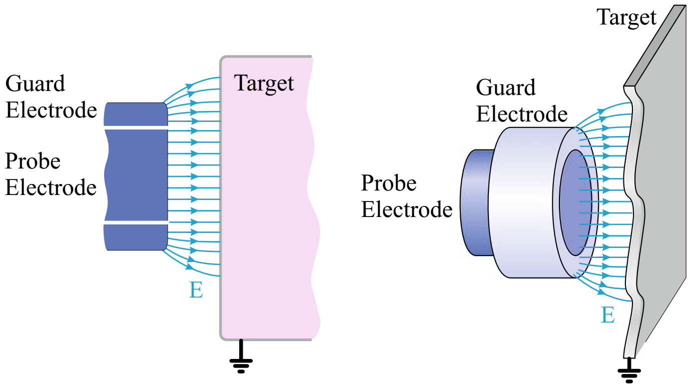


Fig. 4.5 Capacitive sensor in front of a target surface whose motion has to be sensed. The guard electrode leads to a more uniform electric field between the probe electrode and the target (left side view, right perspective view)

strain sensors have to be glued to the piezo element or a strained part of a flexure stage.

A second type of strain sensor is the piezoresistive semiconductor strain sensor. It consists of a planar n-doped silicon resistor with heavily doped contacts. When the sensor is strained, the electron mobility increases, reducing its resistance.

A third type of strain sensor is a piezoelectric strain sensor. In this type of stain sensor the piezoelectric effect is used the other way around than for piezoelectric actuators: A strain in a piezoelectric strain sensor leads to a voltage proportional to the strain, which is measured.

These types of sensors do not measure the distance moved directly, but the strain developing at the surface they are fixed to. Therefore, the stiffness of the strain gauge has to be much lower than the stiffness of the nanopositioner to which they are fixed. This is the case for flexure stage and piezo stacks. In the following we mention some types of position sensors which work contact-less.

In a capacitive sensor, the probe electrode forms a plate capacitor together with a surface of the moving object (grounded), as shown in Fig. 4.5. In order to reduce the stray fields at the edge of the probe electrode, a guard electrode, on the same electric potential as the probing electrode is used. A measurement of the capacitance can be performed for instance by applying an AC voltage to the probe electrode and measuring the resulting current. The distance (change) between the capacitor plates is obtained by the distance dependence of the capacitance of a plate capacitor. Capacitive position sensors measure distances contact-less and can be for instance integrated in flexure stage and in other nanopositioners using piezo stacks.

An eddy current sensor consists of a coil in front of a conducting target. When the coil is excited by an AC current an eddy current is induced in the target. The corresponding AC resistance depends on the magnitude of the eddy current, which in turn is distance dependent.

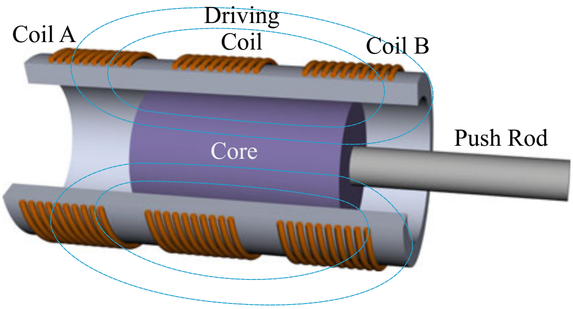


Fig. 4.6 Principle of the structure of a linear variable displacement transformer (LVDT) which uses inductive position sensing (details see text)

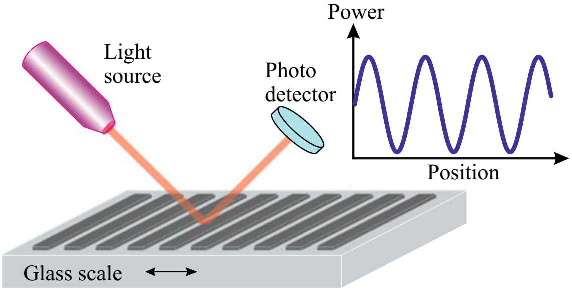


Fig. 4.7 Principle of the simplest kind of optical position sensing. Actual devices are more advanced

Linear variable displacement transformers (LVDT) consist of three coils (or sometimes only two coils) moving relative to a cylinder made from magnetically high permeable material, as shown in Fig. 4.6. An AC driving current in the center coil leads to a magnetic flux in each of the sensor coils. As the core moves, the flux through each sensor coil changes proportional to the length the inner cylinder moves in the coil. Thus, the displacement of the core is proportional to the difference of the voltages induced in the sensor coils.

Here we mention very briefly also two types of optical position sensors, while more information can be found in [3]. Laser interferometers can be used as position sensors; fiber interferometers are used due to their compactness. The very simplest form of an optical linear encoder can be imagined like the AFM beam deflection method with a laser beam reflecting from a scale bar (as shown in Fig. 4.7) and a photo-detector measuring the intensity of the reflected beam. If the scale bar consists of an periodic array of stripes of different reflectivity, the intensity measured in the photo-detector oscillates periodically as the scale bar is moved. Real optical linear encoders are somewhat more elaborated and provide also an absolute distance information. A comparison of the different types of position sensors, also with respect to their resolution can be found in [3].

4.3 AFM Design with a Tube Scanner

The principle of a design of an AFM with a tube scanner, like the multimode AFM (Bruker [4]), is shown in Fig. 4.8. This type of AFM consists of three units: the head with the optical beam deflection unit and the cantilever, the piezo tube scanner with the sample, and the base with the stepper motor used for coarse approach.

The head houses components for the optical beam deflection detection. The beam from a laser diode is reflected and focused on the reflecting back of the AFM cantilever. The laser beam is adjusted under control of an optical microscope which is mounted above the head. The laser beam reflected from the back of the cantilever is adjusted to the center of a position sensitive photo-detector. The cantilever itself is attached to a larger handle part (cantilever chip) which can be inserted using tweezers into a cantilever holder which in turn is inserted into the head. The cantilever holder comprises also a dither piezo element which is used to oscillate the cantilever in the

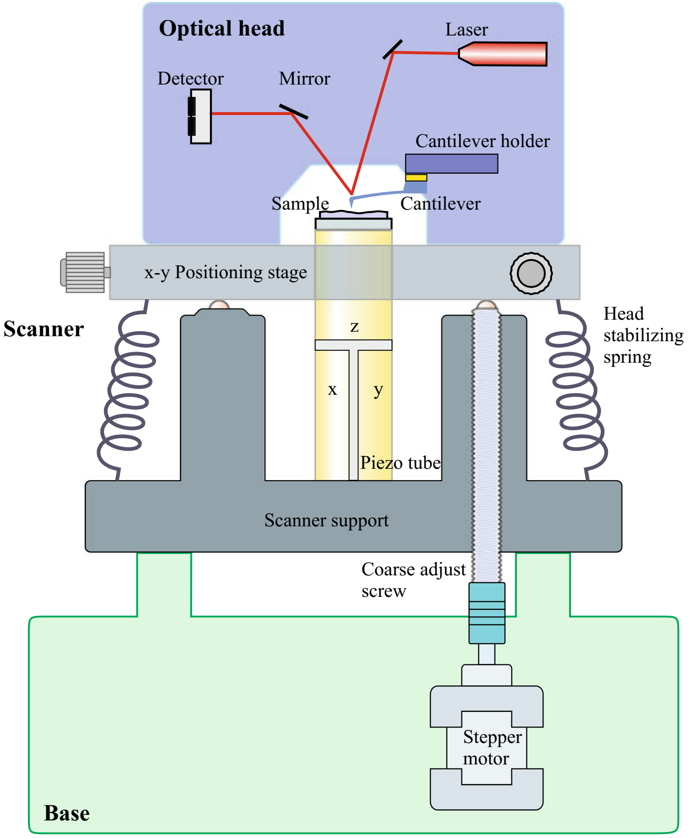


Fig. 4.8 Design principle of an AFM with a tube scanner. The main parts are the optical head, the scanner and the base unit

dynamic modes (yellow in Fig. 4.8). On the bottom of the head an xy-motion stage is attached which can move the cantilever tip with respect to the sample and thus allows AFM imaging of the desired location on the sample.

The top part of the scanner housing has three balls on which the bottom of the head is fixed by springs. One of these balls can be moved up and down by a fine pitch thread allowing for the coarse motion between cantilever tip and sample. The scanner unit consists of a cylinder which houses a tube scanner, which is used for x, y, and z fine motion. The tube scanner is segmented into two sections along its axis. The segment closer to the cantilever (upper part) is used to control the vertical tip-sample motion (z-direction), while the lower part is segmented into four quadrants, which allows lateral motion (xy-scanning) (Fig. 4.8). The z-extension part of the tube piezo element acts as a lever to enhance the lateral motion. Different scanners can be inserted with different xy-scan ranges, the smaller scan ranges (less than a micrometer) allow for higher resolution operation down to atomic resolution, while scanners with a larger xy-scan range are required if correspondingly larger areas of the sample have to be imaged. The sample is mounted on top of the piezoelectric tube scanner. The scanner unit is mounted on top of the base unit of the instrument which houses the stepper motor used to drive the coarse approach.

4.4 AFM Design with Scanners Operating in Closed Loop

The scanner described above is open loop, i.e. without position sensing, however, also closed loop scanners are available [5, 6]. The advantage of closed loop operation is obvious: Not relying on the voltage applied to the piezo element, but having an independent information about the position of the object to be moved, based on an actual measurement of the position. Due to this, image distortions resulting from hysteresis and creep of the piezo actuators are avoided.

The design principle of an AFM with closed loop operation is shown in Fig. 4.9. The sample is attached to an xy-flexure stage with integrated position sensors. The optical detection stage is attached below a z-flexure stage which is driven by a piezo element on top, as shown in Fig. 4.9. A position sensor of one of the types mentioned above is used to operate the z-positioning of the cantilever tip relative to the sample in closed loop. This closed loop operation allows to measure the true sample topography in all three directions.

The coarse approach mechanism is included by a fine thread screw and a stepper motor moving the head relative to the xy-scanning stage.

In this design the whole optical system, which has a considerable weight, is moved up and down when following the surface topography together with the cantilever in z-direction. This leads to a relatively low resonance frequency of this z-stage and thus environmental vibrations are transmitted with a larger amplitude to the tip-sample distance, as discussed in Sect. 3.6.

This can be improved by the design used e.g. in the Asylum Cypher AFM [5], or in the Park NX10 AFM [6], as shown in Fig. 4.10. Here the cantilever is coupled

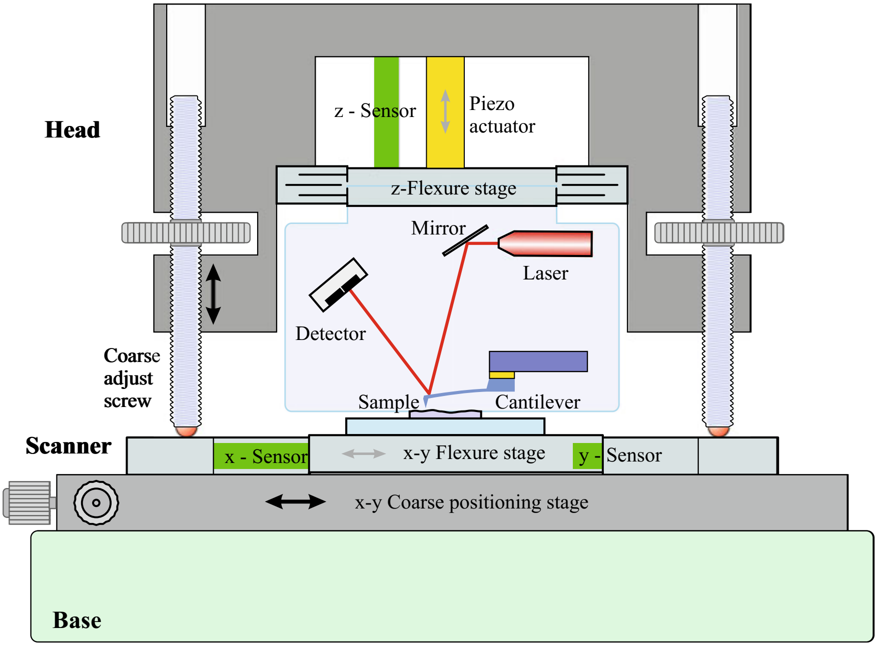


Fig. 4.9 AFM design with closed loop flexure scanners in xy-directions as well as in z-direction

to the scanner/sample stage by a small (low mass) and high stiffness mechanical loop, highlighted in red in Fig. 4.10, avoiding the large mechanical loop shown in blue in Fig. 4.10. In this mechanical loop the fine z-motion for the scanning can consist of a small z-flexure stage (or piezo stack), highlighted in orange, with an integrated position sensor, highlighted in green. Further, a coarse motion between the cantilever stage and the scanner/sample stage is needed in order to disengage tip and sample in order to allow for an exchange of the sample or the cantilever. This z-coarse positioning can be realized with a fine thread screw moved by a stepper motor. In this design the optical system is not moved together with the cantilever and the cantilever moves relative to the optical system during scanning. However, this z-motion of the cantilever does not have a large influence on the sensor signal of the photo-detector: A pure z-motion is suppressed by a large factor relative to the signal on the photo-detector which is resulting from a bending of the cantilever, as it is the case for the topography signal (this is discussed in detail in Sect. 11.5). If the laser beam is directed towards the cantilever from the top along the z-direction, it will always illuminate the same position on the cantilever independent of the height of the cantilever.

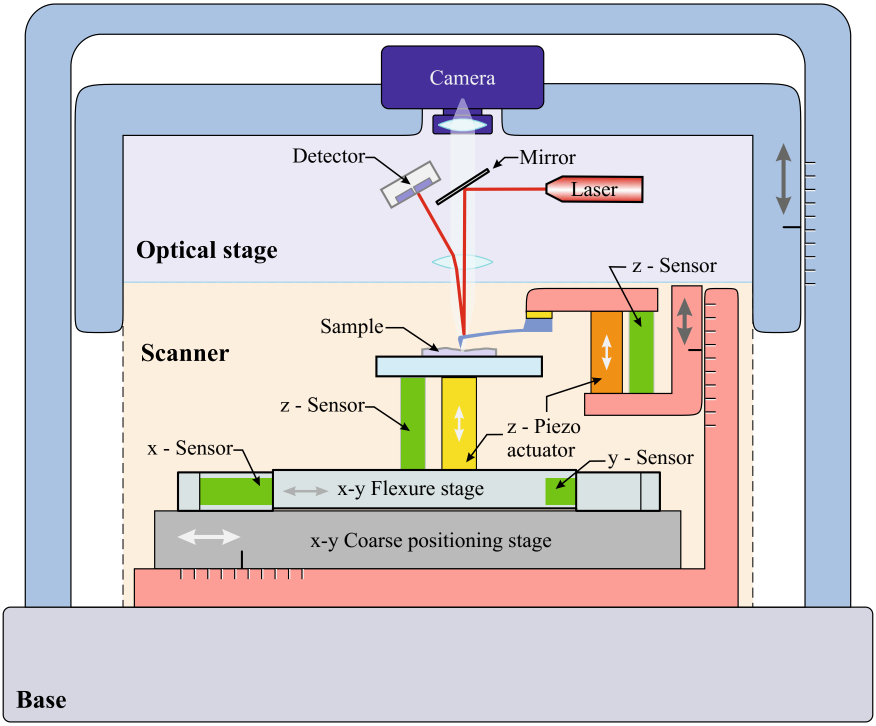


Fig. 4.10 AFM design with closed loop xy-flexure scanners and a compact stiff mechanical loop for the in z-direction

4.5 AFM Designs for Large Samples

In the AFM designs discussed until now the sample was scanned in xy-direction. Due to this the sample size is limited to relatively small samples on the order of 10 mm size. For large (>100 mm) or heavy samples different designs are used in which the sample is fixed and the cantilever (tip) is scanned [4]. In this case also the laser beam which is focused to the back of the cantilever has to be scanned with the cantilever, as the width of the cantilever can be smaller than the scan width. This does not necessarily mean that the whole optical system has to be moved in xyz during scanning. It is sufficient that the lens focusing the laser beam on the cantilever is scanned together with the cantilever. The xyz-scanning of the cantilever can be performed either by a tube scanner as the one shown in Fig. 4.8, or by a small xyz flexure stage. In order to allow for closed loop operation position sensors are included.

4.6 AFM Designs for Vacuum Operation

While most AFMs are operated in air, some AFMs are operated in vacuum or even in ultrahigh vacuum (UHV). Here the beam deflection method is less common, as the adjustment of the optical path is more difficult in vacuum and some components used at ambient conditions, like stepper motors, may be not vacuum compatible. In vacuum conditions often quartz sensors are used as AFM sensors instead of cantilevers. The use of piezoelectric quartz sensors for AFM detection is described in more detail in Chap. 18. The oscillation of these quartz sensors can be excited and detected completely electrically without the need of any optical system. Due to this, the design of AFMs using quartz sensors is basically very similar to the design of a scanning tunneling microscope (STM). The only differences are that instead of an STM tip a somewhat larger quartz sensor is used and secondly two electrical connections are required, one for the excitation and one for the detection. Often the inertial sliders discussed in Sect. 4.1.1 are used for AFM designs in vacuum. We discuss here two SPM designs which can be used together with a quartz sensor as an AFM. One design is based on inertial sliders and one design does not rely on inertial sliders.

4.6.1 Pan Slider

The Pan slider is an SPM design with very high rigidity which is mainly used in vacuum and cryogenic environments [7]. It was named after Shuheng Pan, who invented the design. The moving part is a sapphire prism containing a tube piezo scanner which in turn holds a quartz sensor (e.g. tuning fork) plus a tip (cf. Chap. 18).

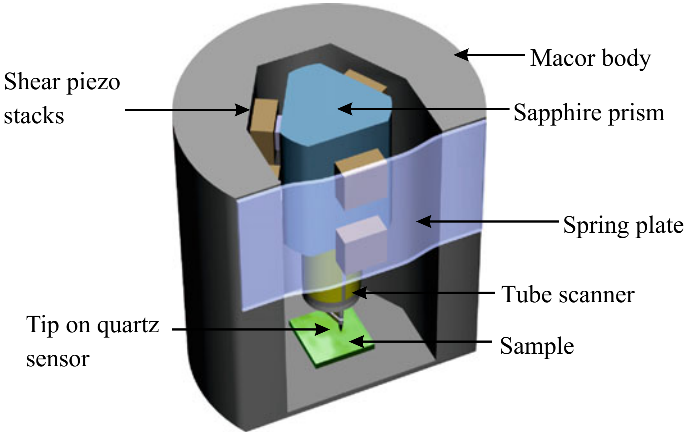


Fig. 4.11 Pan SPM design using shear piezo elements in order to move a sapphire prism on which a tube scanner is mounted

The stepping is actuated by six shear piezo stacks, as shown in Fig. 4.11. Four of the shear piezos are mounted on the interior of a Macor® body. The other two are pressed against the sapphire prism by a spring plate. With this construction the pressure on the six piezo stacks is approximately equalized. The working principle of this walker is inertial motion. First the shear piezo elements are moved quickly, so that the prism does not move (sliding phase). Then the piezos are moved slowly (riding phase). An appropriate material combination for a reliable slip-stick is given by an alumina plate mounted on top of the shear piezos. While the original design did not allow for coarse xy-motion of the sample relative to the tip, it can be upgraded by an xy-moving table below the sample usually constructed using shear piezo elements.

4.6.2 KoalaDrive

The coarse positioning unit takes up most space in a scanning probe microscope. An ultimately small SPM design can be reached if the coarse approach of the tip towards the sample is integrated *inside* a piezo tube scanner. However, here the inertial slider principle is not optimal. In order to function, an inertial slider needs inertia, i.e. a certain mass, which works against the desired miniaturization. Also the large acceleration required to move an inertial slider induces a lot of shaking of the whole mechanism. The KoalaDrive, which avoids all inertial motion was constructed at Jülich by Vasily Cherepanov et al. [8].

The task of the KoalaDrive nanopositioner is to move a rod along its axis, as shown in Fig. 4.12. For use in an AFM a quartz sensor is fixed to the end of the rod. The KoalaDrive consists of two tube piezo elements mounted one after the other, as shown in Fig. 4.12. At the ends and between the two tube piezos, three springs are mounted, holding a central rod. The upper two springs shown in Fig. 4.12

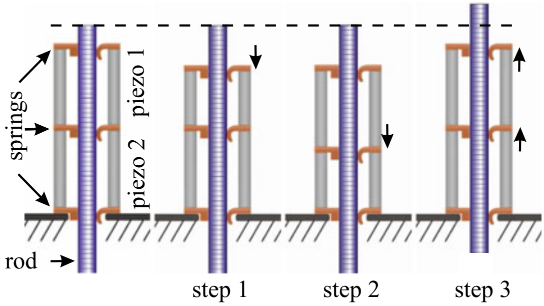


Fig. 4.12 Working principle of the KoalaDrive: concerted interplay between static friction and sliding friction. If only one spring moves, the rod is held stationary by the other two (*step 1* and *step 2*). The motion of the springs during the different steps of a cycle is indicated by *arrows*. If two springs move simultaneously, the central rod moves together with them (*step 3*)

can be moved by an extension or compression of the tube piezos along their axes. The working principle of the KoalaDrive relies on concerted consecutive motions in which the frictional surfaces between a spring and the rod alternate between static friction and sliding friction. Whenever only one spring moves, the other two will hold the rod (by static friction) and only at the single moving spring the frictional engagement will be lifted and sliding friction will occur. One cycle of motion is shown in Fig. 4.12. In step 1 of the cycle, the upper piezo element contracts and the upper spring goes into sliding friction. The central rod is kept stationary by the lower two springs, which stay in static friction with the rod. Subsequently, in step 2 the middle spring moves downwards, while the upper and the lower spring remain in their positions. For the upper spring, this is realized by a simultaneous contraction of the lower piezo element and a corresponding expansion of the upper one, leaving the upper spring unmoved. Also here a single spring (middle one) moves, while the two others keep the rod fixed. Finally, in step 3 the lower piezo extends and moves the two upper springs up simultaneously. In this case, the lower spring goes into sliding friction and the upper two springs move the rod up (static friction). In simplified terms, the working principle follows the rule: “Two are stronger than one”. If two springs move simultaneously, the central rod moves with them. If only one spring moves, the rod is kept stationary by the other two.

One single cycle can induce a motion in the range between several μm and 100 nm, which is ideally suited for a coarse approach in scanning probe microscopy. A long stroke, only limited by the length of the rod, and speeds up to 1 mm/s are possible. Most other nanopositioners used for tip-sample approach in scanning probe microscopy under vacuum conditions use the inertial motion with sawtooth-like signals inducing large accelerations causing vibrations in the system. The operating mode of the KoalaDrive is quasi-static (one cycle can even last several seconds) leading to a continuous motion without shaking, thus avoiding large accelerations. Movies of the motion of the KoalaDrive measured using an SEM during one cycle of motion are available on the internet at www.mprobes.com/koaladrive.html. These real-time movies show the motion of a tip attached to the central rod.

In the next step, the KoalaDrive can be used to build an ultra-compact SPM. The KoalaDrive is used for the tip-sample coarse approach and is integrated into a segmented scanning tube piezo element used for the xyz-scanning fine motion as shown in Fig. 4.13a. The SPM is completed by attaching quartz sensor with a tip (plus a holder) to the central rod and an outer frame, which holds the sample, as shown in Fig. 4.13a. Since the coarse approach mechanism is integrated into the piezoelectric tube scanner, no extra space for the coarse approach is required. Thus, this design leads to an SPM of minimal size: A complete SPM scanner can be integrated inside a piezo tube of 6 mm outer diameter and 12 mm length. In Fig. 4.12b, a photograph of an actual KoalaDrive SPM is shown. The use of the KoalaDrive makes the scanning probe microscopy design ultra-compact and leads accordingly to high mechanical stability.

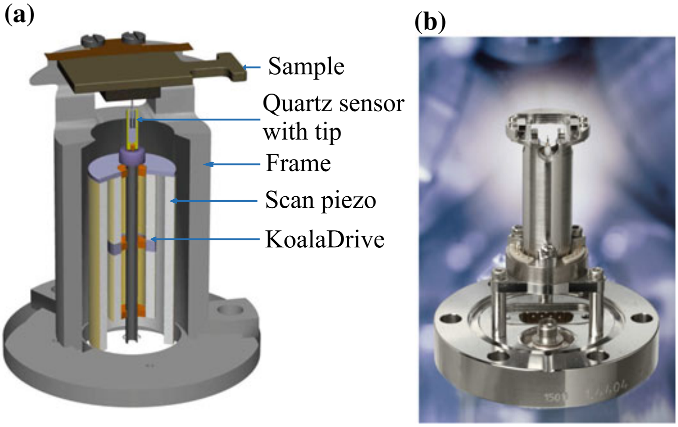


Fig. 4.13 **a** Design of an AFM using the KoalaDrive leading to an AFM of minimal size. **b** Photograph of an actual KoalaDrive SPM

4.6.3 Tip Exchange

Unfortunately, an initially sharp tip at the end of an cantilever, or attached to a quartz sensor degrades when used for some time. If the tip is used under ambient conditions it can be replaced straightforwardly by insertion of a new cantilever with a fresh tip.

When working in vacuum the quartz sensor (cf. Chap. 18) with a tip attached to it is mounted in a sensor holder, which is large enough to be handled during the transfer into the vacuum. The sensor holder (with a tip attached) is inserted into the vacuum AFM using a wobble stick or another kind of manipulator. The easiest way of inserting a sensor holder into an AFM in vacuum is if the receptacle at the AFM includes a small magnet which guides the sensor holder (made of magnetic material) to its desired position. Often a fork mechanism (or gripper mechanism) is used to release the sensor holder from the manipulator when it is in position in the vacuum AFM. Instead of magnetic forces also a spring mechanism can be used to fix a sensor holder in the AFM. Additionally, to the mechanical fixing of the AFM quartz sensor also two electrical contacts have to be provided for electrical excitation and detection in the dynamic AFM modes.

4.7 Summary

- Coarse approach is the approach between the tip and sample from the macroscopic range down to the range covered by the piezoelectric scanner. The coarse approach can be performed by a stepper motor and fine thread screw.

- Another type of coarse approach are inertial sliders. They are actuated by a sawtooth signal applied to the piezoelectric elements. During the slow slope part of the signal, the slider moves together with the support, while during the steep slope part of the signal the slider disengages from the support and does not move together with the support due to its inertia. This leads to a relative motion between slider and support in the micrometer range and below for every step.
- As AFM scanners tube piezo elements, piezo stacks, as well as flexure-guided nanopositioning stages are used. The latter have the advantage that they can be used in closed loop operation.
- In closed loop operation the non-linearity and hysteresis of piezo elements is avoided by measuring the distance actually moved. Different kinds of position sensors provide the required nanoscale position information.
- The designs of different AFM instruments use partially tube scanners and partly flexure guided scanners with closed loop operation.
- AFMs with quartz sensors as alternative sensors to cantilever type AFM sensors are mostly used in vacuum environment and use piezo tubes as scanners and particular designs for the coarse approach, such as the Pan SPM design or the KoalaDrive.

References

1. D.W. Pohl, Dynamic piezoelectric translation devices. Rev. Sci. Instrum. **58**, 54 (1987). https://doi.org/10.1063/1.1139566
2. Patent DE 40 23311 C2
3. A. J. Fleming, A review of nanometer resolution position sensors: operation and performance. Sens. Actuators A: Phys. **190**, 106 (2013). https://doi.org/10.1016/j.sna.2012.10.016
4. http://www.bruker.com
5. https://afm.oxinst.com
6. https://www.parksystems.com/
7. Patent WO 93/19494
8. V. Cherepanov, P. Coenen, B. Voigtländer, A nanopositioner for scanning probe microscopy: the KoalaDrive. Rev. Sci. Instrum. **83**, 023703 (2012). https://doi.org/10.1063/1.3681444

Chapter 5
Electronics and Control for Atomic Force Microscopy

We introduce the time domain and the frequency domain approaches to electronic signals. Then we discuss some basic electronic components, such as voltage divider, low-pass filter, and operational amplifier. We continue to discuss topics more closely related to atomic force microscopy such as the feedback electronics, which in AFM serves to stabilize the tip-sample distance. We close this chapter on electronics by discussing how digital-to-analog converters and analog-to-digital converters work in principle.

5.1 Time Domain and Frequency Domain

The usual representation of a time-dependent (electrical) signal is to analyze the signal, e.g. voltage V as the function of time $V(t)$, as it is appearing on the oscilloscope screen. However, sometimes it is also useful, or even more useful, to consider the “frequency content” of a signal. A periodic signal can be represented as a sum of sine signals of different frequencies having different amplitudes and phases (Fourier series). This is termed: signal representation in the frequency domain.

If the signal is not periodic, the Fourier transform is used to represent the signal in the frequency domain. The transform of a signal from the time domain $V(t)$ to the frequency domain $\hat{V}(\omega)$ is given by

$$\hat{V}(\omega) = \int_{-\infty}^{\infty} V(t) e^{-i\omega t} \mathrm{d}t, \tag{5.1}$$

and correspondingly the transform from the frequency domain to the time domain is given by the inverse Fourier transform as

B. Voigtländer, *Atomic Force Microscopy*, NanoScience and Technology, https://doi.org/10.1007/978-3-030-13654-3_5

$$V(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{V}(\omega) e^{i\omega t} \mathrm{d}\omega. \tag{5.2}$$

In a more sloppy notation we will skip the hat for the frequency domain quantity and write just $V(\omega)$.

5.2 Voltage Divider

One of the simplest electronic circuits is the voltage divider, which is shown in Fig. 5.1a. Applying Kirchhoff's law and Ohm's law to this circuit results in the following equations

$$\begin{aligned} V_{\text{in}} &= V_1 + V_2 = I(R_1 + R_2) && \text{(Kirchhoff's voltage law)} \\ V_2 &= R_2 I = V_{\text{out}} && \text{(Ohm's law).} \end{aligned} \tag{5.3}$$

These equations can be solved for

$$\frac{V_{\text{out}}}{V_{\text{in}}} = G = \frac{R_2}{R_1 + R_2}. \tag{5.4}$$

The output voltage divided by the input voltage is the called transfer function G. We have assumed here that the output voltage is measured with an infinite inner resistance, i.e. no current flows at the output. The limiting cases for the transfer function are $G \approx 1$ for $R_1 \ll R_2$ and $G \approx R_2/R_1$ for $R_1 \gg R_2$.

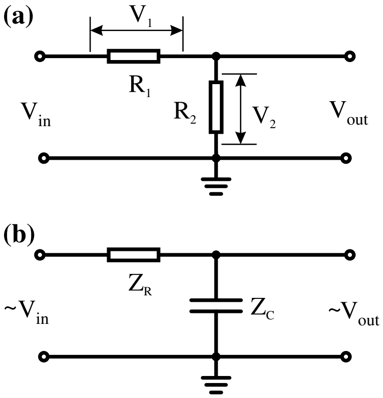


Fig. 5.1 **a** Circuit scheme of a voltage divider. The transfer function is given by $G = V_{\text{out}}/V_{\text{in}} = R_2/(R_1 + R_2)$. **b** This circuit is also a voltage divider, however, now R_2 is replaced by a capacitor and an AC input voltage is considered. Thus, we use the complex impedances Z_{R} and Z_{C} in order to obtain the transfer function

5.3 Impedance, Transfer Function, and Bode Plot

In the previous section, we considered DC voltages and currents. In the AC case, the voltages and currents can be written in the complex notation as

$$V = V_0 e^{i(\omega t + \phi_V)}, \quad \text{and} \quad I = I_0 e^{i(\omega t + \phi_I)}. \tag{5.5}$$

Of course, for ohmic resistors Ohm's law still reads as $V = RI$. For capacitances and inductors the concept of resistance can be extended to a complex impedance, which is defined as

$$\begin{aligned} Z_\mathrm{C} &= \tfrac{1}{i\omega C} && \text{for a capacity } C, \text{ and} \\ Z_\mathrm{L} &= i\omega L && \text{for a inductance } L, \text{ and of course} \\ Z_\mathrm{R} &= R && \text{for a resistor } R. \end{aligned} \tag{5.6}$$

For the impedances, the equivalent of Ohm's law applies as $V = ZI$. For AC circuits, including several impedances Z, the usual Kirchhoff laws apply, and the rules for parallel and series resistors also hold for impedances, if the quantities are represented in a complex form.

As an example, we consider the circuit shown in Fig. 5.1b, which is similar to the voltage divider, except that one resistor is replaced by a capacitor and an AC input voltage is applied. Thus, we consider the complex impedances Z_R and Z_C. The transfer function (now dependent on the frequency) can be calculated in analogy to (5.4) as

$$G(\omega) = \frac{V_\mathrm{out}}{V_\mathrm{in}} = \frac{Z_\mathrm{C} I}{(Z_\mathrm{R} + Z_\mathrm{C}) I} = \frac{\frac{1}{i\omega C}}{R + \frac{1}{i\omega C}} = \frac{1}{1 + i\omega RC}. \tag{5.7}$$

For linear systems the output signal is a sinusoidal signal at the same frequency as the sinusoidal input signal, however with modified amplitude and phase. Thus, the transfer function (output signal divided by input signal) is a complex quantity with amplitude and phase. In the Bode plot, the absolute value (modulus) of the complex transfer function and the phase difference between output voltage and input voltage are plotted, as shown in Fig. 5.2. The corresponding equations are

$$|G(\omega)| = \frac{|V_\mathrm{out}|}{|V_\mathrm{in}|} = \frac{1}{\sqrt{1 + \omega^2 R^2 C^2}}, \quad \text{and} \quad \phi_V = \arctan(-\omega RC). \tag{5.8}$$

For frequencies lower than the corner frequency $\omega_\mathrm{c} = 1/(RC)$, the absolute value of the transfer function approaches unity, i.e. gain $|V_\mathrm{out}| \,/\, |V_\mathrm{in}|$ is one. For frequencies much larger than ω_c the absolute value of the transfer function decreases as $1/\omega$. At the corner frequency, the gain has the value $1/\sqrt{2}$ (which corresponds to -3 dB). In conclusion, the circuit shown in Fig. 5.1b is a low-pass filter, which transmits signals up to the frequency ω_c with gain one and suppresses signals with higher frequencies. Another way to express this is that this circuit corresponds to a low-pass filter with a bandwidth of $\omega_\mathrm{c} = 1/(RC)$.

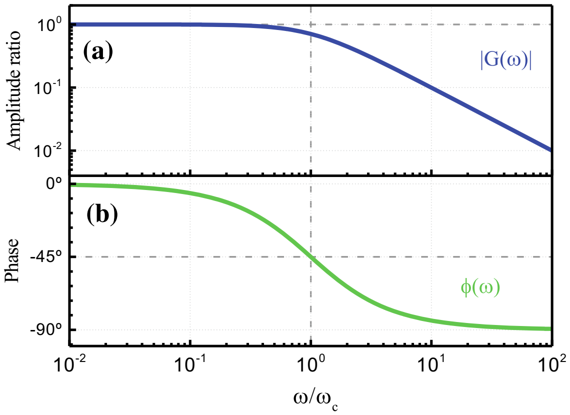


Fig. 5.2 The Bode plot shows the absolute value of the complex transfer function also called gain or amplitude ratio (**a**) and the phase shift of the output relative to the input signal (**b**). The figure shows the Bode plot of the circuit in Fig. 5.1b. The behavior of the absolute value of the transfer function (amplitude ratio) approaches the value one for frequencies lower than the corner frequency, and decreases for higher frequencies, which is the characteristic of a low-pass filter

The phase behavior of this low-pass is shown in Fig. 5.2b. The phase shift is zero for frequencies much lower than the corner frequency and approaches $-90°$ for frequencies much larger than the corner frequency.

The analysis of the low-pass circuit was one simple example, another one is if the resistor and the capacitor in Fig. 5.1b are exchanged. This circuit corresponds to a high-pass filter. Also more complicated circuits can be analyzed using Kirchhoff's laws or the rules for impedances in parallel or in series. One requirement for the type of analysis described in this section is that the input signal V_{in} is a sinusoidal signal. If the transfer function for all frequencies is known this characterizes the behavior of the circuit at all frequencies. This is a basis to obtain the output signal for all functions via Fourier methods.

5.4 Output Resistance/Input Resistance

In Fig. 5.3a we consider a device connected to a voltage source. Any kind of signal source can be replaced by an ideal voltage source with an resistor in series, which we call output resistance R_{output}, as shown in Fig. 5.3a. If the output of the signal source is connected to the input of a device, this can change the output voltage V_{out}, being no more identical to the ideal voltage source V_{signal}. The voltage V_{out} depends also on R_{input}, the input resistance of the device connected to the source. The circuit shown in Fig. 5.3a is (again) a voltage divider. Using (5.4) the output voltage V_{out} can be written as

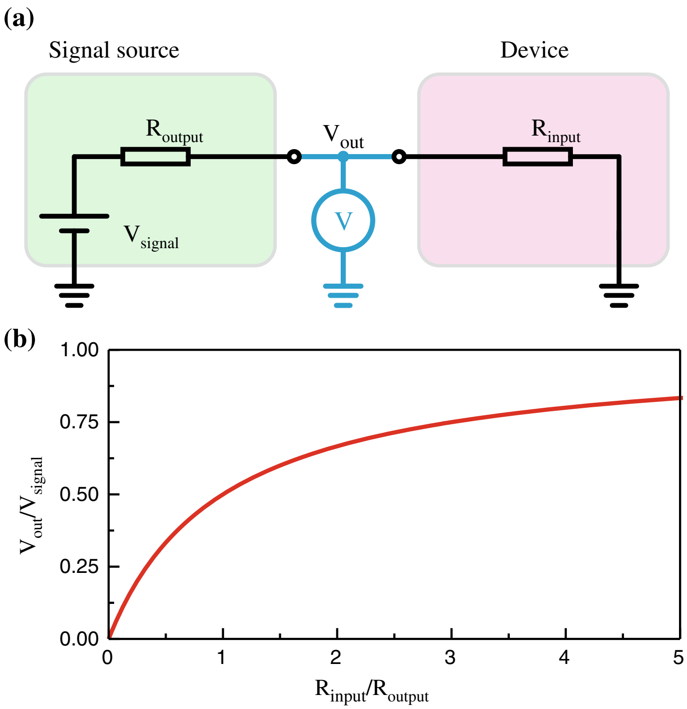


Fig. 5.3 **a** Signal source, consisting of an ideal voltage source V_{signal} and an output resistance R_{output}, connected to a device with an input resistance, characterized by the resistance between the input and the ground R_{input}. **b** The output voltage for this circuit approaches V_{signal} only if $R_{\mathrm{input}} \gg R_{\mathrm{output}}$

$$V_{\mathrm{out}} = V_{\mathrm{signal}} \frac{R_{\mathrm{device}}}{R_{\mathrm{signal}} + R_{\mathrm{device}}}, \tag{5.9}$$

and is shown in Fig. 5.3b. It can be seen that the output voltage approaches the signal voltage if $R_{\mathrm{input}} \gg R_{\mathrm{output}}$.

However, in relevant cases of small signal sources of sensors like photodiodes (in the case of atomic force microscopy), the inner resistance of the signal R_{output} is high. In such cases a impedance converter is used, which we discuss in Sect. 5.6.1 in order to convert the high output resistance of the signal source to a very low output resistance at the output of the impedance converter, which can be connected to devices with a modestly low input resistance, always maintaining the relation $R_{\mathrm{input}} \gg R_{\mathrm{output}}$.

The concept of output resistance and input resistance can be applied in sequence when connecting electronic circuits one after another. We can assign to each device in a sequence of devices an input resistance and an output resistance. In order to avoid the input of the next device modifying the output of the previous device, the relation the relation $R_{\text{input}} \gg R_{\text{output}}$ should be always maintained.

Here we considered the DC, however, the concept of output and input resistances can be extended to the AC case using the impedance replacing the resistance. Furthermore, this concept can also be used for active devices like circuits with operational amplifiers, discussed in Sect. 5.6.

5.5 Noise

If we consider a DC electric signal with some time-dependent fluctuations such as the current $I(t)$ or the voltage $V(t)$, it can be characterized by its average

$$\langle V \rangle = \lim_{T \to \infty} \frac{1}{T} \int_0^T V(t) \mathrm{d}t. \tag{5.10}$$

Fluctuations of the voltage around this average are called the noise as $\Delta V(t) = V(t) - \langle V \rangle$. This is still a time-dependent quantity and its average is zero. If the noise is due to random fluctuations, it is usually characterized by the following time independent quantity

$$\sqrt{\langle \Delta V^2 \rangle} = \sqrt{\lim_{T \to \infty} \frac{1}{T} \int_0^T (V(t) - \langle V \rangle)^2 \mathrm{d}t.} \tag{5.11}$$

also called root mean square (RMS) noise.

The above considerations about the noise were in the time domain, i.e. considering the time-dependent signal $V(t)$ and the time-dependent noise $\Delta V(t)$. Often it is more convenient to consider the frequency components of a (noise) signal using the Fourier transform of the time signal. This has the advantage that instead of the complicated time dependence of a (noise) signal simple sinusoidal frequency components are considered. If the (noise) signal is fed into a linear system, as for instance a low-pass, the output frequency component is the input multiplied by the transfer function $G(\omega)$. In the following, we will consider the frequency dependence of a (noise) signal, which is often named analysis in the frequency domain. The transition between the time domain and the frequency domain is performed by the Fourier transform.

An important quantity describing a noise signal in the frequency domain is the power spectral density (PSD). In a theoretical treatment the PSD is defined as the Fourier transform of the auto-correlation function. A more practical equivalent definition defines the PSD as proportional to the absolute square of the Fourier transform of the time signal $\Delta V(t)$ [1, 2].

The single sided power spectral density (PSD) of the noise $\Delta V(t)$ (or generally of a time-dependent signal) is

$$N_{\mathrm{V}}^2(f) = \lim_{T\to\infty} \frac{1}{T} \left| \int_{-\infty}^{\infty} \Delta V(t) e^{-2\pi i f t} \mathrm{d}t \right|^2 . \tag{5.12}$$

An important property of the single sided power spectral density of the noise is that it relates to the mean square noise as

$$\left\langle \Delta V^2 \right\rangle = \int\limits_0^{\infty} N_{\mathrm{V}}^2(f)\mathrm{d}f = \frac{1}{2\pi} \int\limits_0^{\infty} N_{\mathrm{V}}^2(\omega)\mathrm{d}\omega = \int\limits_0^{\infty} N_{\mathrm{V}}^{\prime 2}(\omega)\mathrm{d}\omega, \tag{5.13}$$

where $N_{\mathrm{V}}^2(f)$ is the natural frequency PSD, while $N_{\mathrm{V}}^{\prime 2}(\omega)$ is the angular frequency PSD. As measurements are usually performed using the natural frequency, we will use f in the following.

If a detection scheme is used which measures the noise variable only within a certain bandwidth $B = f_2 - f_1$ between f_1 and f_2, the band limited mean square noise can be written as

$$\left\langle \Delta V^2 \right\rangle_B = \int\limits_{f_1}^{f_2} N_{\mathrm{V}}^2(f)\mathrm{d}f. \tag{5.14}$$

The noise PSD indicates how much power the noise signal carries in a small region around f. The spectral density of the noise is defined as $N_{\mathrm{V}} = \sqrt{N_{\mathrm{V}}^2}$. If the noise spectral density is constant between f_1 and f_2 (5.14) reduces to

$$\left\langle \Delta V^2 \right\rangle_B = (f_2 - f_1) N_{\mathrm{V}}^2(f), \tag{5.15}$$

and we obtain

$$\sqrt{\left\langle \Delta V^2 \right\rangle_B} = N_{\mathrm{V}} \sqrt{B}. \tag{5.16}$$

The spectral density of the noise variable ΔV is expressed in the unit of the noise variable per $\sqrt{\mathrm{Hz}}$, for instance volt/$\sqrt{\mathrm{Hz}}$. If the angular frequency is used the angular frequency bandwidth $B_\omega = \omega_2 - \omega_1$ is in units of rad/s, not cycles/s. Correspondingly, the unit of the noise spectral density $N_{\mathrm{V}}(\omega)$ is volt/$\sqrt{\mathrm{rad} \cdot \mathrm{Hz}}$.

The power spectral density can be measured using a device named spectrum analyzer. However nowadays, hardware spectrum analyzer instruments are less frequently used in favor of analog to digital conversion of the measured signal followed by a subsequent software discrete Fourier transform (DFT). As the PSD is proportional to the absolute square of the Fourier transform a signal proportional to the PSD is easily obtained. However, the correct calibration is no more "included" as it was in the hardware spectrum analyzer. Since the discrete Fourier transform transforms just n numbers to n new numbers, the user has to take care about the necessary calibration steps, and errors may occur in this calibration. While this is straight forward in theory, it involves a number of non-trivial details. In the signal processing from the time series of the signal to the spectral density several proportionality factors are involved, for example: RMS amplitude or peak amplitude, two-sided spectrum or single-sided spectrum, natural frequency PSD or angular frequency PSD, window type, etc. So one has to consider all these factors carefully. Due to this non-trivial proportionality factors, complementary also an experimental calibration of the power spectral density is very desirable and will be considered in Appendix B.

5.6 Operational Amplifiers

Since operational amplifiers (op-amp) are the basic building blocks of analog electronics a brief introduction to their operation is given. An operational amplifier can be considered as a "gain block" amplifying the difference between the input voltages (ideally possessing very high gain). The voltage at the output is the amplified voltage difference at the inputs. Outside of the gain block there is a feedback-loop network (e.g. consisting of resistors), which controls the actual gain. Operational amplifiers operated close to DC have typically the following properties:

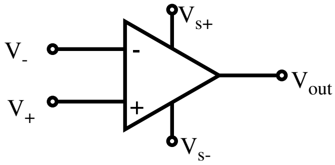
- Very high input resistance, with a typical input current of a few pA,
- Very low output resistance, typically a few ohm,
- Very large open-loop voltage gain G (10^4–10^6).

We will show that if these properties of an operational amplifier are met the characteristics of the amplifier are determined by the feedback-loop network only, not the gain block itself. We are not concerned with the inner working of the operational amplifier. A block diagram of an operational amplifier is shown in Fig. 5.4. The output voltage is the difference of the input voltages multiplied by the open-loop gain G as

$$V_{\text{out}} = G(V_+ - V_-). \tag{5.17}$$

Due to the very high open-loop gains of operational amplifies, they are usually not operated in an "open" configuration, because any voltage difference exceeding the sub-millivolt range will saturate the output voltage which is limited to the supply voltage V_{s}.

Fig. 5.4 Block diagram of an operational amplifier showing the supply voltages V_{s}, the input voltages $V_{\pm}$ and the output voltage V_{out}



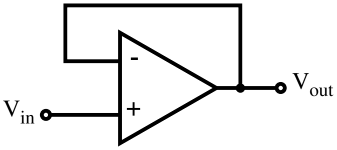


Fig. 5.5 Operational amplifier wired as a voltage follower. A negative feedback is realized by connecting the output to the negative (inverting) input

5.6.1 Voltage Follower/Impedance Converter

If we connect the output of an operational amplifier to its negative (inverting) input (Fig. 5.5) and apply a voltage signal to the non-inverting input, we will find that the output voltage of the op-amp closely follows that input voltage.

In order to find an expression for V_{out} for the circuit in Fig. 5.5 we start from (5.17) which states that the output voltage is the difference of the input voltages times the open-loop gain. In our case the positive input voltage V_+ is V_{in} and the negative feedback voltage V_- is due to the negative feedback V_{out}. Thus (5.17) reads

$$V_{\text{out}} = G(V_{\text{in}} - V_{\text{out}}), \tag{5.18}$$

which leads to

$$V_{\text{out}} = V_{\text{in}} \frac{G}{1+G}. \tag{5.19}$$

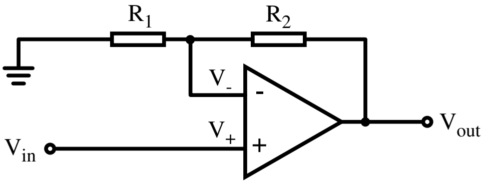
For a large open-loop gain, the output voltage is approximately equal to the input voltage $V_{\text{out}} \sim V_{\text{in}}$.

Taking the output voltage of the operational amplifier and coupling it to the inverting input is a technique known as negative feedback. In this circuit the operational amplifier has the capacity to work in a linear mode, as opposed to merely being fully saturated (due to the high gain) with no feedback for voltage differences exceeding the mV range.

Here, as in the other operational amplifier circuits we will discuss, the actual gain (which is one here) is not determined by the open-loop gain of the operational amplifier but by the outer feedback circuit (which is just a simple connection between V_{out} and V_-). It might be imagined that an amplifier with a gain of one is useless. However, this circuit acts as an impedance converter, since a high input resistance/impedance (being an intrinsic property of an op-amp) is converted to a low output resistance/impedance (being another intrinsic properties of an op-amp).

While having "only" a voltage gain of one, the voltage follower has a power (current) gain. The voltage follower is often used as "buffer" to interface a large impedance output signal to device with a low impedance (input) load. The voltage follower as impedance converter acts as interface device, drawing almost no current from the source supplying its input (because of its high input resistance), and it can supply a large amount of current to loads with low (input) impedance.

Fig. 5.6 Operation principle of non-inverting amplifier



5.6.2 Voltage Amplifier

If we add a voltage divider to the feedback wiring (Fig. 5.6) only a fraction of the output voltage is fed back to the inverting input. In this case the output voltage is a multiple of the input voltage.

The gain of this circuit can be calculated taking the basic equation (5.17) into account. If the output is connected to the inverting input, via a voltage divider network, V_- can be written (using Ohm's and Kirchhoff's laws[1]) as $V_- = V_{\text{out}} \frac{R_1}{R_1+R_2} = V_{\text{out}} K$, and V_{in} is connected to the positive input V_+, then

$$V_{\text{out}} = G(V_{\text{in}} - KV_{\text{out}}). \tag{5.20}$$

Solving this equation for $V_{\text{out}}/V_{\text{in}}$, we find

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{G}{1+KG}. \tag{5.21}$$

If G is very large the gain becomes

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{1}{K} = 1 + \frac{R_2}{R_1}. \tag{5.22}$$

We can change the voltage gain of this circuit just by adjusting the values of R_1 and R_2 (changing the ratio of output voltage which is fed back to the inverting input).

While we have used in the basic equation for the operational amplifier (5.17) together with the analysis of the feedback circuit using Ohm's and Kirchhoff's laws, the analysis of operational amplifier circuits can be simplified using two simple rules. The rule that the input current of an operational amplifier vanishes we have already used in our analysis. Due to the very high gain G, the difference between the inputs V_+ and V_- approaches zero. This is a general rule, leading to the following two "golden rules" which simplify the analysis of circuits with operational amplifiers.

- The input current to an operational amplifier vanishes (high input impedance).
- The difference between the inputs V_+ and V_- approaches zero.

[1] $V_{\text{out}} = V_1 + V_2 = I(R_1 + R_2) = (V_1/R_1)(R_1 + R_2) = V_- \frac{R_1+R_2}{R_1}$.

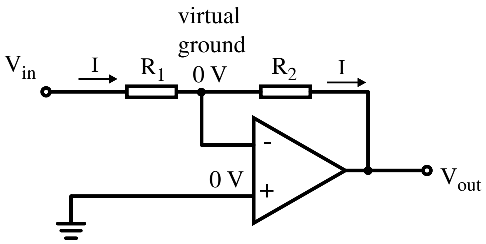


Fig. 5.7 Circuit of an inverting amplifier realized with an operational amplifier

In the following we calculate the output voltage for the circuit shown in Fig. 5.7 using above "golden rules" for operational amplifiers. In this circuit a negative feedback is provided through a voltage divider, but the input voltage is applied to the inverting input and the non-inverting input is grounded. The second "golden rule" tells us that the voltage at the inverting input is zero. Thus, the inverting input is referred to in this circuit as a *virtual ground*, being kept at ground potential (0 V) by the feedback, yet not directly connected to (electrically common with) ground. Since the input current to the operational amplifier is zero (first "golden rule"), the current through R_1 and R_2 are the same. By applying Ohm's law to the two resistors the gain can be calculated as

$$\frac{V_{out}}{V_{in}} = \frac{-IR_2}{IR_1} = -\frac{R_2}{R_1}. \tag{5.23}$$

Note that the output voltage always has the opposite polarity of the input voltage. For this reason, this circuit is referred to as an inverting amplifier.

5.7 Current Amplifier

In AFM detection the current of a photo diode, corresponding to the deflection of a cantilever is converted to a voltage by a current amplifier. Such amplifiers are also called transimpedance amplifiers and already the circuit shown in Fig. 5.7 can serve as such a current-to-voltage converter. If we consider the voltage source plus the resistor R_1 as a current source, a current of $I_{in} = V_{in}/R_1$ flows to the virtual ground. Since the input current of the operational amplifier is practically zero (high input resistance), this current flows through the feedback resistor R_2. In the actual current amplifier shown in Fig. 5.8, the input current I_{in} has to flow through the resistor R_{FB}. Therefore, $I_{in} = I_{FB} = -V_{out}/R_{FB}$. Or

$$V_{out} = -I_{in}R_{FB}. \tag{5.24}$$

The input current is converted to an output voltage with R_{FB} as proportionality factor. As an example: If the feedback resistor has a value of $R = 1\ \text{G}\Omega$, one nanoampere

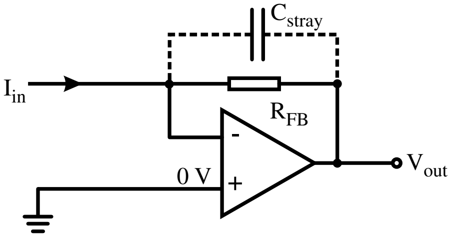


Fig. 5.8 Circuit used as current amplifier, e.g for the current of a photo diode. The gain (actually transconductance in V/A) is proportional to the resistance of the feedback resistor R_{FB}. The bandwidth of this current amplifier is limited by the stray capacitance C_{stray}

of input current results in an output voltage of 1 V. Due to the high input resistance of an operational amplifier and its low output resistance, a high input impedance is converted to a low impedance output which can be processed further.

Up to now we have considered the operational amplifier circuits as DC circuits. In the following, we consider the AC performance of the current amplifier shown in Fig. 5.8 and will show that its bandwidth is limited by the stray capacitance C_{stray} parallel to the feedback resistor. We use the complex impedance to analyze this AC circuit. The complex impedances for a resistor R and a capacity R are $Z_{\mathrm{R}} = R$, and $Z_{\mathrm{C}} = 1/(i\omega C)$, respectively. Since the two impedances in the feedback arm of the operational amplifier are in parallel, the following expression results for the total (complex) impedance Z as

$$\frac{1}{Z} = \frac{1}{Z_{\mathrm{R}}} + \frac{1}{Z_{\mathrm{C}}} = \frac{1}{R} + i\omega C. \tag{5.25}$$

The absolute value of the complex impedance results as

$$|Z| = \frac{R}{\sqrt{1 + (\omega RC)^2}}. \tag{5.26}$$

Replacing according to (5.24) $V_{\mathrm{out}} = -ZI_{\mathrm{in}}$, and identifying R with R_{FB}, as well as $C = C_{\mathrm{stray}}$ results in

$$V_{\mathrm{out}} = \frac{-I_{\mathrm{in}} R_{\mathrm{FB}}}{\sqrt{1 + \left(\omega R_{\mathrm{FB}} C_{\mathrm{stray}}\right)^2}}. \tag{5.27}$$

This frequency dependence of the output voltage of the current amplifier is the same as that of a simple passive low-pass with a resistor and a capacitor. The corner frequency of such a low-pass at which the output voltage drops by $1/\sqrt{2}$ is $f_{\mathrm{corner}} = 1/\left(2\pi R_{\mathrm{FB}} C_{\mathrm{stray}}\right)$. As an example, if by careful design the stray capacitance can be reduced to 0.1 pF a bandwidth of 1.5 kHz is obtained for a feedback resistance of 1 GΩ. The bandwidth of the amplifier is the frequency range which is amplified

Table 5.1 Gain, bandwidth and noise for a current amplifier with $R_{\mathrm{FB}} = 100\,\mathrm{M}\Omega$ and $R_{\mathrm{FB}} = 1\,\mathrm{G}\Omega$

$C_{\mathrm{stray}} = 0.5\,\mathrm{pF}$	$R_{\mathrm{FB}} = 100\,\mathrm{M}\Omega$	$R_{\mathrm{FB}} = 1\,\mathrm{G}\Omega$
Gain	10^8 V/A	10^9 V/A
Bandwidth	3 kHz	300 Hz
Noise	0.3 pA	0.1 pA

without significant loss of the signal (i.e. from DC to $f_{\mathrm{corner}} \sim 1/(2\pi R_{\mathrm{FB}} C_{\mathrm{stray}})$. It can be seen that the gain which is proportional to R_{FB} and the bandwidth proportional to $1/R_{\mathrm{FB}}$ are opposing figures of merit. Increasing the amplification means decreasing the bandwidth and vice versa. Some numerical examples are given in Table 5.1.

Another figure of merit for amplifiers is the noise. The (RMS) noise induced by the thermal excitation of the electrons in a resistor R is called Johnson noise [3, 4] and can be calculated as

$$I_{\mathrm{noise}} = \sqrt{\frac{4k_{\mathrm{B}}TB}{R_{\mathrm{FB}}}}. \tag{5.28}$$

with B being the bandwidth of the measurement (from DC to a maximum frequency) and k_{B} the Boltzmann constant. In Table 5.1 some numerical values are given.

5.8 Feedback Controller

In atomic force microscopy, a feedback controller is used to follow the surface topography. Before we come to the application of a feedback controller to AFM, we will consider feedback controllers in general. A general model for a feedback loop is shown in Fig. 5.9. In the control loop, the system output x (measured constantly by a sensor) is fed back to the input side, and compared to the setpoint w by subtraction $w - x = e$. Depending on this error signal e, the controller determines a system input (control signal) y, which is fed into the system in order to adjust the system output x to the setpoint value w. This whole operation of the controller acts in a closed feedback loop and fulfills the task of adapting the system output to the setpoint in the presence of an external disturbing signal d. We consider the controller as well as the system as linear systems. If the system is non linear, only so small deviations from the working point are considered that the system response can be approximated as linear.

Since the treatment of the feedback loop is often quite abstract and formal, we will first consider a simple example: the heating system of a house in winter. The simplest example of a feedback system is the on-off controller. On your thermostat you set a certain desired temperature (setpoint) w. If the measured temperature x is lower than w the controller gives a signal y to the system. For the case of the heating system of a house, y is the heating power which is turned on from zero to a certain power; thus the radiators heat the rooms until the set point temperature w

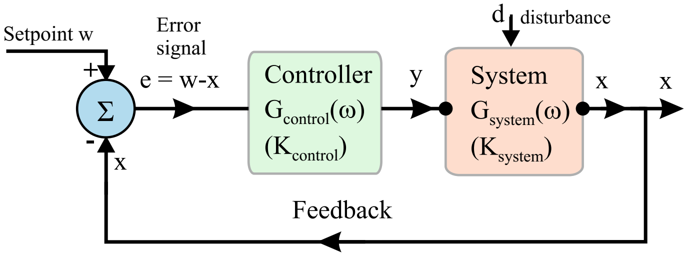


Fig. 5.9 General model for a feedback loop with the set point w as input parameter, a controller which has the error signal $w - x$ as input, and the system with its output x, which is fed back to the input and subtracted from the setpoint. The controller and the system can be described by their transfer functions $G_{\text{control}}(\omega)$ and $G_{\text{system}}(\omega)$, respectively. In the very simplest case the transfer functions do not depend on the frequency and are simple proportionality constants K_{control}, and K_{system}, respectively

is reached. Due to the inertia of the system (i.e. the time delays) the temperature in the rooms will continue to rise for some time after the heating has been switched off (temperature overshoot), because the radiators are still warm. You can easily imagine how this cycle continues. For instance, when the measured temperature x falls below the setpoint temperature w it will take some time after the heating is started before the radiators become warm. In conclusion, the actual temperature x fluctuates around the desired temperature w. What controller theory is all about is to find a smart way to keep x as close as possible to w.

In the formal language of control theory the controller, as well as the system can be described by their transfer functions as function of the frequency. $G(\omega)$ = output signal/input signal, as shown in Fig. 5.9. The transfer function can be measured by applying a sinusoidal input signal of frequency ω and measuring the amplitude and phase of the output signal (the sinusoidal signals are represented in the complex notation, e.g. $y = y_0 e^{i(\omega t + \phi)}$. The transfer function $G(\omega)$ is a complex function with amplitude and phase. The transfer function can be graphically represented by the Bode plot. The Bode plot for a low-pass as example was shown in Fig. 5.2a.

In the following we calculate the transfer function of the closed-loop feedback system from the (open-loop) transfer functions $G_{\text{control}}(\omega)$ and $G_{\text{system}}(\omega)$. We make use of the rule that the total transfer function of two systems is the product of the two individual transfer functions. We start from the output of the feedback loop x and work backwards: The output signal of the loop x is equal to the input of the system y times the system transfer function $G_{\text{system}}(\omega)$. The input of the system y is also the output of the controller and thus $y = (w - x)\, G_{\text{control}}(\omega)$. Thus, repeating these steps in equations, the closed-loop system output can be written as

$$\begin{aligned} x &= y\, G_{\text{system}}(\omega) \\ &= (w - x)\, G_{\text{control}}(\omega)\, G_{\text{system}}(\omega). \end{aligned} \tag{5.29}$$

As the closed-loop transfer function T is output x divided by input w, the closed-loop transfer function can be obtained from (5.29) as

$$T(\omega) = \frac{x}{w} = \frac{G_{\text{control}}\, G_{\text{system}}}{1 + G_{\text{control}}\, G_{\text{system}}}. \tag{5.30}$$

It should be mentioned that this transfer function is the steady-state transfer function, after initial transients due to the initial conditions have decayed. If the initial transients should be considered, numerical simulations can be used, or advanced concepts of control theory like the complex frequency and the Laplace transform have to be used, which are beyond the scope of the current treatment [5, 6].

We consider in the following the very simple case that the system can be described by a frequency independent constant transfer function K_{system}, such that $y = x\, K_{\text{system}}$. One of the simplest controllers is the proportional controller which is described by another frequency independent constant $G_{\text{control}} = K_{\text{control}} = K_{\text{P}}$ and will be discussed in the following.

5.8.1 Proportional Controller

If in the example of the heating system of a house, a heater with a continuously variable heating power is available (not just on or off), a proportional controller (P controller) can be realized. For the P controller the output of the controller y is proportional to the error signal $w - x(t)$, as

$$y(t) = K_{\text{P}}(w - x(t)). \tag{5.31}$$

The proportional constant K_{P} is called proportional gain. Since the heating power is now proportional to the error signal it is obvious that the temperature can be controlled much better with much less overshoot than for the on-off controller. (Actually, the on-off controller is a P controller with infinite gain K_{P}, in which the heating power is limited by the maximum heating power of the heater.) Since the output of the controller is (ideally) instantaneously proportional to the error signal, the P controller is a fast reacting type of controller.

In the frequency domain the P controller has a very simple constant transfer function $G(\omega) = K_{\text{P}}$. Thus, the Bode plot of the transfer function as function of the frequency has the constant value K_{P} for the gain and a frequency independent phase of $0°$.

One problem with the proportional controller is that a pure proportional control will not settle at the setpoint value w, but will retain a steady-state error, which depends on the proportional gain. This can be qualitatively understood as follows. If in the example of our heating system we have continuous losses of heat (in winter it is outside cooler than inside), therefore, we need a non-zero heating power in order to maintain the setpoint temperature, even if the error signal is zero. However, the pure proportional controller does not provide this. According to (5.31) the control signal y is zero for zero error signal $w - x$. In the reverse conclusion this means that the pure proportional controller cannot reach the setpoint w. The higher the external disturbance (i.e. the cooler it is outside) the greater is the deviation from the setpoint value. Increasing the proportional gain can reduce the deviation but it never goes to zero and high gain can lead to instabilities (oscillations) in the feedback loop.

In the following we derive an expression for the steady-state error of a P controller ($K_{\text{control}} = K_{\text{P}}$) and a system characterized by a constant transfer function K_{system}. We include also a disturbance signal d acting on the system, as shown in Fig. 5.9, which contributes to the output signal of the system as $K_{\text{system}}\, d$. The steady state error can be written as

$$\begin{aligned} e_{\text{steady}} &= w - x \\ &= w - (K_{\text{system}}\, y + K_{\text{system}}\, d) \\ &= w - K_{\text{system}} K_{\text{P}}\, e_{\text{steady}} - K_{\text{system}}\, d. \end{aligned} \tag{5.32}$$

This equation can be solved for e_{steady}, resulting in

$$e_{\text{steady}} = \frac{1}{1 + K_{\text{system}} K_{\text{P}}} \left(w - K_{\text{system}}\, d \right). \tag{5.33}$$

The system transfer function K_{system} is system immanent, however, the controller transfer factor K_{P} can be increased in order to minimize the steady-state error e_{steady}. However, as we will see later, a high proportional gain can lead to an instability of the feedback loop. Therefore, there are limits for the increase of the proportional gain. Equation (5.33) also shows that a disturbance signal d acting on the system is equivalent to a change of the setpoint w. Therefore, we will mimic below a disturbance signal by changing the setpoint, which can be realized easier. Here we discussed the steady-state behavior of the feedback system, while the initial transient, i.e. the behavior of the system when it approaches towards the steady-state will be discussed below.

In summary the advantage of the P-controller its fast reaction time, the controller output is instantaneously directly proportional to the error signal. The disadvantage of the P controller is the steady-state deviation of the system output from the desired setpoint value.

5.8.2 Integral Controller

The integral controller provides a control signal proportional to the accumulated deviations from the setpoint. The contribution from the integral term is proportional to both the magnitude of the error and the duration of the error. Summing the instantaneous error over time (integrating the error) corresponds to an accumulated effect that should have been corrected previously. For the I controller the output of the controller y is written as

$$y(t) = K_\mathrm{I} \int_0^t (w - x(\tau))\mathrm{d}\tau. \tag{5.34}$$

The proportional constant K_I is called integral gain. The integral controller eliminates the residual steady-state error that occurs with a proportional controller. A disadvantage of this type of controller is the slower reaction to changes of the input signal, due to the integration. For small times the value of the integral is small. Of course also the I controller can be made faster (shorter reaction time) by increasing K_I, however, this also increases the tendency towards unstable and oscillating behavior. In a variant of the I controller, the integration is not performed from zero, but over a time interval Δt prior to the current time.

Let us now discuss the steady-state behavior of the I controller. We assume a step in the setpoint (or alternatively a step in the disturbing signal). If, after an initial transient, the I controller has adjusted the output signal such that the momentary error signal vanishes, i.e. $w - x = 0$, this vanishing error signal will be kept in the steady-state. If the error signal vanishes, no new contribution adds to the integral and the controller output signal y corresponding to the vanishing error signal is maintained constant in the steady-state.

In order to derive the transfer function of the I controller in the frequency domain, we start from the expression (5.34) for the output signal of the I controller and insert an oscillatory (complex) input signal as input signal $e(\tau)$ of the I controller. This results in

$$\begin{aligned} y &= K_\mathrm{I} \int_0^t e(\tau)\mathrm{d}\tau \\ &= K_\mathrm{I} \int_0^t e_0 e^{i(\omega\tau+\phi)}\mathrm{d}\tau \\ &= \tfrac{K_\mathrm{I}}{i\omega} e(t). \end{aligned} \tag{5.35}$$

The transfer function in the frequency domain results as

$$G(\omega) = \frac{y}{e} = \frac{-iK_\mathrm{I}}{\omega}. \tag{5.36}$$

The Bode plot resulting from this transfer function leads to an amplitude ratio (gain) of $|G| = K_\mathrm{I}/\omega$ and a constant phase of −90° (c.f. Fig. 5.11). The high gain at low frequencies, proportional to $1/\omega$, leads to a small error signal of the closed feedback loop for low frequencies, which we have seen before as the vanishing steady state error. However, the gain decreases at high frequencies due to the $1/\omega$ behavior and thus high frequency deviations from the setpoint cannot efficiently compensated by the I controller.

In summary the advantage of the I controller the vanishing the error signal in the steady-state. The disadvantage of the I controller is the its slower reaction time due to the integration of the error signal.

5.8.3 Proportional-Integral Controller

In a PI controller the P and the I control signals are added up, as shown in Fig. 5.10. In this controller, the advantages of both the P and I controllers are combined, while avoiding their individual disadvantages. Short-term deviations from the setpoint are compensated fast by the proportional controller and long-term deviations are compensated by the integral controller. This type of controller can regulate the error signal to zero in steady-state. The output signal can be written as

$$y(t) = K_\mathrm{P}(w - x(t)) + K_\mathrm{I} \int\limits_0^t (w - x(\tau))\mathrm{d}\tau. \tag{5.37}$$

Often also a differential controller is added resulting in total in a PID controller. However in atomic force microscopy the noise is usually so large that a D controller (which is particularly prone to noise) is not used and the PI controller is standard in AFM.

The transfer function of the PI controller in the frequency domain is obtained by adding the transfer functions of the P controller and the I controller. This is the case, as the P controller and the I controller are connected in parallel and add up, as shown in Fig. 5.11. Thus, the transfer function of the PI controller results as

$$G(\omega) = K_\mathrm{P} + \frac{K_\mathrm{I}}{i\omega}, \tag{5.38}$$

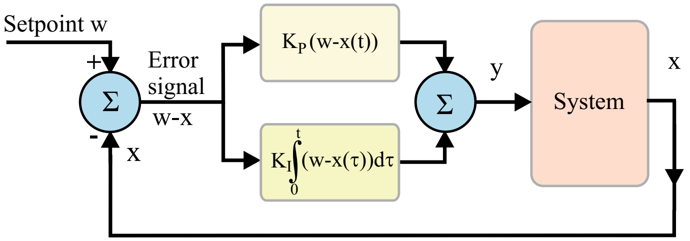


Fig. 5.10 Schematic of a PI controller in which the control signals of the P controller and the I controller are added

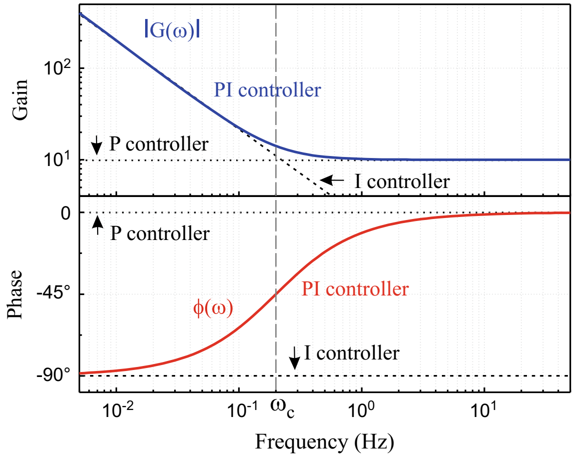


Fig. 5.11 Bode plot of a PI controller. At low frequencies the I controller part is dominant, resulting for instance in a vanishing steady state error. At high frequencies the P controller is dominant controlling small high frequency deviations from the setpoint. The transition between the P and the I behavior occurs at the cross-over frequency ω_c. The behavior of a pure P and a pure I controller is indicated as dotted and dashed lines, respectively

and the corresponding amplitude ratio (gain) and phase result as

$$|G(\omega)| = \sqrt{K_P^2 + \frac{K_I^2}{\omega^2}}, \quad \text{and} \quad \phi = \arctan\left(\frac{\text{Im}(G)}{\text{Re}(G)}\right) = \arctan\left(\frac{-K_I}{K_P\omega}\right). \quad (5.39)$$

Using these equations, the Bode plot for the parameters $K_P = 10$ and $K_I = 2$ is shown in Fig. 5.11. The Bode plot of the PI controller follows for low frequencies the behavior of the I controller, with a $1/\omega$ gain and a phase of $-90°$. At large frequencies the behavior of the PI controller converges towards that of a P controller, i.e. constant gain and a phase of 0°. The transition between both regimes occurs at the corner frequency $\omega_c = K_I/K_P$, as indicated in Fig. 5.11.

In case that the constant gain at high frequencies is undesirable, a low-pass filter with the transfer function shown in (5.7) and Fig. 5.2 can be added after the PI controller in order to damp the high frequency response. In this case the respective transfer functions multiply.

5.8.4 Time Discrete Implementation of a PI Controller

Up to now we have considered a continuous system output signal and a continuously acting controller. However, nowadays large parts of the feedback loop are implemented digital. The system output is sampled digital and the controller is implemented digital as well. The sampling occurs periodically with a sampling time t_{sample} and also the controller output signal is calculated with the corresponding frequency. This time discrete implementation of the controller corresponds to an additional part of the controller with a corresponding transfer function, which has to be multiplied with the transfer function of e.g. the PI controller. The transfer function corresponding to the time discrete implementation maintains the amplitude of the signal (this is not changed by the sampling), while it results in a phase shift of the signal due to the sampling, as shown in the following. If the output signal of the controller y is calculated from the error signal in one time step, this corresponds to a time delay of the output signal of t_{sample} relative to the input signal. This time delay corresponds according to (2.7) to a phase shift of $\phi(\omega) = -\omega t_{\text{sample}}$. This phase shift can be neglected for low frequencies $\omega \ll 2\pi/t_{\text{sample}}$. The phase shift (absolute value) increases linearly with the frequency ω and becomes sizable at high frequencies approaching $1/t_{\text{sample}}$. If the phase shift becomes 180° the sign of the signal inverts and the negative feedback will turn into a positive feedback and this can lead to an instability of the feedback loop. In order to prevent this, signals of this frequency have to be suppressed, e.g. by a low-pass added to the controller.

The pseudocode of a time discrete implementation of a PI controller is given in the following.

```
start
read measured_signal
error_signal = set_point - measured_signal
integral = integral + error_signal * dt
controller_output = K_P * error_signal + K_I * integral
goto start
```

Such a digital algorithm of the feedback controller can be used to analyze the behavior of a controller with a spreadsheet. One way to analyze the performance of controllers is analyze their step response. Step response means that the setpoint w (or alternatively the disturbing signal) is changed instantaneously from e.g. zero to one and the reaction of the controller and the whole system to reach the new setpoint is monitored.

As a first example, we use the simplest controller, a P controller characterized by its gain K_{P} and the simplest possible system characterized by a proportional gain $K_{\text{system}} = 1$. Additionally, we include a low-pass to the system with a time constant of $20 \times t_{\text{sample}}$. The step response of this simple feedback system is shown in Fig. 5.12a for two different values of K_{P}. The steady state error can be clearly seen and corresponds to the steady state error calculated from (5.33). With increasing gain K_{P} the steady state error becomes smaller and the time to reach the steady state becomes shorter. However, beyond a certain maximum value of the gain K_{P} the feedback loop becomes unstable, i.e. monotonously increasing oscillations result.

The behavior of an I controller in the time discrete case is shown in Fig. 5.12b for two different values of the gain K_{I} and a system consisting again of a proportional gain of one and a low-pass behavior. It can be seen that for an I controller the set point value is reached in the steady state. Furthermore, for larger values of the integrator constant K_{I} a faster reaction of the controller is observed. Figure 5.12c shows the step response of a time discrete PI controller for two different sets of K_{P} and K_{I}. In the PI controller the advantages of both types of controller are combined: The setpoint is reached and the response time is reasonably fast.

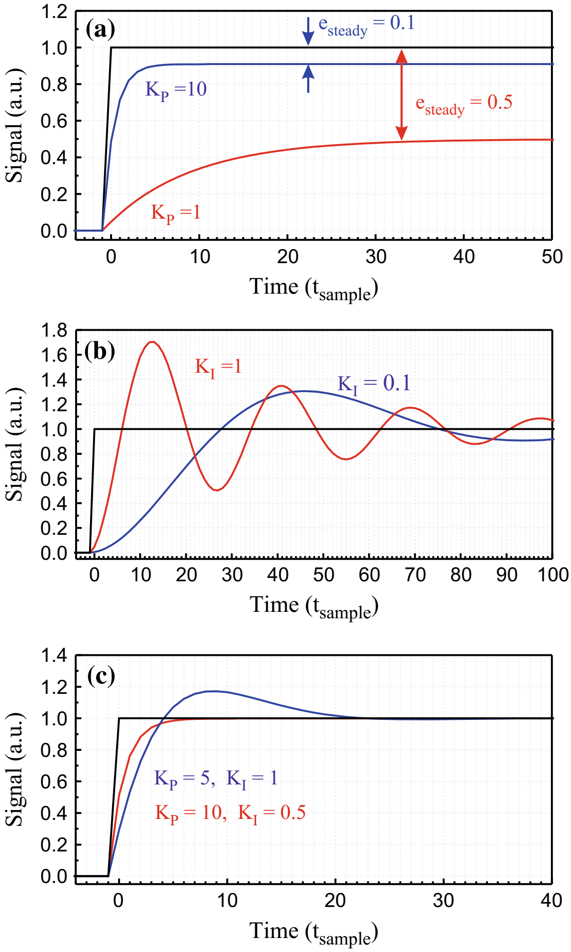


Fig. 5.12 **a** Step response of a P controller and a system with a proportional gain of unity and a low-pass behavior of $20 \times t_{\text{sample}}$. The setpoint signal with a step at time $t = 0$ is shown in black and the resulting system output signal $x(t)$ induced by the controller in blue and red. Due to the steady state error inherent to the P controller, the system output does not reach the setpoint value of one, but remains at a lower value. The steady state error calculated from (5.33) corresponds to the values shown in the graph. **b** Step response of an I controller for two different constants of K_I. An overshooting of the system output signal is observed in both cases and the speed of the controller increases for the higher value of K_I. **c** Step response of a PI controller for two different sets of constants. The PI controller combines the advantages of the P and the I controller: vanishing steady state error and fast response time. Note that the time scales in **a**–**c** are different

5.8.5 Instabilities of a Feedback Loop

A feedback system is considered as stable if a bounded input signal results into a bounded output signal for all times. A statement about the stability of the *closed-loop* feedback system can be obtained from the Bode plot of the *open-loop* system, i.e. the feedback path in Fig. 5.9 is not closed. The Bode stability criterion can be formulated as follows [7]:

> A *closed-loop* system is stable if the corresponding *open-loop* system is stable and the frequency response of the open-loop transfer function (Bode plot) has an amplitude ratio of less than unity at all frequencies corresponding to $\phi = -180° - n \cdot 360°$, where $n = 0, 1, 2, \ldots, \infty$.

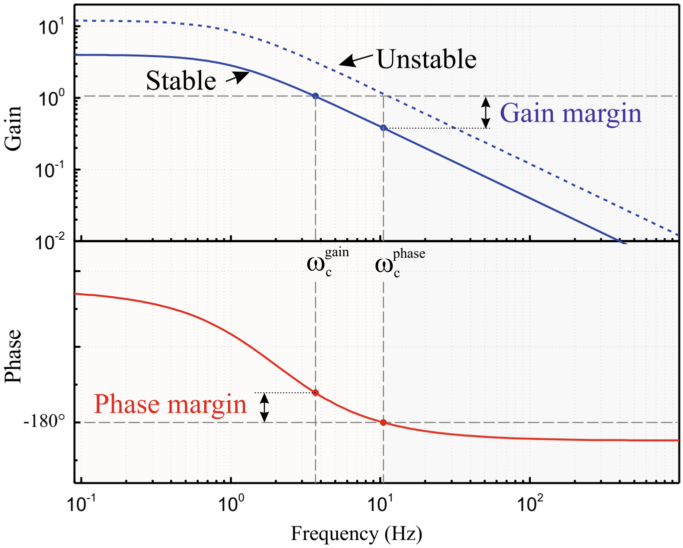


Fig. 5.13 Open loop transfer function of a feedback system. The corresponding closed loop system is stable if the gain remains smaller than one at the phase crossover frequency ω_c^{phase}

Since this criterion is a bit hard to understand we explain it in the following for some examples. In many cases and the ones we consider in the following only the $n = 0$ case is relevant. The Bode plot of an open loop system consists of two graphs (a) gain or amplitude ratio and (b) phase as function of the frequency for a sinusoidal input signal. We discuss first the Bode plot shown as solid lines in Fig. 5.13 (with a hypothetical monotonously decreasing frequency behavior). The phase crossover frequency ω_c^{phase} is defined as the frequency at which the phase has the value of $-180°$. If the gain of the open loop system at this frequency is smaller than one, as it is the case for the transfer function shown as solid line in Fig. 5.13, the closed loop system is stable. Moreover, the gain margin describes how much the gain of the open-loop transfer function can be increased before the system becomes unstable, as indicated by the dashed line in Fig. 5.13. The gain crossover frequency ω_c^{gain} is defined as the frequency at which the gain has the value of one. The phase margin (usually defined as positive value) is how much the phase has to be decreased to become $-180°$, as shown in Fig. 5.13. This gain and phase margins indicate how far the feedback loop is from the transition to an unstable behavior.

An example for an unstable closed loop feedback system is shown as dashed curve in Fig. 5.13. In this example the gain curve is shifted up to higher gains, while it is assumed that the phase behavior remains the same (solid line phase curve). In this case the open loop transfer function has a gain of one at the phase crossover frequency ω_c^{phase} and the closed loop feedback system will be unstable according to the above Bode stability criterion. Strictly speaking the above Bode stability criterion is a sufficient criterion, but not a necessary condition for stability [7].

5.8.6 Measurement of Transfer Functions

We have seen that transfer functions, complementary to the analysis of a step response, are important to characterize a feedback system. The transfer function of the controller e.g. a PI controller can be calculated as (5.38). For the transfer function of the system simple models can be used as we have done it before, e.g. a P system with a low-pass. However, more realistically the transfer function of an AFM can be much more complicated than the one of a simple model system.

The open loop transfer function of a system can be measured as follows. First the feedback system is brought to operation at a desired working point. For the case of an AFM this means that the tip is engaged to the sample and the sensor signal (e.g. the oscillation amplitude) has reached its desired setpoint value. Then the controller is stopped at this working point and a sinusoidal modulation signal with a certain frequency $V_{\mathrm{in}}(\omega)$ is applied (added) to the input of the system, as shown in Fig. 5.14. If the system is linear (as we assume throughout this treatment) the resulting output signal $V_{\mathrm{out}}(\omega)$ of the system is also a sinusoidal oscillation at the frequency ω, however, with a different amplitude and phase as the input signal. The output signal of the system is measured and analyzed with respect to its amplitude (gain) and phase. The resulting amplitude ratio $V_{\mathrm{out}}(\omega)/V_{\mathrm{in}}(\omega)$ and the phase of the output signal relative to the input signal correspond to two points of the Bode plot at the frequency ω. After the values of the Bode plot have been obtained at one particular frequency, the feedback is enabled again in order to compensate for a drift from the desired working point. In an AFM system, tip and sample may have drifted to another position and the system output may be somewhat different from the setpoint value. After the setpoint value is restored by the feedback, a new measurement at a slightly higher frequency of $V_{\mathrm{in}}(\omega)$ is performed. In this way the Bode plot and thus the transfer function of the system can be measured in a desired frequency range.

The controller can be also included into the system. In this case the modulation voltage $V'_{\mathrm{in}}(\omega)$ is applied to the input of the controller (e.g. as a modulation of the setpoint value as shown in Fig. 5.14). This corresponds to the measurement of the open loop transfer function of the feedback system, if the system output signal x is not fed back to the input of the controller.

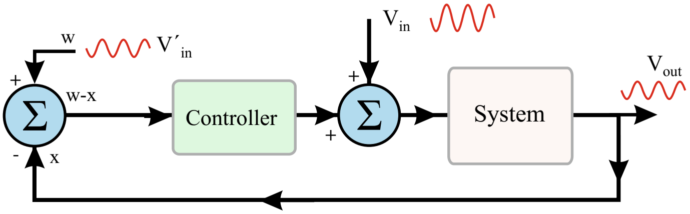


Fig. 5.14 Scheme of the measurement of transfer functions. The open loop system transfer function is measured by modulation of the system input signal with the controller halted. The transfer function of the closed loop feedback system is measured by modulating the set point

When measuring system transfer functions, it should be noted that a system transfer function might not be constant as function of time and various outer conditions. Different kinds of system variations can happen. For instance for the case of an AFM the system transfer function can change if the tip form changes (tip switch), or the transfer function close to a step edge can be different from the one on a free terrace, or the transfer function on different materials can be different. However, other parts of the system transfer function like the transfer function of the high voltage amplifiers driving the piezo elements, or the amplifier measuring the signal amplitude will stay constant.

The transfer function of the closed loop feedback system can be measured by modulation of the setpoint value with a sinusoidal signal $V_{\mathrm{in}}^{'}(\omega)$, while the controller is in operation and the feedback loop is closed (Fig. 5.14). The Bode plot is obtained by analyzing amplitude and phase of the output signal $V_{\mathrm{out}}(\omega)$ relative to the input signal $V_{\mathrm{in}}^{'}(\omega)$.

5.9 Feedback Controller in AFM

Up to now we have discussed feedback controllers from a very general perspective. Now we will apply these concepts to the case of an AFM system. In AFM the elements in the above-mentioned feedback loop have the following correspondence (Fig. 5.15).

- The setpoint w corresponds to a voltage representing the desired AFM sensor signal x, e.g. deflection, amplitude, or frequency shift.
- The (digital) controller calculates the error signal and determines the system input (control variable) y, which is in AFM the z-voltage controlling the tip-sample distance using the z-piezo element.
- The most complex part of the feedback loop is the system itself. In the case of AFM, it consists of DA converters, converting the digital value of the control variable y to an analog voltage, the high-voltage amplifiers (HVA) for the z-piezo voltage, the z-piezo element for the vertical positioning of the sample (or tip), and the tip-sample interaction, as well as the measurement of the AFM sensor signal, which depends on the AFM detection mode. In the static mode it is the cantilever deflection, in the dynamic AM mode it can be the oscillation amplitude, or in the FM mode the frequency shift of the sensor resonance frequency. Any of these signals is converted to a corresponding voltage by an amplifier. This voltage is called the AFM sensor signal. Finally, the sensor signal is converted by AD converters and corresponds to the system output x which is fed to the controller.
- Various kinds of noise arise due to external mechanical vibrations, the noise of the amplifiers, the sensor thermal noise, and the noise of DA and AD converters.
- Also the topography of the sample corresponds to a disturbance changing the tip-sample distance and thus the sensor signal.

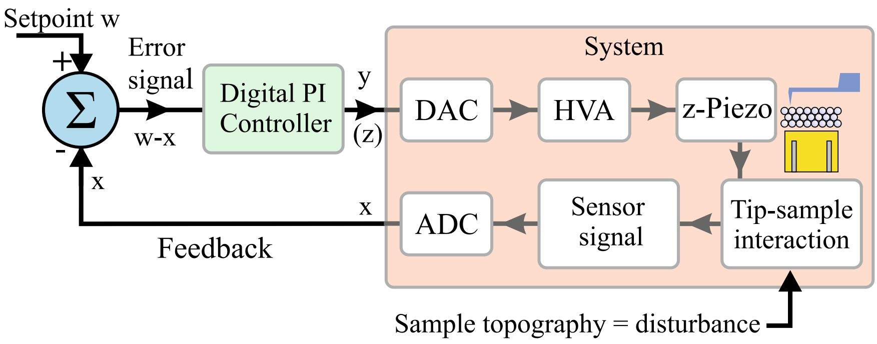


Fig. 5.15 Model of an AFM feedback loop with a digital (time discrete) PI controller and the system consisting of a digital to analog converter (DAC), a high voltage amplifier (HVA), *z*-piezo elements, the tip-sample system, detection of the sensor signal and a analog to digital converter (ADC)

In AFM the surface topography corresponds to a disturbance of the system and the controller compensates for this disturbance. A step in the topography of the sample can be emulated in the control system by a step-like change of the setpoint value for the sensor voltage (step response). Since the disturbance due to the topography can have quite high values (several steps, scanning slope), this leads according to (5.33) to a large steady state error if only a P controller would be used. Due to this, in AFM the I controller is the most important controller, as this controller leads not to a steady state error, even for large values of the topography signal. The I controller provides a constant output in response to a vanishing error signal (no further contributions to the integral due to vanishing integrand). This is useful to maintain a new height level past a step edge in the topography.

In AFM, the P-part of the controller regulates fast deviations from the setpoint such as small/atomic corrugations. Moreover, the I-controller has the advantage that it is less prone to noise. Depending on the conditions, the measured sensor signal can be quite noisy. While the P-controller reacts immediately to a noise spike of the measured signal, an I controller acts as a low-pass averaging out noise spikes.

A differential controller is rarely used in AFM, as this type of controller is most prone to noise and the implementation of another controller adds a further dimension in the space of control parameters to be optimized.

The gain constants of the PI controller are optimized in AFM operation as follows. Starting from an initial working point with default (conservative) gains and with the sample approached to the tip, the integrator gain K_I is increased until, ringing (slight oscillations) of the controller output signal z is observed. Then K_I is decreased until the controller output signal becomes stable again. Then the proportional gain K_P is increased until ringing occurs and decreased until a stable condition is reached. This procedure can be repeated in order to do fine tuning of the parameters. Subsequently, scan is started and the parameters can be further optimized according to the obtained

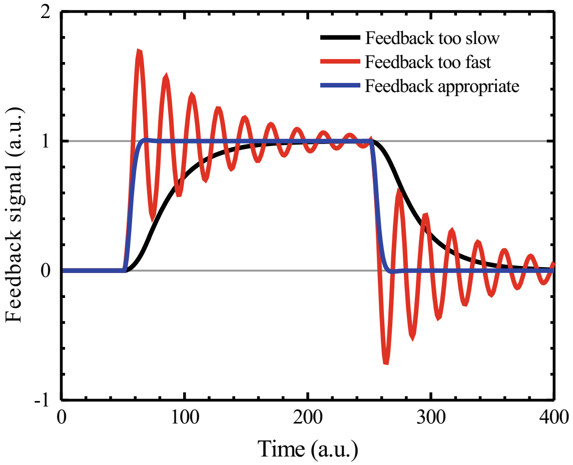


Fig. 5.16 Schematic (exaggerated) step response of the AFM feedback signal to a modulation of the setpoint with a square wave. For too slow feedback settings (*black line*), too fast feedback settings (*red line*), and appropriate feedback settings (*blue line*)

image. For instance a sharp feature in the topography should be imaged as a sharp feature without ringing.

Alternatively to real scanning of a surface topography, the setpoint signal w can be changed in order to emulate a topography signal and observe the reaction of the controller to this signal. Modulating the setpoint with a square signal corresponds to the analysis of the step response of the feedback system. In Fig. 5.16 the setpoint changes from zero to one at time = 50 and back to zero at time = 250. The reaction of the AFM feedback signal y or z to this is shown schematically (exaggerated) for too slow feedback settings (black line), too fast feedback settings (red line), and appropriate feedback settings (blue line). A scan in the reverse direction will show the opposite signatures. If the feedback parameters are not optimized this can lead to artifacts in the acquired images. If the feedback is too slow this will lead to blurred images; if the feedback is too fast this may lead to a feedback overshoot when the tip encounters sudden height changes in the topography of the sample, as discussed also in Chap. 8.

In AFM, there is an effect which exerts a high load to the feedback controller. Usually the sample is not oriented perfectly parallel to the xy-directions given by the piezoelectric scanner. This slope (scanning slope) is usually the largest height signal in the original AFM data and will be removed by appropriate background subtraction in the final image, as discussed in Chap. 7. However, the feedback has to follow this (scanning) slope. As a quantitative example, if the xy-plane of the scanner and the sample surface are 3° off relative to the sample surface, this slope corresponds to a height of 500 Å for a 1 μm wide scan. This is usually by far the largest topography height signal in an image. Also here the I controller can follow a constant slope better than the P controller. In principle another controller type can be added which particularly compensates for this slope disturbance signal [8]. However, since this adds another parameter and makes thus the parameter tuning of the controller more difficult, this is rarely done. Another way to cope with the scanning slope is to set the system input (voltage to the z-piezo) already to the value it had at this position in the previous scan line. In this case the PI controller has only to compensate for the changes in the topography relative to the previous scan line. As the scanning slope usually remains constant from scan line to scan line this puts less load to compensate for on the PI controller.

5.10 Implementation of an AFM Feedback Controller

Feedback controllers are realized via a digital feedback loop nowadays. The sensor signal is measured by an amplifier and then the corresponding voltage is digitized by analog-to-digital converters (ADC), as shown in Fig. 5.17. These converters can have, for instance, an accuracy of 20 bit in a range of ±10 V corresponding to a step width of 20 μV, which is usually far below the noise in the system and therefore sufficient for all practical purposes.

The actual feedback loop is often realized by a digital signal processor (DSP) or with field programmable gate arrays (FPGA) (Fig. 5.17). A DSP is a own computer on which a single user single-task real-time program runs. From the measured (digitized) sensor signal and the sensor signal setpoint, the output, i.e. the actuator voltage for the z-piezo motion, is calculated using a digitized version of a PI controller. Using a digital feedback loop has several advantages. First, it is very easy to stop the feedback and to perform spectroscopic measurements, and also to measure the transfer function. Another advantage is that the feedback mode can be changed just by changing the software. The controller algorithm can be changed by a few lines

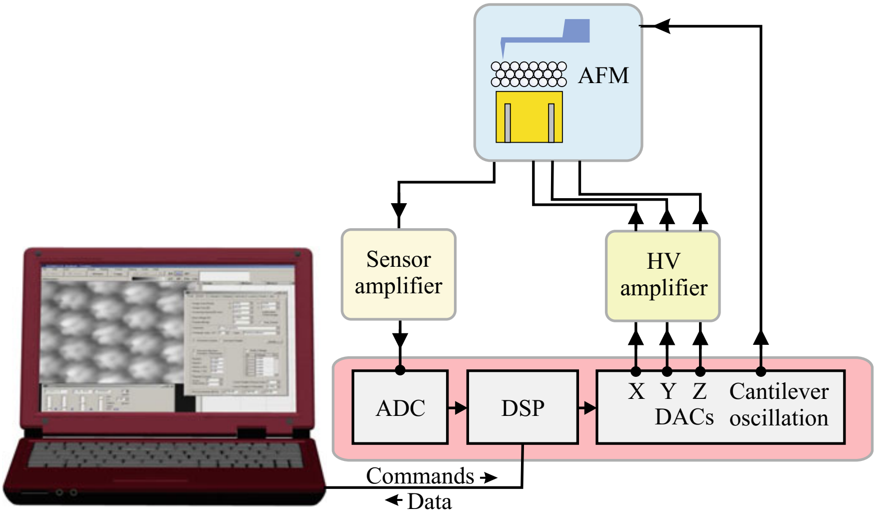


Fig. 5.17 Implementation of computer controlled AFM electronics

in the DSP program. Furthermore, non-linear algorithms for noise reduction can be implemented.

Once the controller output (new z-voltage) is calculated, this number is converted into an actual voltage by (for instance) 20 bit digital analog converters (DAC). This z-voltage (range: ±10 V) is then amplified by a high-voltage amplifier to a range of e.g. ±200 V (Fig. 5.17). This is enough to reach the necessary amplitude of the piezo actuators of a few micrometers. Regarding the resolution, the following reasoning can be applied: For a piezo constant of 60 Å/V and a high-voltage amplifier gain of 20 one DAC unit converts to a z-distance of 2 pm, which is usually more than enough. If a higher resolution is required, the gain of the high-voltage amplifier can be reduced. This means that with the high resolution DA and AD converters available today the digitization of the input and output quantities is no longer a problem since it is far below the usual noise limits. For the dynamic AFM modes also the oscillation voltage can be supplied from the computer via a DAC to piezo driving the oscillation of the sensor.

When scanning an AFM image, the DSP sends the xy-scan data to the DAC. The voltages for the x- and y-electrodes are finally amplified by the high-voltage amplifiers. The data about the height of the tip above the surface, i.e. the voltage applied to the z-piezo, generated by the feedback algorithm running on the DSP, is sent to the PC. The measurement program takes the height of the AFM tip above the surface and displays it as an image, i.e. in gray scale as a function of x and y.

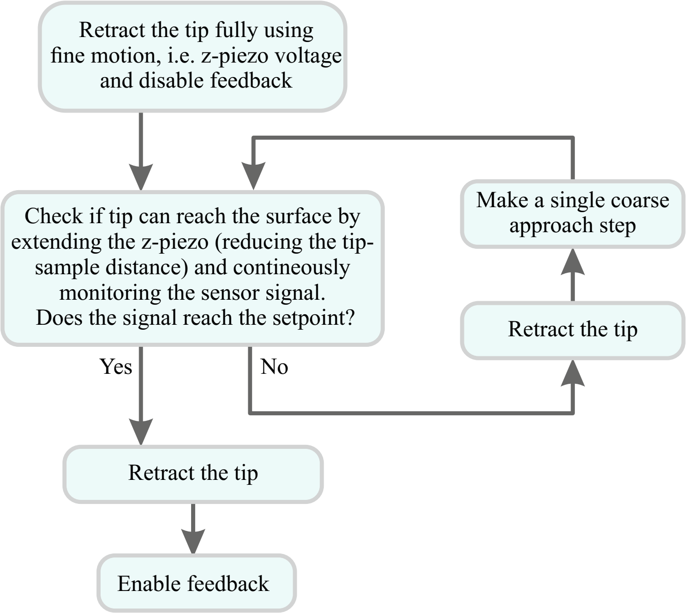

Retract the tip fully using fine motion, i.e. z-piezo voltage and disable feedback
Check if tip can reach the surface by extending the z-piezo (reducing the tip-sample distance) and contineously monitoring the sensor signal. Does the signal reach the setpoint?
Make a single coarse approach step
Retract the tip
Yes
No
Retract the tip
Enable feedback

Fig. 5.18 Flow chart of the automatic approach procedure used in atomic force microscopy

The digital control of the AFM also allows an automated procedure to be used during the coarse approach of the tip towards the sample. This procedure consists of alternating steps of fine z-approach and a subsequent coarse approach step, if a tip-sample contact is not reached within the z-fine position range. In this procedure the coarse positioning step has to be (for safety) smaller than the z-fine positioning range. A flow chart for an automated control could be as shown in (Fig. 5.18). After the automatic coarse approach a desired setpoint for the sensor signal is chosen and scanning can be started.

5.11 Digital-to-Analog Converter

In a computer controlled data acquisition and control system, analog data have to be read to the computer and digital data generated by the computer have to be converted to analog signals. For instance, in atomic force microscopy the xy-scan signals are generated by a computer program (digital values) and have to be converted to analog signal driving the piezo elements. For this task a digital-to-analog converter (DAC) is used. Here we describe the principle of how such a device can operate. However, actual digital-to-analog converters are more sophisticated than the basic idea explained here.

We assume that the digital signal is already present as voltages (high/low) at several wires of a connector. As an example, we will consider a four-bit signal in the following. In Fig. 5.19, the digital signal is represented by switches either open or closed (−5 V). Each of the lines (switches) has a different weight from 2^0 to 2^3 corresponding to the weight of the bit in the binary digital code. If all switches are open this corresponds to zero (0000), if all wires are connected to −5 V this

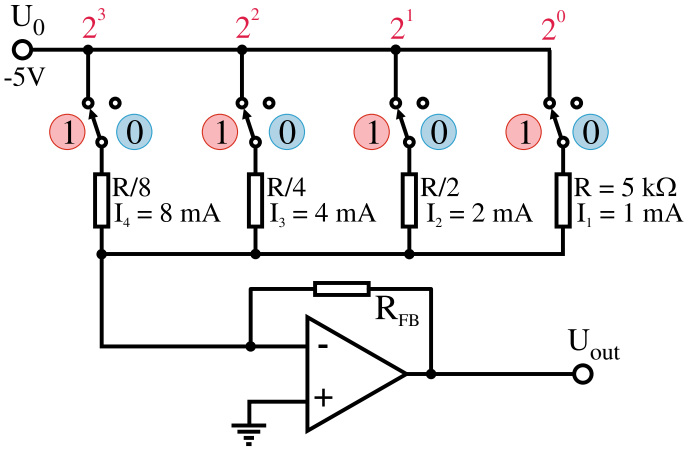


Fig. 5.19 Operating principle of a digital-to-analog converter

corresponds to (binary 1111, i.e. 15). The task is now to convert the digitally coded voltage values present at the four connectors to 16 analog voltages relative to ground, ranging, for example, from 0 to 10 V. The resistor following each switch is chosen such that the current through it (when flowing to ground) corresponds to the weight of that bit. The least significant bit (2^0) has, for instance, a 5 kΩ resistor, corresponding to a current of 1 mA to ground, while the most significant bit (2^3) has an 8 times smaller resistor corresponding to an 8 times higher current of 8 mA in this line. All the lines are routed to the inverting input of an operational amplifier acting as a transimpedance amplifier. Since the positive input of the operational amplifier is on ground, the negative input is the virtual ground, as we have considered before. At the point where all these lines are brought together the sum of all the currents flows through R_{FB}. According to (5.24), the analog output voltage at the operational amplifier is

$$U_{\mathrm{out}} = -R_{\mathrm{FB}} U_0 \sum_{\text{i=all closed switches}} \frac{1}{R_{\mathrm{i}}}. \tag{5.40}$$

The maximum output voltage can be chosen using a proper value for R_{FB}.

5.12 Analog-to-Digital Converter

In atomic force microscopy, the analog voltage corresponding to the sensor signal has to be converted to a number (e.g. 16-bit value) proportional to the analog voltage (sensor signal). For this task, an analog-to-digital converter (ADC) is used. An ADC can be realized by the comparison of the analog signal (to be digitized) to a voltage from a digitally generated voltage ramp. The principle of operation of one simple ADC is shown in Fig. 5.20. A digital voltage ramp is generated and converted to an analog voltage ramp using a DAC. The value of the generated voltage ramp is compared to the analog input signal to be digitized using a comparator. This comparator has a low digital signal as long as the voltage ramp has a lower voltage than the input voltage. A comparator can be realized by an operational amplifier without external feedback network. Due to its large open-loop gain the output will always be maximally positive as long as the negative input voltage is smaller than the voltage at the positive input. The comparator signal changes to logically high if the voltage ramp exceeds the voltage to be measured (Fig. 5.20). This end of conversion signal is then fed to the ramp controller in order to stop the ramp and to read the actual (digital) ramp value. With this digital value of the ramp, a digital value of the analog input signal is saved and the conversion is stopped. Instead of ramping up all digital values from zero, also some interval-based algorithm can be also used in order to find the value closest to the analog input.

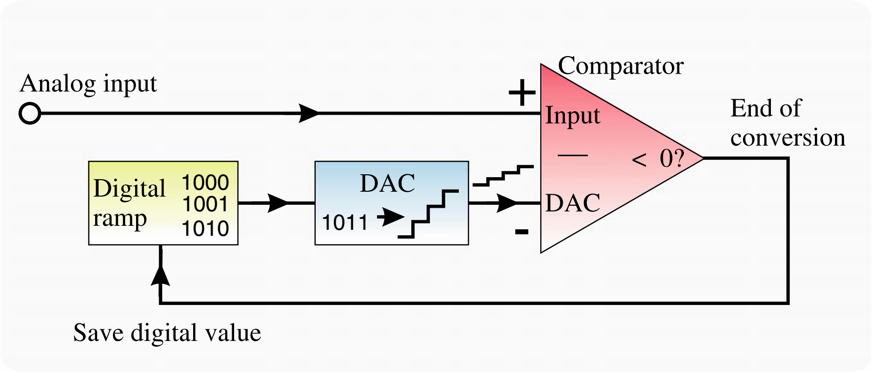


Fig. 5.20 Operating principle of an analog-to-digital converter

5.13 High-Voltage Amplifier

High-voltage amplifiers are needed to drive the piezo elements since the voltages supplied by the digital-to-analog converters are usually only in the range up to ±10 V and are not high enough to generate sufficient extensions of the piezo elements of several micrometers. Therefore, the DAC voltages are amplified up to about 200 V, which generates the required piezo extensions. We assume here again piezo tubes as piezo elements. Much higher voltages are not advisable because they can lead to a depolarization of the piezo material. A reasonable upper limit for the required bandwidth of the high-voltage amplifiers is the resonance frequency of the piezo element. You cannot move a piezo element at a frequency higher than its resonance frequency. Therefore, 50 kHz is an upper limit for the required bandwidth. In practice, the AFM feedback loop often has a much lower bandwidth in the range between 1 and 10 kHz. In this case, a low-pass filter at the output of the high-voltage amplifier can be used to reduce the noise. The output noise of the high-voltage amplifiers should be less than 1 mV. With a typical z-piezo constant of about 50 Å/V, this corresponds to a noise in the extension of the piezo in the z-direction of 0.05 Å, i.e. 5 pm.

The piezo motions during scanning are relatively slow. In order to move inertial sliders (Sect. 4.1.1), saw-tooth signals are applied to the piezo elements and the steepest possible slope of the piezo motion is required. This means a high slew rate (voltage change per time) of the high-voltage amplifier is required. The achievable slew rate depends on the capacitive load at the output of the amplifier, i.e. the capacity of the piezo elements. A high piezo capacity means that a lot of charge has to be pumped to or from the piezo element. If this has to be done in a short time, a high current has to flow. Therefore, high-voltage amplifiers driving piezo elements with a high capacity have to supply a high current in order to achieve a high slew rate. This can lead to problems of high power dissipation in the leads. This problem with the high capacitance occurs mostly for monolithic stacks of piezo elements. They have capacitances in the μF range, while piezo tubes, for instance, have only capacitances in the nF range.

5.14 Summary

- Operational amplifiers are characterized by a very large input resistance, a very low output resistance and a very large open-loop gain.
- The actual gain of an operational amplifier including a feedback network is determined by the characteristics of the feedback network, not by the operational amplifier.
- Two golden rules can be applied when analyzing an op-amp circuit: (i) The input current vanishes. (ii) The voltage difference between the inputs is zero.
- A current amplifier converting the input current to an output voltage can be built using an operational amplifier. The output voltage depends on the feedback resistance as $V_{\text{out}} = -I_{\text{in}} R_{\text{FB}}$.
- In the proportional controller, the actuating variable is proportional to the error signal. In the integral controller the actuating variable is proportional to the time integral over to the error signal.
- The transfer function, output signal divided by the input signal (including amplitude and phase), is used to characterize the frequency response of electronic components.

References

1. F. Reif, *Fundamentals of Statistical and Thermal Physics* (Waveland Press Inc., Long Grove, 1965). ISBN: 1577666127
2. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd edn. (Cambridge University Press, Cambridge, 2007). ISBN: 9780521880688
3. J. Johnson, Thermal agitation of electricity in conductors. Phys. Rev. **32**, 97 (1928). https://doi.org/10.1103/PhysRev.32.97
4. H. Nyquist, Thermal agitation of electric charge in conductors. Phys. Rev. **32**, 110 (1928). https://doi.org/10.1103/PhysRev.32.110
5. R.G. Lyons, *Understanding Digital Signal Processing*, 3rd edn. (Pearson Education, London, 2011). ISBN: 8131764362
6. R.E. Best, *Phase Locked Loops*, 6th edn. (Mc Graw Hill, New York, 2007). ISBN: 0071493751
7. J. Hahn, T. Edison, T.F. Edgar, A note on stability analysis using Bode plots. Chem. Eng. Educ. **35**, 208 (2001)
8. D.Y. Abramovitch, S. Hoen, R. Workman, in *American Control Conference on Semi-automatic Tuning of PID Gains for Atomic Force Microscopes* (2008), p. 2684. https://doi.org/10.1109/ACC.2008.4586898

Chapter 6
Lock-in Technique

A lock-in amplifier measures a signal amplitude hidden in a noisy environment. An AC modulation is used to measure the signal in a very narrow frequency range. Using the lock-in technique the noise can be even much larger than the signal which can nevertheless be measured precisely. In dynamic atomic force microscopy is used for instance to detect the oscillation amplitude.

6.1 Lock-in Amplifier–Principle of Operation

In order to see what the task is for a lock-in amplifier Fig. 6.1, shows an AC signal with different levels of noise superimposed. The original signal is shown in red and an increasing amount of noise amplitude is added to the signal from Fig. 6.1a, b. It may seem hopeless to try and recover the original signal amplitude in Fig. 6.1b, which is buried by a large noise signal. Two important requirements are needed for the lock-in technique to accomplish this task. First, the frequency of the AC signal has to be known and, second, the phase of the signal has to be stable.

In order to explain how a lock-in amplifier works, we look to the product of two harmonic signals. The following mathematical identity holds for the product of two harmonic functions at two different frequencies

$$\begin{aligned} &A\cos(\omega_1 t+\phi)\times B\cos(\omega_2 t)\\ &\quad=\frac{1}{2}AB\left\{\cos\left[(\omega_1+\omega_2)t+\phi\right]+\cos\left[(\omega_1-\omega_2)t+\phi\right]\right\}, \end{aligned} \tag{6.1}$$

where A and B are the amplitudes of both harmonic functions, respectively and ω_1 and ω_2 are the corresponding angular frequencies, respectively and ϕ a phase difference.

We now discuss the result for two cases. If $\omega_1=\omega_2$ the first cos term results in a harmonic signal (AC component) with frequency $\omega_1+\omega_2=2\omega_1$. The cos

B. Voigtländer, *Atomic Force Microscopy*, NanoScience and Technology, https://doi.org/10.1007/978-3-030-13654-3_6

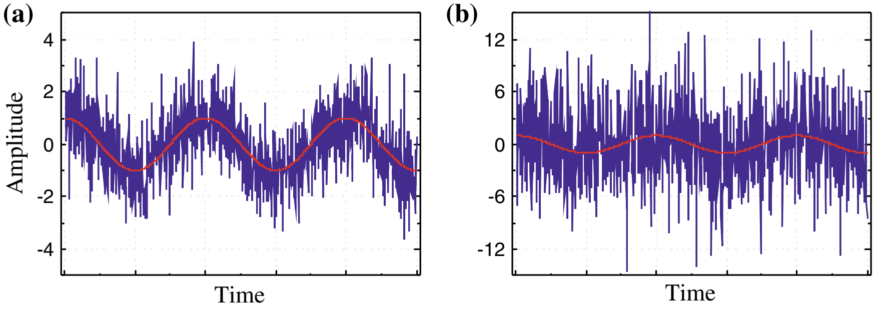


Fig. 6.1 Sinusoidal AC signal (*red*) and sinusoidal signal plus noise (*blue*). The noise increases from **a** to **b**. The amplitude of the harmonic signal is always one

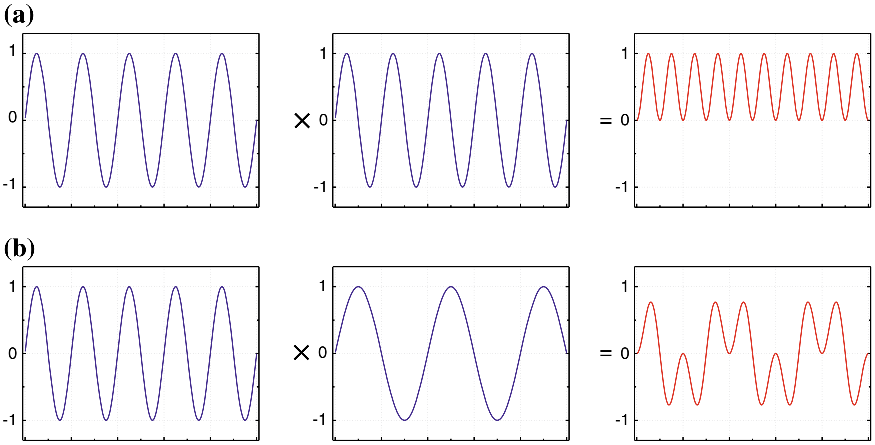


Fig. 6.2 **a** Product of two phase-coherent harmonic functions with identical frequency $\omega_1 = \omega_2$ results in a DC component plus a harmonic component. **b** Product of two phase-coherent harmonic functions with different frequencies $\omega_1 \neq \omega_2$ results in a harmonic signal without DC component

term containing the frequency difference results in a DC component of the value $\frac{1}{2}AB\cos\phi$. The sum of both terms (AC component and DC component), corresponding to the product of the two harmonic functions, is also visualized in Fig. 6.2a. Thus, the product of two harmonic signals of the same frequency results in a DC component plus a harmonic signal.

If $\omega_1 \neq \omega_2$ the product of the two harmonic signals can be written as the sum of two harmonic signals oscillating with the sum and the difference of ω_1 and ω_2. In this case, the product signal has no DC component, as shown in Fig. 6.2b.

In the next step of the lock-in detection, the DC component of the product signal is extracted by time averaging or low-pass filtering of the product signal as

$$\lim_{T\to\infty} \frac{1}{T} \int\limits_0^T A\cos(\omega_1 t + \phi) \times B\cos(\omega_2 t)\,\mathrm{d}t = \begin{cases} \frac{1}{2}AB\cos\phi & \omega_1 = \omega_2 \\ 0 & \omega_1 \neq \omega_2 \end{cases} \tag{6.2}$$

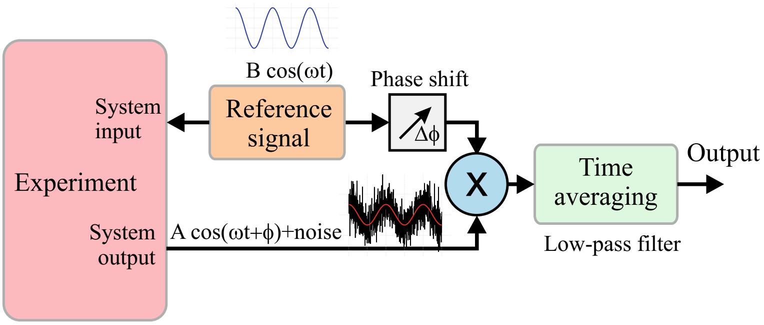


Fig. 6.3 Schematic of a lock-in amplifier consisting of a reference oscillator which modulates (via the experimental setup) the output signal of the system. This signal serves as input for the lock-in amplifier and is multiplied by the reference signal and then low-pass filtered. Due to this, only the frequency component close to the modulation frequency survives and all noise components at other frequencies are suppressed by this modulation technique

For the case $\omega_1 \neq \omega_2$ the signal is a harmonic signal without DC component. Therefore, the averaging results in the signal vanishing completely. For the case $\omega_1 = \omega_2$ the time averaging filters out just the DC component of the product signal $\frac{1}{2}AB\cos\phi$, which is proportional to the signal A that we want to measure. Additionally, the result is proportional to the phase difference between the input signal and the reference signal. Due to this, the lock-in technique is also called phase-sensitive detection.

The phase sensitive detection can also be used to suppress parasitic signals with a fixed phase relation to the desired signal. Let us assume as an example the measurement of a (resistive) AC current signal. The signal wire will form a capacitor with the surrounding ground, leading to a parasitic capacitive current. Since this parasitic signal component has a phase difference of 90° (relative to the desired resistive signal), the capacitive signal can be suppressed by adjusting the phase appropriately according to (6.2).

In conclusion: by time averaging, all (noise) frequency components with $\omega_1 \neq \omega_2$ are filtered out and only the frequency component at the reference frequency ω_2 survives with an amplitude proportional to the signal to be measured. The noise frequency components (for instance 50/60 Hz line frequencies) are filtered out by the lock-in amplifier. A schematic diagram of a lock-in amplifier is shown in Fig. 6.3. In the first stage of a lock-in amplifier, the input signal of amplitude A (which is the signal amplitude to be measured modulated by the reference signal at frequency ω plus a lot of noise) is multiplied by the reference signal (of known amplitude B). In a second stage the time averaging filters out the high-frequency component.

While the lock-in amplifier is *very* effective in noise reduction, noise components with a frequency close to the reference frequency result in low frequency contributions in the product signal at a frequency $(\omega_1 - \omega_2)$. Long integration times of about $\tau \approx 2\pi/(\omega_1 - \omega_2)$ are required in order to average these low frequency components

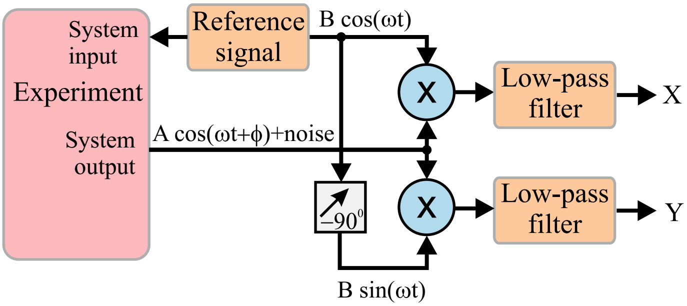


Fig. 6.4 Schematic of a two channel lock-in amplifier. Measuring X and Y and subsequently applying some arithmetic calculations leads to the simultaneous determination of the absolute value of the amplitude and the phase

out. The reference frequency of the lock-in amplifier is usually chosen in a frequency range where the noise signal has the smallest spectral density. These considerations apply for coherent noise. Noise components with an unstable phase $\phi_{\text{noise}} \neq \text{const.}$ average out even if they are at the reference frequency.

Also a DC offset added by the experimental apparatus to the measurement signal is suppressed by lock-in detection. If this constant signal component is multiplied by the reference signal a harmonic signal oscillating around zero results, which is averaged out by the time averaging.

If the measured signal has a phase shift ϕ relative to the reference signal induced by the experiment, the output of the lock-in amplifier is also proportional to $\cos\phi$. This phase shift can be compensated by a corresponding phase shift of the reference signal in the lock-in amplifier, as shown in Fig. 6.3. The phase shift is optimized in order to obtain a maximal output signal amplitude (or better vanishing output and subsequently applying a phase shift of $\pm\pi/2$.

The absolute values of the amplitude and the phase can also be measured simultaneously. A scheme for performing such a measurement is shown in Fig. 6.4. In one channel the usual measurement is performed (channel X), while in the second channel phase of the reference signal is shifted additionally by $-90°$ (channel Y). The term for the reference signal in channel Y becomes $B\cos(\omega t - 90°) = B\sin(\omega t)$. If we neglect the constant factor $1/2\,B$ this results for the channel X in $X = A\cos\phi$, as discussed before for the single channel lock in amplifier. For the channel Y, we apply the reasoning for the lock-in amplifier as in (6.1), however, using the identity for sin times cos, results (after time averaging) in the signal $Y = A\cos(\phi - \pi/2) = A\sin\phi$. Expanding this to complex variables $\tilde{X} = Ae^{i\phi}$ and $\tilde{Y} = Ae^{i\phi - \pi/2}$ as shown in Fig. 6.5 helps to calculate amplitude and phase. The absolute value of the amplitude A and the phase shift ϕ can be determined from the measured values X and Y as $A = \sqrt{X^2 + Y^2}$ and $\phi = \arctan(Y/X)$. In digital lock-in amplifiers, the measured values X and Y are available as numbers and the computation can be performed arithmetically.

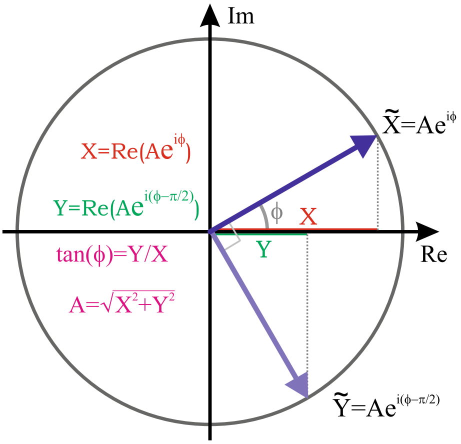

Im
Re
$\tilde{X}=Ae^{i\phi}$
$\tilde{Y}=Ae^{i(\phi-\pi/2)}$
$X=\mathrm{Re}(Ae^{i\phi})$
$Y=\mathrm{Re}(Ae^{i(\phi-\pi/2)})$
$\tan(\phi)=Y/X$
$A=\sqrt{X^2+Y^2}$
ϕ
X
Y

Fig. 6.5 Simultaneous determination of the amplitude A and the phase shift ϕ of the signal by a measurement with an additional phase shift of 90°, using a two-channel lock-in amplifier

A lock-in amplifier is used for the measurement of small AC signals with virtually arbitrary noise reduction (determined by the integration time), provided that the AC signal is coherent (stable phase) and the frequency is known.

Up to now, we have considered the measurement of amplitude and phase of an AC signal. However, the lock-in technique can also be used if the signal to be measured is a DC signal. In this case the DC signal is converted to an AC signal by modulation, i.e. multiplied by an AC reference signal to obtain a phase stable AC signal of a known frequency.

It could be assumed that a DC signal can be measured with high precision using long averaging times, i.e. low bandwidth without lock-in technique. However, at DC a particular type of noise, the $1/f$-noise (occurring in many electronic devices) which becomes very large at small frequencies impedes high precision DC measurements. The modulation of a DC signal with an AC reference signal transfers the DC amplitude to an AC amplitude, avoiding the $1/f$-noise problem. In the frequency domain the signal is transferred from DC, where it is prone to $1/f$-noise to a higher frequency AC signal at which the lock-in detection technique can be applied.

6.2 Summary

- The lock-in technique is an AC modulation technique used to detect small AC signals hidden in a noisy environment.
- Multiplication of the measurement signal by the reference signal results in a DC component proportional to the amplitude of the measured signal at the modulation frequency. For all other frequency components of the measurement signal, multiplication by the reference signal results in an AC component, which is averaged out by time averaging.

Chapter 7
Data Representation and Image Processing

Scanning probe microscopy data usually have the form of a matrix where the topography (height) or some other signal such as the phase in dynamic AFM is measured as a function of the lateral xy-position on the surface. Data representation is the task to map the heights (i.e. the output of the z-controller) to gray levels in an image in an optimal way. Image processing is used in order to enhance the image representation further, i.e. by removing image artifacts such as high-frequency noise, noise pixels or noise lines [1, 2].

7.1 Data Representation

A data representation using 8-bit or 256 gray levels (ranging from 0 (black) to 255 (white)) is more than sufficient, since the human eye can distinguish only less than one hundred gray levels. These data are displayed as an image of typically 512 × 512 pixels.

The original data on the height of the tip (z-output signal of the digital feedback loop) are usually set by digital-to-analog converters (DAC) with a certain resolution. In the following, we consider 16-bit converters as an example (≈65,000 levels), while nowadays 24-bit DACs are available. The task for data representation is now to efficiently map the data, which cover a certain range of the 65,000 levels (DAC units), to the 265 gray levels. This task is also called background subtraction. As an example, we will discuss this first for one scan line. However, the same strategies apply for a whole image. As a convention for the gray levels black is assigned to the lowest height and white to the highest. If one were to map the 16-bit data range linearly from the lowest level to the highest level to the 8-bit gray scale from black to white not much of the surface structure would be visible. One scan line usually covers only a small range of the 65,000 levels. As an example, the scan line shown in Fig. 7.1 contains a range of about 800 height levels (DAC units). If the 256 gray levels were

B. Voigtländer, *Atomic Force Microscopy*, NanoScience and Technology, https://doi.org/10.1007/978-3-030-13654-3_7

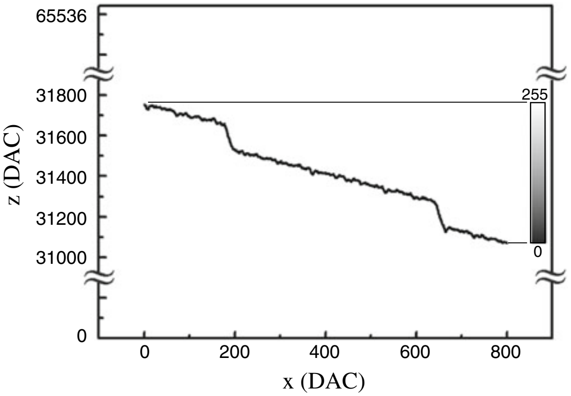


Fig. 7.1 For a good data representation the 256 *gray levels* have to be mapped to the 65,000 DAC levels in a proper way

mapped to the complete range of 65,000 digital-to-analog converter (DAC) levels, (level 0 is black and level 65,000 is white) a range of 65,000/256 = 256 height levels would be mapped to one gray level. It is clear that most of the information contained in the original data is lost by this poor mapping. For our scan line in Fig. 7.1, the 800 height levels in which the image information is contained would be mapped to only 3 gray levels ($800/256 \approx 3$). Therefore, the gray scale should be mapped to a smaller range of the 65,000 digital-to-analog converter (DAC) levels which contain the (height) data of the scan line, as shown in Fig. 7.1.

Another effect is that the actual topographic data are often hidden due to the quite large slope of a scan line. This slope arises because the scanning plane is usually tilted slightly with respect to the sample. This tilt occurs due to an imperfect alignment of the sample relative to the coordinate system of the scanning piezo element. In the following, we term this the *scanning slope*, which can be as large as several degrees. This scanning slope shows up as a tilted base line in the data as shown in Fig. 7.1. Usually, and specifically in atomically resolved images, the range of the real height values on the surface is very small (only a few Å), and the range of the measured height data is dominated by the scanning slope. Here we give two quantitative examples in which we consider a relatively large tilt angle between surface and scanner of 3°. If we consider an image of the size of 1 μm the height difference induced by this slope across the image is $\Delta h = \Delta x \tan \alpha \approx 500\,$Å. This 500 Å on an image size of 1 μm corresponds to a scanning slope which will be present in all images. If we consider, on the other hand, that we have as the real image signal, for instance, 5 atomic steps, each of 3 Å height, the image signal we want to measure (15 Å) resides on a scanning slope of 500 Å. This means that the background height change due to the slope is 30 times larger than the image signal (the steps). In a second example, we take an atomically resolved image of a size of 500 Å, corresponding to a height difference due to the scanning slope of 26 Å. If the atomic corrugation on a single atomic terrace is 1 Å the (atomic corrugation) signal to (scanning slope) background ratio is 1/26 in this case.

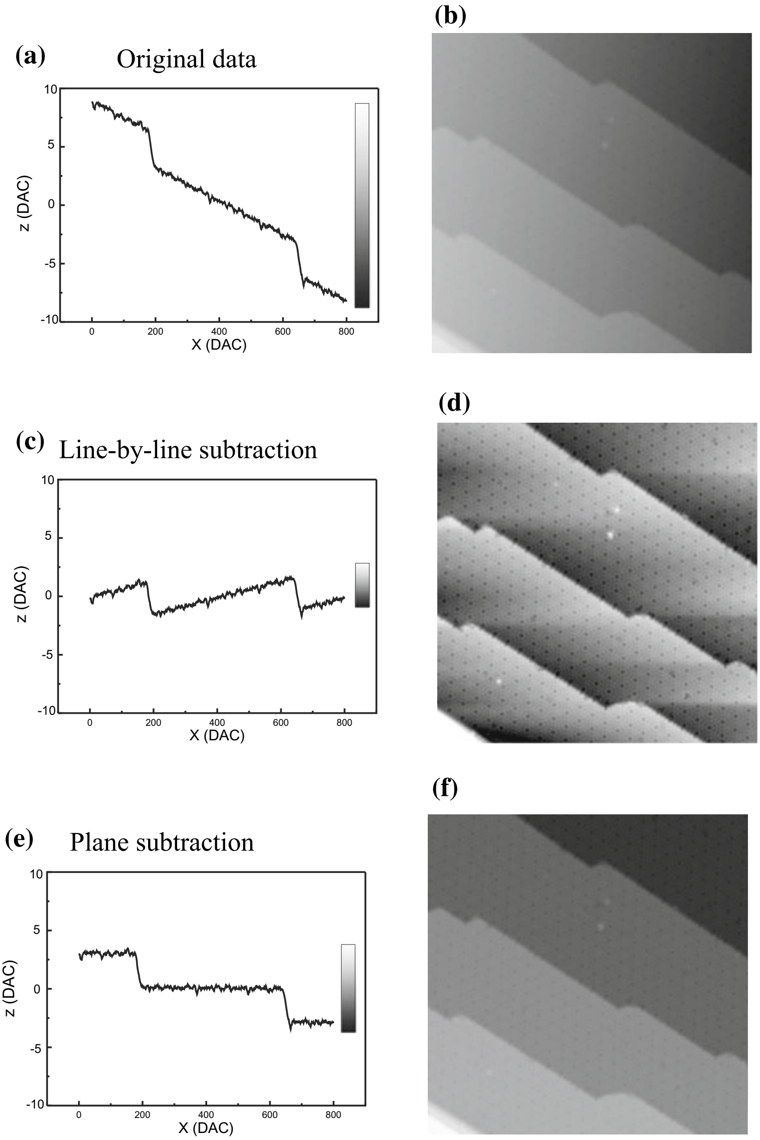


Fig. 7.2 STM data taken on a stepped Si(111) surface with the atomically resolved (7 × 7) reconstruction contained in the data. Comparison of different kinds of background subtraction for a single scan line (*left panel*) and a whole image (*right panel*). **a** and **b** Show the original data without background subtraction. In **c** and **d** a line-by-line background subtraction was applied. In **e** and **f** a plane subtraction relative to one of the terraces, between steps of a single atom height, was applied. The image size is 600 Å. In this image, the scanning slope in the x-direction corresponds to an angle of 0.7° between sample and scanner

We have seen that even a small tilt between sample and scanner leads to a substantial slope in the images. This slope can be eliminated by a background subtraction. This is usually done by fitting a straight line to the data of each scan line and by displaying only the deviations of the data with respect to this fit, as shown in Fig. 7.2c, d. This background subtraction increases the contrast in the image, but also leads to artifacts like the black shadows (i.e. one terrace has no uniform gray level) which arises due to some higher parts of the scan line which pull the fitted line up. The next higher approximation is to use a fit to a quadratic function as background. This can also remove the part of the background that arises from the scanner bow (and non-linearities of the piezo elements) in large scans. The scanner bow arises because the xy-motion induced by the tube scanner is approximately a motion on a sphere with a radius of the piezo tube length.

Another kind of background subtraction is not taking each line individually into account, but the whole matrix of measured data as one entity. Here the obvious approaches are to fit a plane or square function (paraboloid) to the data for background subtraction. Another approach is that the user can define points in an image which are known to belong to one specific height (for instance one atomic terrace). The background subtraction is then performed relative to this user-defined plane. An example of this background subtraction relative to a user defined plane is shown in Fig. 7.2e, f. The different methods of background subtraction each have their advantages and disadvantages. The advantage of the (user-defined) plane subtraction is that locations of the same height on the surface are displayed by the same gray level. The advantage of line-by-line subtraction is that the contrast is higher and the small height corrugations due to the atomic structure of the Si atoms are more easily visible. As another variant the whole contrast range from black to white can be used for one atomic terrace, leaving however all lower terraces black and all higher ones white. This is also called clipping. If you see larger areas in an image either white or black, the real data are outside the contrast range and are clipped to black or white. A helpful tool to see the distribution of the gray levels contained in an image over the 256 available gray levels is an histogram of the gray levels in an image. Such a histogram of the gray levels shows if the gray levels are evenly distributed or if some gray levels are (almost) not occupied in the respective image.

Apart from the gray scale images considered so far, it is, of course, also possible to use color in the image representation. In the false color representation, the 8-bit gray scale palette is replaced by a color palette. The most popular one is the fire palette ranging from black via red and yellow to white. In Fig. 7.3a a gray scale representation (subtracted line-by-line) of a stepped Si(7×7) surface is used, while in Fig. 7.3b a false color representation with the fire palette is used. In Fig. 7.3c a plane subtracted representation of the same image is shown in gray scale and false color representation using a palette with several colors is shown in Fig. 7.3d. Here the palette was chosen such that each terrace has a specific color. In Fig. 7.3e a 3D image representation of the same image is shown. Here techniques like rendering and ray tracing are used to give a plastic impression of an actual three dimensional landscape of the measured data. While such images look like the real morphology of a landscape it must be kept in mind that the z-scale in SPM images is almost always

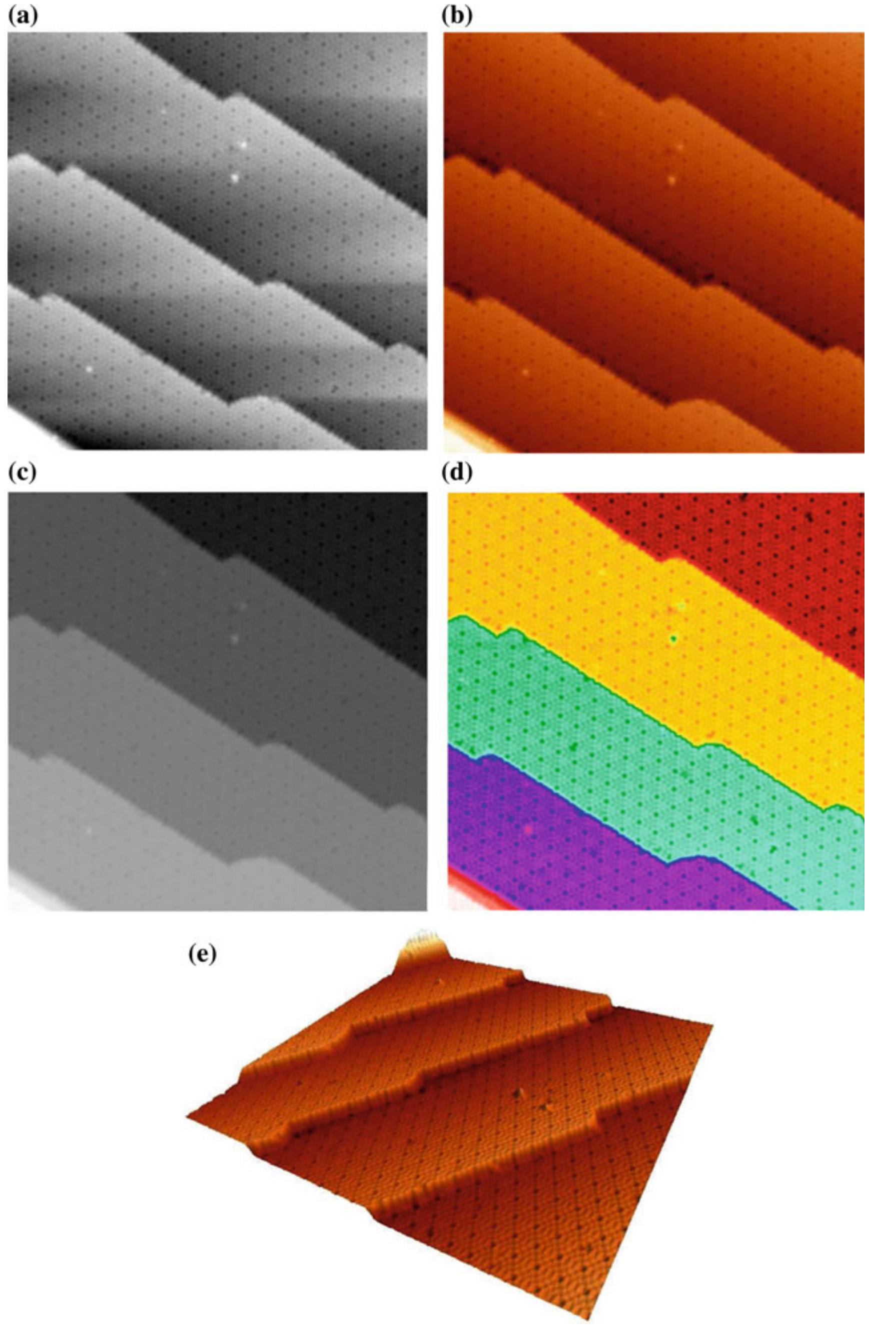


Fig. 7.3 STM image of a Si(111)-7 × 7 surface shown in different representations. Line-by-line background subtraction using **a** a *gray scale* palette and **b** a color palette. Plane background subtraction on one terrace **c** with *gray scale* palette and **d** a color palette with different colors for each atomic terrace. **e** Three dimensional representation of the same image

quite exaggerated relative to the lateral scale. For the example in Fig. 7.3e, the z-scale in the image is only 12 Å, while the image size is 600 Å. Going one step further a fly-by movie through the atomic or nano canyons at the surface can be generated. With all these different kinds of image representations it should not be forgotten that they are only different representations of the same initial data matrix. The appropriate image representation should always be chosen for the respective purpose. An elaborated image representation with a lot of colors may be well suited to impress laypeople but may obscure the visibility of important details. Therefore, a simple gray scale representation is often sufficient to convey the scientific information.

7.2 Image Processing

The application of image processing filters has two purposes. First, to enhance the image representation contrast above that possible with simple background subtraction and, second, to remove image artifacts such as high-frequency noise, noise pixels or noise lines. These are often eliminated by simple matrix filters. These filters consist of a sum of products of nearby pixel values with elements of a weighting matrix.

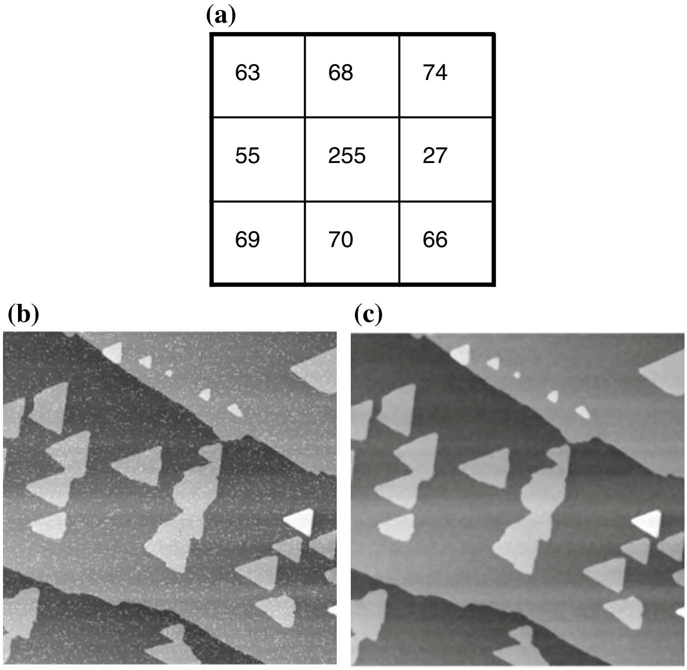
Matrix or convolution filters are used (a) to remove noise from the images, (b) to sharpen (high-pass), or (c) to smoothen (low pass) the images. The following algorithm describes the 3 × 3 convolution of image pixels. The measured value of an image pixel in the image matrix $z(x, y)$ is replaced by a modified value $z'(x, y)$

$$z'(x, y) = \frac{\sum_{i=x-1}^{x+1} \sum_{j=y-1}^{y+1} W_{(i-x+2, j-y+2)} z(i, j)}{\sum_{i=1}^{3} \sum_{j=1}^{3} |W(i, j)|}. \tag{7.1}$$

Depending on the properties of the matrix W high-pass, low-pass and other kinds of filters can be realized [3].

Another very simple and effective filter is the median filter. It removes speckle noise in the images, i.e. pixels which have, a very different gray value than the neighboring pixels. The advantage of this filter is that it does not lead to a pronounced blurring of sharp edges in the image, as other averaging filters do. For a median-filtered pixel consider the 8 pixels surrounding one pixel plus the center (original) pixel (9 pixels) and take as the new (gray) value for the center pixel the median of these nine pixels. The median is not the mean of the 9 pixels but the 5th highest value (i.e. the middle value, which is 68 in the example in Fig. 7.4a). The same procedure is applied to all pixels in the image. Median filtering is robust with respect to outlier pixels which would influence the mean considerably but not the median. In Fig. 7.4b, an image with white noise pixels is shown and Fig. 7.4c shows the image after median filtering.

Another frequently applied method for filtering SPM images is Fourier filtering. However, this kind of filtering is often not very useful for "improving" images. From



63	68	74
55	255	27
69	70	66

Fig. 7.4 **a** Example of the median filter showing gray values in a matrix of 8 pixels around a center pixel. When applying the median filter, the value of the center pixel is replaced by the fifth highest value (68 in the example). Thus, the outlier value of 255 is replaced by the more reasonable value of 68. **b** STM image of triangular Si islands on Si(111) with speckle noise. **c** After median filtering this noise is removed

the 2D Fourier transform of an image some parts considered to be noise are cut out and a reverse transformation is performed. With this procedure the image information in the respective frequency range is removed also. The emphasis in Fourier filtering is on enhancing the periodic part of the image, while in SPM often the defects and deviations from a periodic ideal lattice are interesting. Strong Fourier filtering can highlight the periodic part so strongly that atoms are "produced" by Fourier filtering and defect sites are "filled" by atoms.

One useful application of Fourier analysis for SPM images is the identification of a long-range periodic corrugation signal in the image which may be hidden by noise in the original image. Another application of a Fourier transform is to compare quantitatively two different periodicities which are present in one image, for instance the atomic lattice and an additional periodic long-range modulation, as for instance a Moiré pattern.

It is important to mention in detail in presentations and publications which kind of image processing algorithms have been applied to the original data.

7.3 Data Analysis

There are a whole range of image analysis procedures which are often very specific to the problem under study. For instance, if in studies of epitaxial growth, island populations are analyzed, questions arise like: What is the island density per area? Also other questions about the distribution of the volume, the width, or the height of islands can be answered using AFM data. In principle, all questions related to the morphology of the surface can be answered, since the complete surface morphology is measured. Such analysis tasks can be performed more or less automatically. However, such data analysis procedures are very specific to the problem considered and we will not discuss them further here.

A more general example of data analysis is the measurement of the roughness of a surface. The complex 3D information contained in a topographic image of a sample is condensed in a single number. The usual quantity characterizing the roughness of a surface is the RMS roughness defined as the standard deviation of the heights $h(x, y)$

$$\sigma = \sqrt{\left\langle (h(x, y) - \overline{h})^2 \right\rangle} = \sqrt{\frac{1}{N} \sum\nolimits_{i=1}^{N} (h(x, y) - \overline{h})^2}, \tag{7.2}$$

with N being the total number of pixels in the image, and $\bar{h}$ the average height. Another quantity describing the roughness is R_a which is defined as

$$R_a = \frac{1}{N} \sum\nolimits_{i=1}^{N} \left| h(x, y) - \overline{h} \right|. \tag{7.3}$$

A necessary requirement for a correct determination of the roughness is a good background subtraction of the scanning slope. Further, a roughness present on length scales larger than the image considered for the evaluation of the roughness does not enter in quantities for the roughness considered above. On the other end towards roughness at the smallest length scales, any roughness on length scales smaller than the radius of the AFM tip is not captured properly (cf. Fig. 8.2).

A simple and general procedure for data analysis is the line scan. By mouse clicking on the image, a line is defined in an image on the computer screen and the height levels along this line (sometimes averaged over a certain width perpendicular to this line) are displayed and can be used for high-accuracy measurements of topographic heights as shown in Fig. 7.5, or horizontal spacings of features. Also the slopes of facets of surface features such as islands can be determined. Note that in AFM line scans like the one in Fig. 7.5b the vertical scale displayed is much smaller than the horizontal scale, leading to apparent facet angles much larger than the real ones.

In the following we discuss a flaw that can occur when facet angles of an island are to be determined with a high precision using a line scan. This flaw occurs if a scanning slope is present, i.e. an angle α between the reference plane of the sample and the xy-plane of the AFM scanner. We will show in the following that the facet angle β of an island like the one shown for example in Fig. 7.6a are measured by a

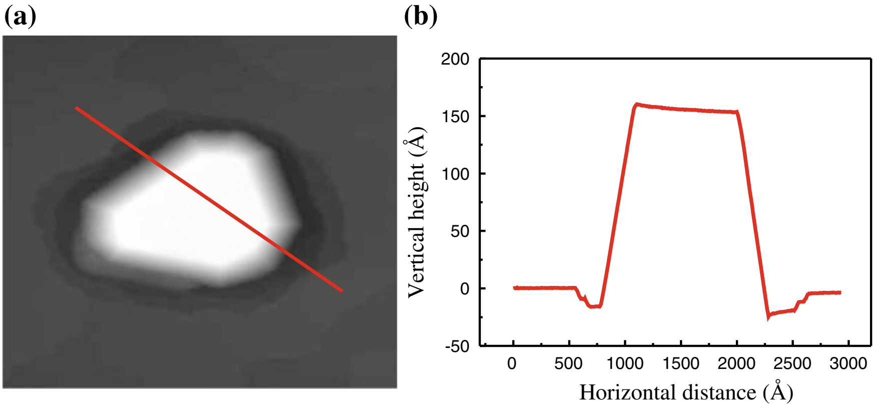


Fig. 7.5 **a** Gray scale image of a 3D Ge island. **b** Line scan across this island

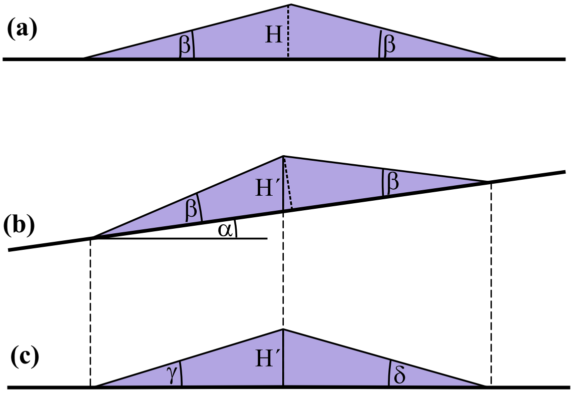


Fig. 7.6 The measured facet angles of an island depend (after a linear background subtraction) on the scanning slope α. **a** If there is no scanning slope ($\alpha = 0$), the correct facet angle β is measured. **b** Raw data obtained by an AFM, if a scanning slope is present. **c** Result of a line scan across the island after a linear background subtraction of the scanning slope. In this case measured facet angles γ and δ different from the real facet angle β result

line scan, after background subtraction of the scanning slope, as facet angles γ and δ instead of the correct angle β.

In Fig. 7.6a an island with two facets with angle β relative to the sample reference plane (x-direction) outside of the island is shown. If the x-direction of the sample is not inclined relative to the x-direction of the AFM scanner, the line scan of the AFM image measures the correct facet angle β of the island. However, usually the sample xy-plane is inclined up to several degrees relative to the xy-scanning plane, which corresponds to the scanning slope introduced in Sect. 7.1. This case is shown in Fig. 7.6b in which the sample x-direction is inclined by the angle α relative to the

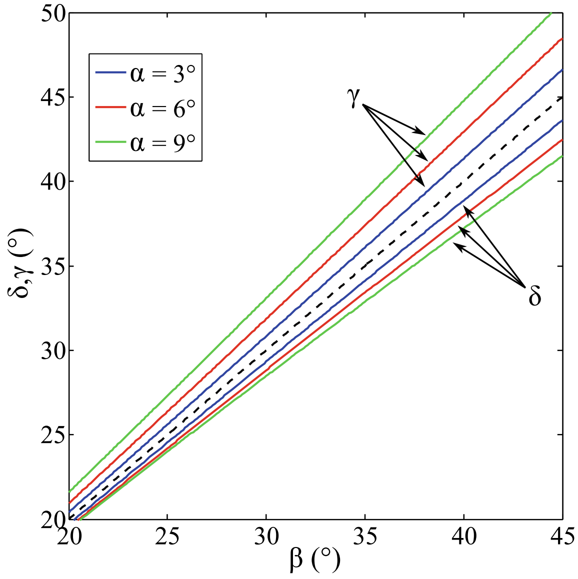


Fig. 7.7 **a** Facet angles γ and δ according to (7.4) after a linear background subtraction. The lines represent three different angles of the scanning slope α as function of the facet angle β. For larger values of α and β the facet angles measured after a linear background subtraction deviate from the real facet angle β. For $\alpha = 0$, γ and δ are equal to β

AFM scanner x-direction. The topography shown in Fig. 7.6b corresponds to the raw data measured by an AFM.

If now a background subtraction is made in a way that scanning slope is removed and the surface plane outside the island becomes horizontal, as shown in Fig. 7.6c, this leads to facet angles γ and δ differing from the correct facet angle β. Working out the trigonometric relations results in

$$\tan\gamma = \frac{\sin\beta}{\cos\alpha\cdot\cos(\alpha+\beta)}, \quad \text{and} \quad \tan\delta = \frac{\sin\beta}{\cos\alpha\cdot\cos(\alpha-\beta)}. \tag{7.4}$$

The facet angles γ and δ, are plotted in Fig. 7.7 for three different angles of the scanning slope α as function of the facet angle β. Without scanning slope $\gamma = \delta = \beta$ (straight line in Fig. 7.7). It can be seen that for larger values of α and β the measured facet angles after a linear background subtraction (γ and δ) deviate from the real facet angle β. In order to determine the actual facet angle β from the measured angles γ or δ, the equations (7.4) have to be inverted, resulting in

$$\beta = \arctan\left(\frac{\cos^2\alpha}{\cos\alpha\cdot\sin\alpha+\frac{1}{\tan\gamma}}\right), \quad \text{and} \quad \beta = \arctan\left(\frac{\cos^2\alpha}{-\cos\alpha\cdot\sin\alpha+\frac{1}{\tan\delta}}\right). \tag{7.5}$$

We have discussed here the measurement of the facet angle in the presence of a scanning slope α using a one-dimensional line scan, as this is the procedure applied in most cases. We have seen that the measured facet angles (after linear background subtraction) do not (exactly) correspond to the real facet angle. An alternative approach to determine the real facet angle is to take the measured raw data Fig. 7.6b and preform a *rotation* of the entire image (or the line scan) instead of a linear background subtraction. A *rotation* in order to remove the scanning slope preserves the facet angle β. The above considerations were performed for an ideal tip shape. Of course the tip shape (and other effects like drift and piezo creep) can additionally influence the measured facet angles.

7.4 Summary

- Data representation is the task to map the measured heights (DAC values) to gray levels in an image in an optimal way.
- Line-by-line background subtraction and plane background subtraction are commonly used.
- Matrix filters can be used to sharpen, or smooth the images, or to remove outlier pixels.
- In order to measure heights, width, or slopes of topographic features line scans can be used as data analysis tool.

References

1. P. Klapetek, *Quantitative Data Processing in Scanning Probe Microscopy*, 2nd edn. (Elsevier, Amsterdam, 2018). ISBN 9780128133477
2. P. Eaton, K. Batziou, Artifacts and practical issues in atomic force microscopy, in *Atomic force microscopy*, ed. by N. Santos, F. Carvalho. *Methods in Molecular Biology*, vol. 1886 (Humana Press, New York, 2019). ISBN 978-1-4939-8893-8. https://doi.org/10.1007/978-1-4939-8894-5_1
3. J.C. Russ, F.B. Neal, *The Image Processing Handbook*, 7th edn. (CRC Press, Boca Raton, 2017). ISBN 9781138747494

Chapter 8
Artifacts in AFM

The ideal AFM tip is (from the point of view of surface imaging) a sharp needle which can image even surface features with high aspect ratio. If the tip has a broader shape, artifacts occur due to a convolution of the tip shape with the surface features. Nearby micro tips can lead to a doubling of surface features in the acquired AFM image. Other kinds of artifacts in atomic force microscopy [1–3] include thermal drift, feedback overshoot, piezo creep, and electrical noise. While the images shown in this chapter are STM images, they show examples of generic effects of SPM artifacts which appear also in AFM images.

8.1 Tip-Related Artifacts

The geometrical shape of the tip will always influence the AFM images taken with it. The most common artifacts in atomic force microscopy occur due to tips which are not sharp (enough). Roughly topographic features present on the surface which have a larger aspect ratio than the tip are not imaged correctly. The acquired image is a convolution of the probing tip shape and the sample topography. Due to this effect, topographic features protruding from a flat surface are broadened. In extreme cases, if sharp asperities are present on the surface the tip shape is imaged by the surface asperities (tip image). The principle of how the tip shape influences the image of a sharp surface feature is shown in Fig. 8.1a. A sharp asperity on the surface is only imaged properly with an equally sharp (or sharper) tip.

An example of this is shown in Fig. 8.1b, where carbide clusters with a high aspect ratio are imaged on a Si surface. Each carbide cluster is imaged as a small high protrusion surrounded by a much larger "halo". All clusters appear with the same shape, which is the shape of the tip. In the image in Fig. 8.1c, we can see that the tip form changes during the image acquisition. In the upper part of the image the carbide clusters appear larger due to a blunt tip, while the tip changes to a somewhat sharper shape in the middle of the image. This occurred during a tip-sample contact. Traces of this are visible in the left part of the image. However, the tip shape is still not ideal

B. Voigtländer, *Atomic Force Microscopy*, NanoScience and Technology, https://doi.org/10.1007/978-3-030-13654-3_8

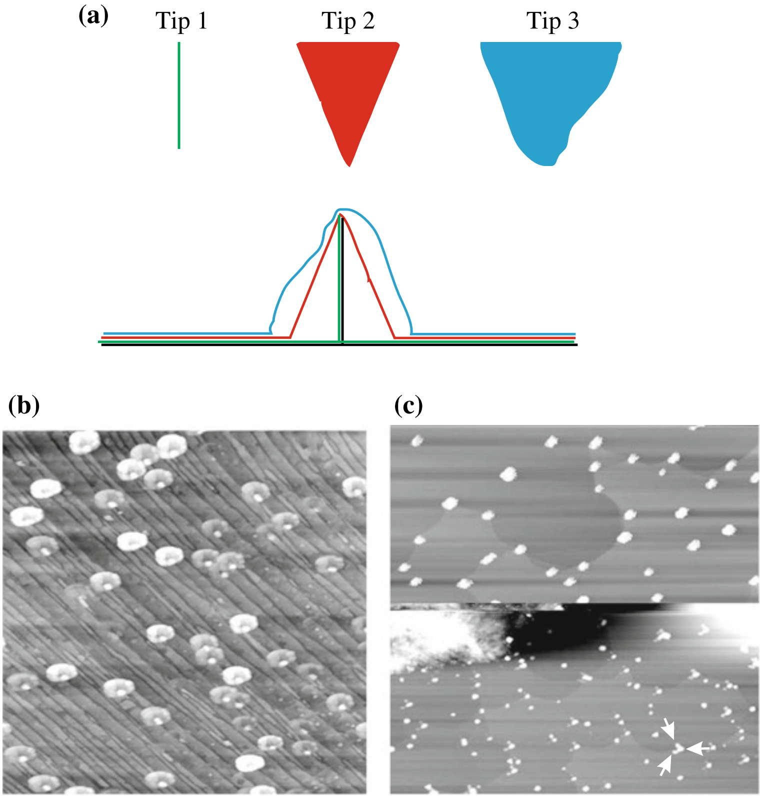


Fig. 8.1 **a** Sketch of the principle of how the tip shape influences the image of a sharp asperity present on the surface. **b** Example in which high aspect ratio carbide clusters are imaged by a blunt tip. All imaged clusters have a similar apparent shape: the tip shape. **c** Image of carbide clusters showing a change of the tip shape in the middle of the image

in the lower part of the image, as higher clusters are imaged as three protrusions, due to the tip shape, as indicated by arrows in Fig. 8.1c. AFM cantilevers have usually tips with a pyramidal shape. Thus, if in an image pyramidal facets appear, this can be due to the AFM tip shape. In some cases facets with the slope of the pyramidal tip occur rather than artifacts with complete pyramids [2].

Generally, (since delta function like tips do not exist) in images influenced by artifacts due to the tip shape the imaged structures have larger apparent lateral sizes than the real structures, if they are protrusions above a flat surface level (Fig. 8.1). If depressions below the average surface level are imaged with a blunt tip, they appear (due to the tip shape) correspondingly smaller in the AFM image. While the lateral width of surface structures is influenced by tip-related artifacts, the height is measured correctly, if the tip returns to the average surface height, as it is the case in Fig. 8.1.

In images influenced by the tip shape the imaged structures do not necessarily have all the same shape. This is only the case if the imaged structures have all the same shape (or are much smaller than the tip), as it is the case in Fig. 8.1. However, induced by a particular tip shape, the imaged structures often obey similarities due to the convolution with the tip shape, e.g. they appear elongated in one particular direction.

A test which can be made in order to verify if a particular shape occurring in a similar way for different imaged structures occurs due to the tip shape is to rotate the sample by 90°. If the images structures appear also to "rotate" by 90°, then the shapes of the imaged structures are a true property of the sample. If the imaged structures do not rotate, they are an artifact of the tip shape. To rotate the scan direction can not be used to distinguish between a tip-related artifact and a real topographic feature: In the corresponding AFM images the imaged features will show up as rotated in both cases, as the relative orientation between tip and surface does not change by scanning a rotated image.

As a rule of thumb, all topographic features which have a radius of curvature smaller than the radius of curvature of the scanning tip, are not imaged properly. Many attempts have been made to use a mathematical deconvolution to recover the real surface topography. However, such attempts are limited due to the following three reasons: (a) Even for a known tip shape a full recovery of the true topography by deconvolution is not completely possible at sharp trenches or close to sharp asperities, because there are "dead zones", i.e. parts of the surface topography which are never reached by the tip as shown schematically in Fig. 8.2. (b) Most importantly the tip shape is generally unknown and a "measurement" of the tip shape at sharp needle-like structures on the surface is often not really practicable. (c) The tip shape may change often. Therefore, any tedious measurement of the tip shape does not last for long. Probably not until deconvolution is attempted. Due to the occurrence of dead zones the height of a structure measured with an AFM is smaller or equal to the real height of this structure.

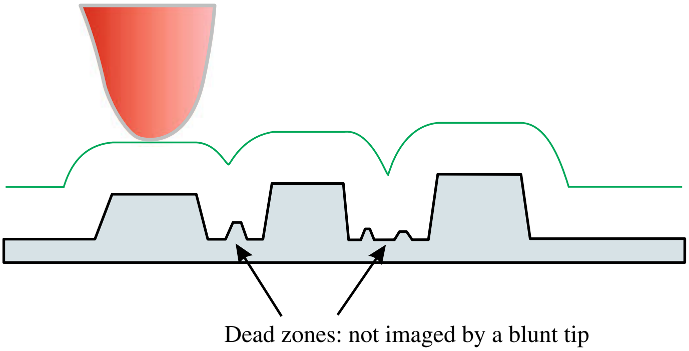


Fig. 8.2 Schematic showing the occurrence of "dead zones" due to the blunt shape of the tip

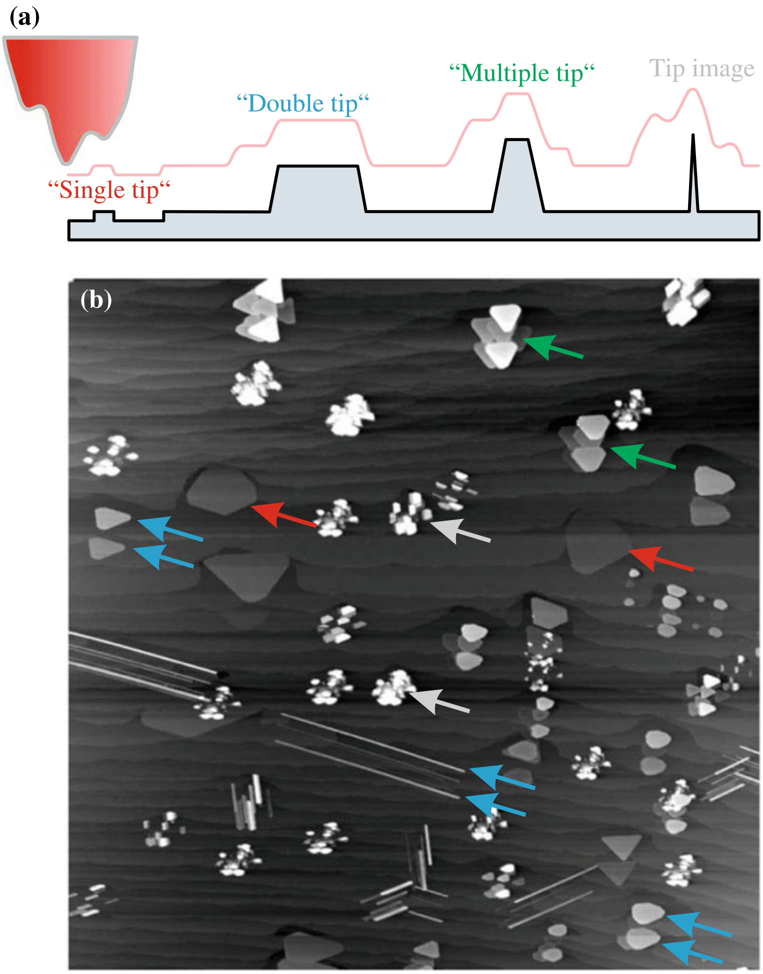


Fig. 8.3 **a** Sketch of a double (multiple) tip giving rise to doubled (multiple) imaging of surface features. The *light red line* shows the trace of the tip above the surface. **b** Example of silicide nano islands and nano wires imaged. The higher the structures which are imaged, the stronger is the tendency towards double (multiple images). For structures of one atomic height a single tip apex images (*red arrows*), somewhat higher structures are imaged by a double tip apex (*blue arrows*). Even higher structures are imaged by even more micro tips (*green arrows*). Narrow and high structures result in an image of the tip structure instead of the surface feature (*gray arrows*)

One particular case of a blunt tip is a double/multiple tip, as shown schematically in Fig. 8.3a. Such a double tip gives rise to double imaging of features on the surface as the islands and nanowires. These double images always occur at the same mutual distance and orientation in the images as indicated by blue arrows in Fig. 8.3b. Depending on the height of the imaged features, the tip acts as a single tip for features

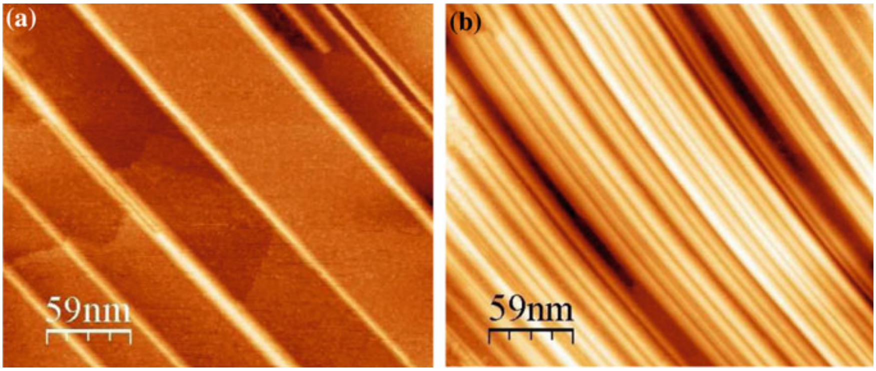


Fig. 8.4 Image of 5 Å yttrium deposited on Si(110). **a** Silicide nanowires imaged with a sharp tip. **b** The same surface imaged with a blunt tip leads to much higher apparent coverage due to multiple images of the silicide nanowires

of a single atomic height (indicated by red arrows in Fig. 8.3b), as a double tip for somewhat higher features (indicated by blue arrows in Fig. 8.3b), or as five or sixfold tip for even higher features (indicated by green arrows in Fig. 8.3b). Narrow and high structures present on the surface result in an image of the tip structure instead of the surface feature (gray arrows).

The images in Fig. 8.4 show that a (blunt) tip can give rise to a completely wrong estimate of the deposited coverage in thin film growth experiments. In Fig. 8.4a, a Si(110) surface is imaged on which 5 Å yttrium was deposited, which can be seen as elongated silicide wires on the surface. The *same* surface (however, not exactly the same area) was also imaged in Fig. 8.4b, with a different blunt tip. Here the silicide coverage *appears* to be much higher. This is not real, but an effect of a blunt tip where the silicide nanowires appear to be multiply imaged by several micotips forming the blunt tip.

The lesson from the previous considerations should not be that you cannot believe any AFM images, but rather you should always critically reflect on your AFM measurements and to reproduce measurements with different tips in order to exclude tip artifacts as carefully as possible.

How to Identify Tip-Related Artifacts

- If all (or many) features on the sample have the same shape, or all the features have an elongated shape in the same direction this is an indication of a blunt tip which is "imaged" by the surface. AFM cantilevers have usually tips with a pyramidal shape. Thus, if in an image pyramidal facets appear, this can be due to the AFM tip shape.
- If the sample is rotated (e.g. by 90°) and the image "rotates" with the sample, the imaged features are not influenced by the shape of the tip.

How to Avoid Tip-Related Artifacts

- The only way to avoid tip related artifacts is to use a tip which is sharp enough that the images are not influenced by the shape of the tip. What means “sharp enough” depends on the features to be imaged.
- If there are doubts about the sharpness of the tip, the tip should be exchanged.

8.2 Scanner-Related Artifacts

Scanner artifacts appear due to the non ideal behavior of the piezoelectric actuators, such as non-linearity, creep and hysteresis, as discussed in Sect. 3.5. These effects lead to the behavior that the distance moved by the actuator is not linearly related to the applied voltage. In AFM images these effects result in image distortions, which are most apparent if periodic grid structures are imaged. On a sample with no regular pattern these distortions are not easily recognized, but still present and can have a magnitude of up to 25%. If a closed loop scanner is used (Sect. 4.2.2), these artifacts are avoided.

An example of image distortion due to a non-linearity in the piezo extension is shown in Fig. 8.5. A silicide nanowire, which is known to be straight due to its crystallographic structure, is imaged as bent. If a tube scanner is used, the tip will follow an arc, resulting in a non-linearity in the lateral directions, as well as the vertical direction [4, 5].

When discussing problems of piezo actuators in Sect. 3.5, we have seen that the new position is not reached instantaneously after the corresponding voltage change, but is only reached asymptotically. If this creep is not yet finished this leads to an image distortion in the AFM images. Artifacts due to creep appear often at the beginning of an image resulting in a bending of all image structures, as seen in

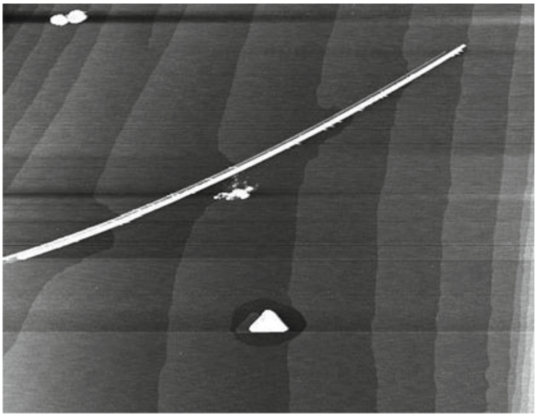


Fig. 8.5 Image of a straight silicide nano-wire, which appears bent in the image due to non-linearities in the piezoelectric actuators

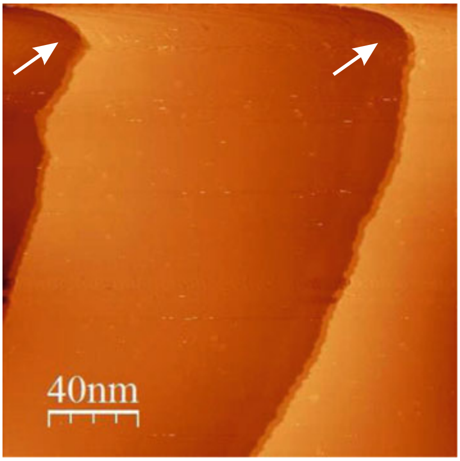


Fig. 8.6 Bending of atomic steps in the beginning of an image of a Si surface highlighted by *arrows*. Additionally to this artifact also an artifact due to a double tip is present in this image

Fig. 8.6. Specifically, when moving to a new lateral position away from the previous one this effect is strong.

How to Identify Scanner-Related Artifacts

- Scanner related artifacts can be identified by imaging structures which are known to be straight or periodic.
- Comparing trace and retrace images with opposite scan directions can help to identify artifacts due to creep and hysteresis.

How to Avoid Scanner-Related Artifacts

- The best way to avoid scanner related artifacts is to work in closed-loop, i.e. to have a sensor which measures the actual distance moved.

8.3 Feedback-Related Artifacts

Artifacts due to the feedback were already discussed in Sect. 5.8. If the feedback is too fast, scanning over a sharp protrusion leads to an overshoot (with oscillations) as shown in schematically Fig. 5.16. Different from this figure the overshoot (when the tip moves sharply up) is not the same as the undershoot (when the tip moves sharply down). This is because the difference between the setpoint value and the actual value is different in both cases. If the setpoint value is e.g. 1 nN in the static mode the lowest measured value is a vanishing force, while in the other direction when approaching the sample high force values of e.g. 10 nN can occur. This makes the feedback in both directions asymmetric and results in stronger overshoots than undershoots. A similar looking artifact with overshoots and undershoots at sharp edges can occur due to creep. If at a sharp upwards edge the height is (initially due to creep) not yet the equilibrium height, this leads to an overshoot of the z-voltage and thus the height.

If the feedback is too slow and when the tip moves sharply down a "flying effect" can occur due to which the tip has no contact to the surface for a certain time.

How to Identify Feedback-Related Artifacts

- These artifacts can be identified due to their opposite behavior on the scan direction (trace and retrace). When scanning a certain area two images can be created one in the forward scan direction and one which the fast scan direction is scanned in the opposite direction. Comparing both images overshoot and undershoot exchange for feedback-related artifacts. This results in the effect that at sharply decreasing edges in the image is blurred.
- These artifacts can be identified by monitoring the error signal i.e. the deviation of the measured signal from the setpoint. The measured signal is the cantilever deflection in the contact mode, or the oscillation in the dynamic AM mode. Larger values of the error signal, particularly at steep steps in the topography, indicate feedback-related artifacts.

How to Avoid Feedback-Related Artifacts

- Feedback-related artifacts can be avoided by optimizing the feedback parameters as described in Sect. 5.8 or by reducing the scan speed.

8.4 Artifacts Due to Periodic Noise

Noise with a high amplitude at a specific frequency will show up as stripes superimposed onto the true topography of the surface. Electrical noise from the power line is 50 Hz (or 60 Hz) noise, which can be recognized as stripes in the images, as shown in Fig. 8.7. Changing the scan speed will change the ratio of the 50 Hz noise to the frequency at which the scan lines are acquired. This has a massive influence on the angle of the observed stripe patterns. To remove electrical noise, careful debugging of the electronics has to be performed, including the removal of ground loops. Vibrational noise can be acoustic noise or floor vibrations of the building. In the Sect. 3.6 on vibration isolation, we discussed how to avoid this kind of noise.

How to Identify Artifacts Due to Periodic Noise

- Artifacts due to periodic noise can be distinguished from topographic features by changing the scan speed. Real topographic features are independent of the scan speed while artifact related features due to periodic noise change their apparent size (periodicity) with the scan speed (observed stripes change their angle).

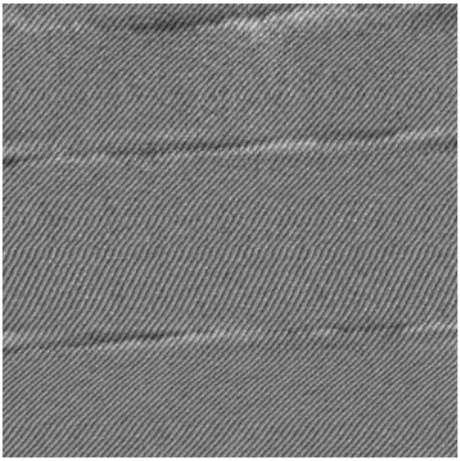


Fig. 8.7 Example of an image which is strongly influenced by 50 Hz noise. The three horizontal atomic step edges are hardly visible due to the strong 50 Hz noise

How to Avoid Artifacts Due to Periodic Noise

- Artifacts due to periodic noise can be avoided by removing the respective source of this type of noise: mechanic vibrations, acoustic noise, or electronic noise, which is often a difficult task.
- Sometimes the apparent visibility of this type of noise in the images can be suppressed (not removed) by changing the scan speed.

8.5 Thermal Drift

If tip and sample are at different temperatures or if the temperature in the room changes, thermal drift occurs. Already small temperature changes of less than 1 °C can result in a substantial drift on the nanoscale.

How to Identify Artifacts Due to Thermal Drift

- Artifacts/distortions due to thermal drift can be identified by repeated scanning of nominally the same area. By the apparent shift of the same topographic features in subsequent images the drift speed can be measured.

How to Avoid Artifacts Due to Thermal Drift

- After starting a measurement by inserting a sample into the AFM, some time should be waited until sample and AFM have equilibrated.
- Temperature variations should be minimized. Particularly sun light shining directly or indirectly to the AFM can induce thermal drift.

- Fast scanning can minimize distortions due to thermal drift, however, other artifacts like feedback related artifacts increase with the scanning speed.

8.6 Laser Interference

If the light from the laser diode (beam deflection AFM) reaches not only the backside of the cantilever, but also the sample, interference between both beams occurs. This type of artifact appears as oscillations in the images as well as in the baseline signal when tip and sample are approached. The periodicity corresponds to the wavelength of the laser light.

How to Identify Artifacts Due to Laser Interference

- Stripes in the image which do not change with the scan speed.
- Oscillations in the baseline signal upon tip-sample approach before tip-sample interaction are visible.

How to Avoid Artifacts Due to Laser Interference

- The laser beam should be focused to the middle of the cantilever width, so that almost no light shines on the sample.

8.7 Summary

- The shape of the tip influences the AFM images, resulting in multiple images. The combination of sharp surface features with a blunt tip leads to the tip shape being imaged.
- When imaging with a blunt tip, parts of the features at the surface are not imaged: “dead zone”.
- Piezo non-linearity, creep, and hysteresis leads to distorted images, which can be avoided by a closed loop operation of the scanner.
- Power line noise, feedback overshoot, thermal drift, and laser interference are further sources of image artifacts.

References

1. P. Klapetek, *Quantitative Data Processing in Scanning Probe Microscopy*, 2nd edn. (Elsevier, Amsterdam, 2018). ISBN 9780128133477
2. F. Golek, P. Mazur, Z. Ryszka, S. Zuber, AFM image artifacts. Appl. Surf. Sci. **304**, 11–19 (2014). https://doi.org/10.1016/j.apsusc.2014.01.149

3. P. Eaton, K. Batziou, Artifacts and practical issues in atomic force microscopy, in *Atomic Force Microscopy*, ed. by N. Santos, F. Carvalho. Methods in Molecular Biology, vol. 1886 (Humana Press, New York, 2019). ISBN: 978-1-4939-8893-8. https://doi.org/10.1007/978-1-4939-8894-5_1
4. M. Hannss, W. Naumann, R. Anton, Performance of a tilt compensating tube scanner in atomic force microscopy. Scanning **20**, 501 (1998). https://doi.org/10.1002/sca.1998.4950200703
5. C. Wei, A circular arc bending model of piezoelectric tube scanners. Rev. Sci. Instrum. **67**, 2286 (1998). https://doi.org/10.1063/1.1146934

Chapter 9
Work Function, Contact Potential, and Kelvin Probe AFM

We already used the term work function when we introduced the tunneling barrier height in STM. The work function can be considered as the energy difference between the vacuum level and the Fermi level of a metal. Here we will see that also a surface term contributes to the work function. The work function is a measurable quantity and the operative definition of the work function is that it is the energy required to remove an electron from the bulk Fermi level of a metal to a certain distance from the solid.[1]

Subsequently, we introduce the contact potential between two metals with different work function, which is used by the Kelvin method for the measurement of work function differences. In spite of the fact that we have not yet introduced AFM in depth, in this chapter we already present the principles of Kelvin probe scanning force microscopy (KFM), which is the nanoscale variant of the Kelvin method.

9.1 Work Function

The work function Φ of a metal can be defined as the difference between the energy of an electron at some distance d outside of a solid E_{out} and the energy of the highest occupied electron level (at zero temperature), i.e. the Fermi energy, thus

$$\Phi(d) = E_{\mathrm{out}}(d) - E_{\mathrm{F}}. \tag{9.1}$$

This corresponds to an operative definition of the work function as the minimum energy to bring an electron from the solid to some distance d outside the solid. The kinetic energy of the electron outside the solid is considered as zero. Note that with this definition the work function depends on how far the electron is removed from the surface.

[1] This distance is specific to the actual type of measurement performed.

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3_9

As a limiting case, the energy to bring the electron from inside the solid to infinity can be considered. Let us consider an infinite crystal filling a half space and being terminated by an infinite surface of specific orientation. If the position of the electron outside of the solid is infinitely far from the solid E_{out} will be the vacuum energy at infinite distance from the surface $E_{\mathrm{vac}}^{\infty}$ and the work function results as

$$\Phi = E_{\mathrm{vac}}^{\infty} - E_{\mathrm{F}}. \tag{9.2}$$

The usual definition of the work function as difference between vacuum energy and Fermi energy hides the fact that the vacuum energy depends on the distance of the electron from the surface.

The work function has two main contributions; one is due to the binding of the electrons inside a solid. Theoretically, one can consider the binding of the electrons inside a solid with different levels of sophistication, from the simple nearly free electron model, the tight binding model, up to ab initio calculations. The essence is always the same: The electrons are bound inside a solid and this bonding corresponds to a lower energy of the electrons in the solid compared to free electrons. A second contribution to the work function arises due to the passage of the electron through the surface layer, which we will discuss in the following.

9.2 Effect of a Surface on the Work Function

Before we consider the effect of the surface on the work function, we note that the effect of the presence of a surface has a negligible effect on the bulk states. Inside the solid the potential of the positive charges of the nuclei is screened very effectively by the electrons at distances larger than the Thomas-Fermi screening length [1]. The Thomas-Fermi screening length is usually very small in metals. For instance, in copper the screening length is only about 0.5 Å. Thus, inside the crystal everything will remain as it was in the infinite bulk crystal since the contribution of the "missing" atoms at the surface is vanishingly small due to the effective screening inside the metal. The energy of the highest occupied electronic level in a metal terminated by a surface will still be E_{F}, as for the infinite crystal.

Now we consider how the changes of the electronic structure at the surface give rise to an additional contribution to the work function, i.e. we consider the work needed to bring an electron through the surface layer. Even if we consider a bulk termination of the surface, which means that the positions of the atom nuclei remain as in the bulk, i.e. undistorted up to the last atom at the surface, as shown for the 1D crystal in Fig. 9.1a, the electron charge distribution near the surface deviates from that in the bulk. Some charge will "spill out" into the vacuum as indicated qualitatively in Fig. 9.1a. This "spill out" of charge is a quantum mechanical effect, as an electron

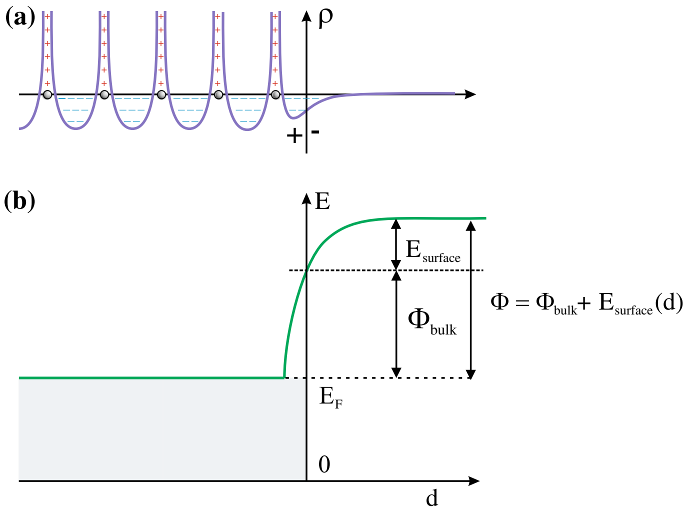


Fig. 9.1 **a** Charge density in a metal crystal which is modified close to the surface and spills out towards the vacuum. This behavior can be described qualitatively by a dipole layer of excess charge density close to the surface. **b** Energy of an electron as function of the distance d from the surface resulting from the charge density given in **a**. The passage of an electron through the dipole layer leads to additional work E_{surface} which has to be done in order to remove an electron from the solid

can reduce its energy when it spreads out over a larger region.[2] The "spill out" of charge at the surface leads to the formation of a charge dipole at the surface with negative charge "spilling out" towards the vacuum and less negative charge (i.e. a positive excess charge) inside the crystal close to the surface as indicated in Fig. 9.1a. The particular way in which the charge distribution at the surface deviates from the bulk structure depends on the crystal structure at the surface (bulk terminated or modified, i.e. known as reconstructed). When an electron is removed from the solid, a contribution to the work function arises from the transfer of the electron through the dipole layer.

The direction of the field in the dipole layer is (usually) such that an additional amount of work E_{surface} has to be done to move an electron through the dipole layer. The total energy to remove an electron at E_{F} from the solid to some distance d consists of a bulk contribution (binding energy) plus the work done by the electron when passing through the dipole layer now reads

[2] This can be seen from a simple 1D particle in a box model, where the energy of an electron state as a function of the quantum number n and size of the box L is

$$E(L) = \frac{\hbar^2 \pi^2 n^2}{2 m_{\text{e}} L^2}. \tag{9.3}$$

With increasing L ("spill out" of charge) the energy decreases.

$$\Phi(d) = \Phi_{\text{bulk}} + E_{\text{surface}}(d). \tag{9.4}$$

The corresponding energy diagram is shown in Fig. 9.1b. Inside the solid the free electron approximation is used with the energy levels filled up to the Fermi energy. When passing through the dipole layer the additional contribution to the energy E_{surface} is added. This surface contribution to the work function can be of the order of up to 1 eV.

The splitting of the work function into different contributions arises from the different approaches used for each effect. A ab initio quantum mechanical theory would include all these effects when an electron is moved from inside the crystal to an distance from the crystal. Besides the influence of the surface which is difficult to calculate with ab ab initio methods, also the electrostatic potential at larger distances from the surface is difficult to calculate quantum mechanically. The correlation and exchange forces outside the surface cannot be calculated quantum mechanically up to large distances of 100 nm. The electrostatic image potential is often used as an approximation of the long-range behavior of the exchange-correlation potential in the vacuum.[3] On the other hand, for short distances the unrealistic divergence of the classical image potential at the surface is avoided by a transition to quantum mechanical calculations, which describe the region close to the surface better.

The work due to the electrostatic image charges (occurring when an electron is moved out of the metal) reduces at the distance of 100 nm to 1 % of the value at 1 nm, and can thus be neglected for larger distances.

In conclusion we have identified three contributions to the work function: the bulk contribution (binding energy), the surface contribution, and the image charge contribution. These are the contributions which enter for a distance of the removed electron up to 100 nm. A further contribution occurs if the electron is removed to distances comparable to the size of the sample, and results due to external electric fields, as will be discussed in the next section.

9.3 Surface Charges and External Electric Fields

Now we consider (different from the semi infinite crystal considered so far) a finite crystal with is terminated by different surfaces, as shown in Fig. 9.2. Different surfaces (with different atomic configurations) terminating a crystal, correspond to

[3] In classical electrostatics it is shown that the force between an electron at distance d from a conducting plate is the same as the force between the electron and a positive elementary charge located at a distance $2d$ from the electron (image charge), i.e. $-e^2/(4\pi\varepsilon_0 4d^2)$. Integrating the negative of this force from infinity to d results in the (image) potential of the electron (relative to a position at infinity) as

$$V_{\text{image}}(d) = \int_{\infty}^{d} \frac{e^2}{4\pi\varepsilon_0 4r}\text{dr} = \frac{-e^2}{4\pi\varepsilon_0}\frac{1}{4d}. \tag{9.5}$$

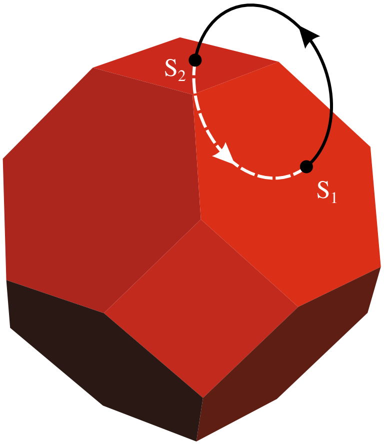


Fig. 9.2 Due to energy conservation, zero total work has to be done in moving an electron along the closed path from inside the metal crystal through surface S_1 and back through surface S_2. This argument shows that the two surfaces S_1 and S_2, which are assumed to have different work functions, have to be at different electrostatic potentials. This different potentials are built up by corresponding surface charges

different "spill out" of charge. This leads to different surface dipoles and therefore also to different work functions at different surfaces of a crystal. In the following, we will show that these different work functions at different surfaces of a finite crystal lead to the presence of net surface charges, and corresponding electric fields.

Let us take an electron on a closed loop from a point inside the crystal to a position outside of the crystal through surface S_1 and back through another surface S_2, as shown in Fig. 9.2. Leaving the crystal through surface S_1 requires work E_1 (surface work to leave the crystal through surface S_1, plus of course also the bulk contribution to the work function, which we leave out here, since it cancels out later). If there were no net surface charges, the electric field outside the crystal would vanish and there would be no work to transfer the electron outside the crystal from surface S_1 to surface S_2. When the electron is inserted back into the crystal through S_2, the work $-E_2$ (negative of the surface work to leave the crystal through surface S_2) is gained. Closing the path inside the metal does not involve energy, since the electric field inside a metal is vanishing. Since the work functions of the two surfaces are different (due to the two different surface contributions to the work function), a perpetuum mobile could be built gaining the energy difference between the two work functions $(E_1 - E_2)$ on each cycle. Since this is clearly impossible, there must be an electric field outside the crystal against which a compensating amount of work is done as the electron is carried from S_1 to S_2. This means the two surfaces must be at two different electrostatic potentials φ_1 and φ_2, satisfying the condition

$$e(\varphi_1 - \varphi_2) = E_1 - E_2 = \Phi_1 - \Phi_2. \tag{9.6}$$

Since dipole layers cannot yield macroscopic fields outside the crystal these fields have to arise from net macroscopic electric charges on the surfaces,[4] which also lead

[4] All net charges are located at the surface of a metal, since the electric field vanishes in the interior of a metal.

to an external electric fields with a range corresponding to the size of the crystal. At larger distances from the crystal these fields vanish.

In the following, we estimate which surface charge density is necessary to "supply" the necessary energy to compensate for the surface-related work function difference of the order of about 1 eV when an electron is transferred macroscopic distances from one metal surface to the other through the outer electric field. For a rough estimate, we consider a plate capacitor arrangement ($d = 1\,\mathrm{cm}$). The surface charge per area A can be expressed as

$$\rho_{\mathrm{surface}} = \frac{Q}{A} = \frac{VC}{A} = \frac{V}{A}\frac{\epsilon_0 A}{d} = \frac{V\epsilon_0}{d}. \tag{9.7}$$

The resulting surface charge corresponds to $\sim 5 \times 10^{-8}$ electrons per surface atom. This shows that even minute charge densities at the surface lead to considerable work, since the distance over which the electric field extends are on the order of the size of the crystal.

Now we will summarize the results on the work to remove an electron from the solid as a function of the distance d. An electron is considered to be removed from the highest occupied level at E_{F}. At very short distances from the surface ($<1\,\mathrm{nm}$), the bulk contribution (bonding energy), as well as the surface contribution are the main contributions to the work. (At surfaces with different electronic structure, the different surface contributions lead to different work functions Φ_1 and Φ_2.) For distances larger than 1 nm from the surface these contributions remain constant. At distances between 1 and 100 nm the work due to the image charge effect is the only distance dependent part of the work function. Between $\sim$100 nm and $\sim$1 mm (a distance corresponding to the sample size) there are no further contributions to the work function. When the distance of the electron removed from the solid becomes close to the sample size, the work due to the external electric fields arising from the previously discussed surface charges contribute to the work.

The work to bring an electron to infinity Φ^{∞} is independent on the work function of the surface through which it passed.[5] Any differences due to the surface work are compensated by macroscopic electric fields created by the surface charges at the different surfaces.

Experimental measurements of the work function are performed at a certain distance. Since most of the experiments are performed in a distance range between 100 and 1 mm, in which the work function is independent of the distance, usually work functions are considered as independent of the distance. An exception is scanning probe microscopy. In scanning tunneling microscopy the distance to which the electron is transferred out of the solid is very small ($<1\,\mathrm{nm}$). Thus, the image potential and even the surface and bulk contributions can be distance dependent at such small distances. The apparent barrier height Φ in STM is more a parameter than directly

[5] It is always assumed that the electron is at rest, i.e. there is no kinetic energy contribution to the work.

corresponding to the work function. Nevertheless, the apparent tunneling barrier height is usually referred as "the work function" and also we will use this not correct wording sometimes.

9.4 Contact Potential

Now we assume two (different) metals with different work functions which are initially not connected to each other Fig. 9.3a.[6] In this case, both metals share a common vacuum level, but their Fermi levels are not aligned, due to the different work functions assumed. Suppose now that these two metals are connected (e.g. by a wire) in such a way that electrons can flow freely from one metal to the other, as shown in Fig. 9.3b. In this case, both metals share a common Fermi level. Since initially the two Fermi levels were not yet aligned, electrons flow through the wire from the metal with the higher Fermi level until equilibrium is reached. However, the charge transfer in order to align the two Fermi levels does *not* occur in such a way that half of the electrons between energy $E_{\mathrm{F},1}$ and $E_{\mathrm{F},2}$ flow from metal 2 to metal 1. A very small transfer of charge builds up a surface charge at the metals and a corresponding electric field $\mathscr{E}$ between them. According to (9.7), over the (macroscopic) distance d these surface charges induce a potential drop V_{contact}, which aligns the Fermi levels of the metals. Due to the macroscopic distance only minute surface charges are needed to build up a voltage on the order of the work function difference.

In equilibrium the condition

$$eV_{\text{contact}} = \Delta\Phi \tag{9.8}$$

holds. The voltage V_{contact} is called contact potential, because it occurs if a contact between the metals is established, for instance by a connecting wire.

9.5 Measurement of Work Function by the Kelvin Method

Equation (9.8) suggests that a simple way to measure the (relative) work function of a metal is to measure the contact potential (relative to a metal with known work function) by connecting a voltmeter between the metals. However, this is not possible since a continuous flow of current (through the voltmeter) would have been produced without a sustaining source of energy. Lord Kelvin proposed a simple way to measure contact potentials by a capacitive method which is described in the following. The two samples are arranged in such a way that the two surfaces form a plate capacitor

[6] We assume semi infinite crystals so that no surface charges are present and thus no electric fields occur outside the crystals. Since in Fig. 9.3a macroscopic distance between both metals is assumed, the work function rises within 100 nm quasi vertically to $E_{\text{vac}} = E_{\text{vac}}^{\infty}$.

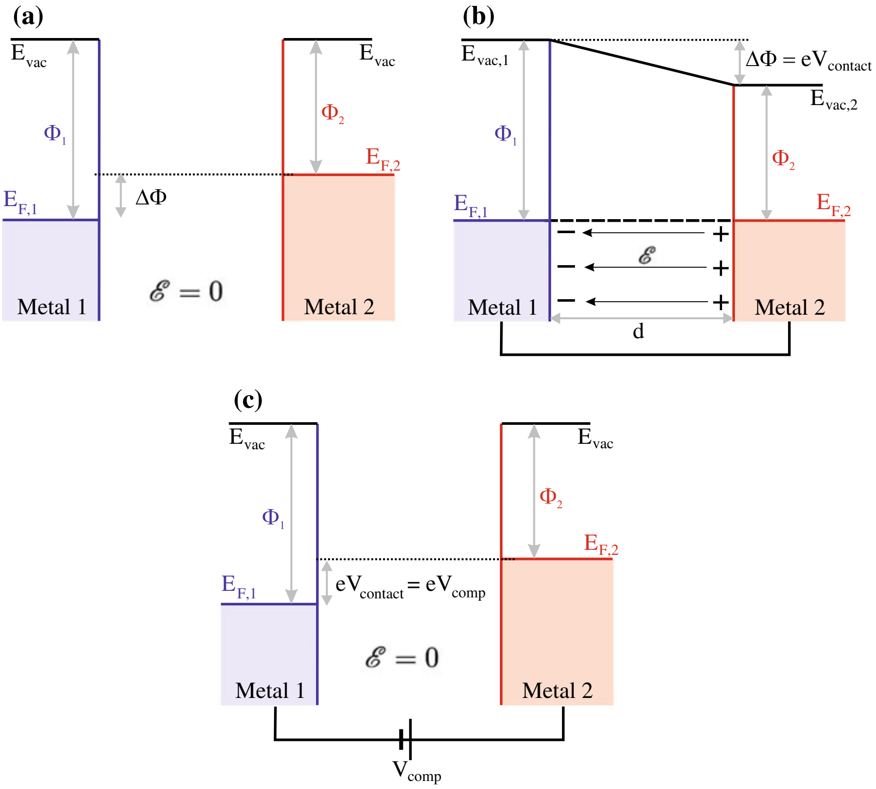


Fig. 9.3 **a** Potential energy diagram for two metals with work functions Φ_1 and Φ_2, which are initially not connected and share thus a common vacuum level. **b** If the two metals are connected by a conducting wire, the Fermi levels of the two metals align. A buildup of surface charge leads to a macroscopic potential gradient compensating the difference between the work functions of the two metals. **c** The surface charges and the corresponding electric field $\mathscr{E}$ vanish if a voltage $V_{\text{comp}} = V_{\text{contact}} = \frac{1}{e}\Delta\Phi$ is applied between the metals

and an outer voltage called the compensation voltage V_{comp} is applied between the surfaces (Fig. 9.4). The total potential difference V can be written as

$$V = V_{\text{contact}} - V_{\text{comp}}. \tag{9.9}$$

The charge on the capacitor is accordingly

$$Q = CV = C\left(V_{\text{contact}} - V_{\text{comp}}\right). \tag{9.10}$$

If the distance between the capacitor plates d is now modulated sinusoidally (for instance by a piezoelectric actuator) with a small modulation amplitude a current results as

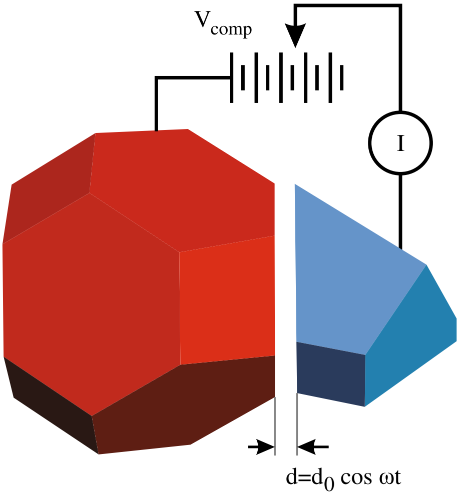


Fig. 9.4 The surfaces of two metals are brought together in a plate capacitor configuration. When the distance *d* between the plates is modulated a charge flow (capacitive current) can be measured. When an external bias potential just compensates the work function no current flows anymore

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t} = \frac{\mathrm{d}C}{\mathrm{d}t}\left(V_{\text{contact}} - V_{\text{comp}}\right), \tag{9.11}$$

since V_{contact} is constant and V_{comp} varies slowly compared to the modulation voltage. Therefore, a capacitive current is only induced by a change in the capacitance of the plate capacitor ($C = \epsilon_0 A/d$). The measured current has linear behavior as function of $V_{\text{contact}} - V_{\text{comp}}$. The current will vanish if V_{contact} or equivalently the work function difference is compensated by the compensation voltage, i.e. if

$$V_{\text{comp}} = V_{\text{contact}} = \frac{1}{e}\Delta\Phi. \tag{9.12}$$

No current flows if this condition is fulfilled and also the electric field between the metals vanishes as shown in Fig. 9.3c. The amplitude of the (capacitive) current can be measured sensitively using the lock-in detection method as a function of the compensation voltage. Using this method, the (macroscopic) contact potential difference between two metals can be measured.

9.6 Kelvin Probe Scanning Force Microscopy (KPFM)

While Kelvin probe scanning force microscopy [2] is the microscopic variant of the Kelvin method, there are also some differences. In the macroscopic Kelvin method the distance between the two metals is modulated and the resulting capacitive current is measured, whereas in Kelvin probe scanning force microscopy the voltage between tip and sample is modulated and the corresponding electric (capacitive) force is measured.[7] For conceptual simplicity we consider a flat surface and the tip is moved at a constant topographic distance over this surface. However, we consider that the surface consists of areas with different work functions which we would like to detect. Our configuration consists of a surface and a tip with a voltage V between them, and a capacitance $C(z)$ for the tip-sample system. Apart from other forces, there is an electrical force between the tip and the sample. If we consider the tip-sample system as a capacitor, the electrical (capacitive) force between tip and sample is the gradient of the potential energy of the capacitor as

$$F_{\mathrm{el}}(z, V) = -\frac{\partial E}{\partial z} = -\frac{1}{2}\frac{\partial C}{\partial z} V^2(t). \tag{9.13}$$

Since we assume a scan at constant tip-sample distance, $\partial C/\partial z$ is a constant. The voltage between tip and sample consists of different contributions: the constant contribution $V_{\mathrm{contact}} - V_{\mathrm{comp}}$, and additionally a voltage component which is modulated at the modulation frequency ω_{mod} resulting in a total voltage between tip and sample as

$$V(t) = V_{\mathrm{contact}} - V_{\mathrm{comp}} + V_{\mathrm{mod}} \cos\left(\omega_{\mathrm{mod}} t\right) \tag{9.14}$$

Thus, the tip-sample force which is proportional to the square of the tip-sample voltage $V(t)$ results as

$$\begin{aligned} F_{\mathrm{el}}(V) &= -\frac{1}{2}\frac{\partial C}{\partial z}\left[V_{\mathrm{contact}} - V_{\mathrm{comp}} + V_{\mathrm{mod}} \cos\left(\omega_{\mathrm{mod}} t\right)\right]^2 \\ &= -\frac{1}{2}\frac{\partial C}{\partial z}\Big[\left(V_{\mathrm{contact}} - V_{\mathrm{comp}}\right)^2 + 2\left(V_{\mathrm{contact}} - V_{\mathrm{comp}}\right) V_{\mathrm{mod}} \cos\left(\omega_{\mathrm{mod}} t\right) \\ &\qquad + V_{\mathrm{mod}}^2 \cos^2\left(\omega_{\mathrm{mod}} t\right)\Big]. \end{aligned} \tag{9.15}$$

The first term in the square bracket is time independent (constant), the second term is a modulation with the frequency ω_{mod}, while the third term consists (after using a mathematical identity) of a constant term plus a component at twice the frequency ω_{mod}. Using the lock-in technique, which we introduced in Chap. 6, the amplitude of the term at the frequency ω_{mod} can be selectively measured. This component vanishes if $V_{\mathrm{contact}} - V_{\mathrm{comp}} = 0$. In the practical implementation, a feedback control of V_{comp} keeps the ω_{mod} component of the force at zero. Thus, by recording the voltage V_{comp}, which nulls the ω_{mod} component of the force signal $\propto \frac{1}{e}\Delta\Phi - V_{\mathrm{comp}}$, the work function difference is measured locally on the nanoscale while scanning over the surface. Due to the modulation of the voltage V, a modulated force is exerted on the cantilever, which induces a cantilever oscillation at the modulation frequency.

So far we have left out the complication that in a practical implementation of an SPM setup the tip-sample distance also has to be measured, and to adapt the setpoint value. In dynamic atomic force microscopy this can be done using a (second) modulation of the cantilever close to its resonance frequency (as we discuss in detail

[7] This is done since the force (not the current) is measured in a AFM setup.

in Chap. 13). Thus, the cantilever is modulated at two (different) frequencies and two lock-in detection units detect the oscillation amplitudes at the respective modulation frequency.

9.7 Summary

- The definition of the work function as the difference between the vacuum level and the Fermi level, includes also a surface contribution to the work function.
- Due to a “spill out” of charge to the vacuum, a charge dipole occurs at the surface. A certain amount of work has to be done to move an electron through this dipole layer. This is the surface contribution to the work function.
- Also a net charge can accumulate at the surface giving rise to a contact potential between metals with different work functions. The contact potential is the difference between the work functions.
- The contact potential can be measured using the Kelvin method by modulating the distance between the surfaces of the metals and measuring the induced capacitive current.
- In Kelvin probe scanning force microscopy (KFM) the work function can be measured locally by modulating the tip-sample voltage.

References

1. H. Ibach, H. Lüth, *Solid-State Physics – An Introduction to Principles of Materials Science*, 4th edn. (Springer, Heidelberg, 2009). https://doi.org/10.1007/978-3-540-93804-0
2. S. Sadewasser, Th. Glatzel (eds.), *Kelvin Probe Force Microscopy - Measuring and Compensating Electrostatic Forces*, 1st edn. (Springer, Berlin, 2012). https://doi.org/10.1007/978-3-642-22566-6

Chapter 10
Forces Between Tip and Sample

The idea behind the atomic force microscope (AFM) is to measure the force between the surface and the scanning tip in order to track the surface topography. Before we describe the atomic force microscopy technique in detail, we consider the forces acting between tip and sample as well as the tip-sample contact mechanics. We consider also the snap-to-contact phenomenon, which can occur due to attractive tip-sample forces.

10.1 Tip-Sample Forces

The total force between tip and sample is composed of several long-range and short-range contributions, which we will discuss in the following. One long-range contribution is the van der Waals force. The van der Waals force in the narrower sense, here specifically the London dispersion force, is a force between neutral atoms or molecules without a permanent dipole moment. It can be described as a spontaneous formation of fluctuating electric dipoles which attract each other. The origin of the van der Waals force is of quantum mechanical nature. There are several levels of approximation for this force, at the most exact level it is a quantum-electrodynamical phenomenon which is called the Casimir–Polder force [1].

For the simple case of two noble gas atoms (distance r) the dipole interaction between them can be treated analytically using some approximations [2], resulting in an interaction potential of

$$U_{\mathrm{vdW}}(r) = -\frac{C}{r^6}. \tag{10.1}$$

The distance dependence with the minus sixth power corresponds to a long-range interaction. The van der Waals interaction is (in this approximation) non-directional

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3_10

(isotropic) and additive, which means that for two groups of atoms the total interaction energy is the sum of all pair potentials. Taking a sample and an AFM tip as an example, not only the atoms in the vicinity of the tip apex contribute to the van der Waals force, but also the forces of atoms in a larger volume of the tip and sample have to be summed up, because of the long range of the force. The total interaction can be obtained by integration. The van der Waals interaction energy between an infinitesimal volume element of the tip $\mathrm{d}V_{\mathrm{tip}}$ and an infinitesimal volume element $\mathrm{d}V_{\mathrm{sample}}$ of the sample can be written as

$$dU_{\mathrm{vdW}} = -\frac{C\rho_{\mathrm{tip}}\rho_{\mathrm{sample}}}{\left|\mathbf{r}_{\mathrm{tip}} - \mathbf{r}_{\mathrm{sample}}\right|^{6}}\mathrm{d}V_{\mathrm{tip}}\mathrm{d}V_{\mathrm{sample}}, \tag{10.2}$$

with ρ_{tip} and ρ_{sample} being the atom densities of tip and sample, respectively. Approximating the tip by a sphere of radius R_{tip} and the sample by a semi-infinite solid results in a van der Waals interaction energy [2] of

$$U_{\mathrm{vdW}} = -\frac{HR_{\mathrm{tip}}}{6d}, \tag{10.3}$$

where R_{tip} is the tip radius, d the tip-sample distance measured from the tip apex, and H is the Hamaker constant. The Hamaker constant is a material property representing the strength of the van der Waals interaction [2, 3]. It is defined as $H = \pi^2 C\rho_{\mathrm{tip}}\rho_{\mathrm{sample}}$, with C being the coefficient in the atom-atom pair potential in (10.1). Typical values for the Hamaker constant are in the range of several eV. The van der Waals force between the tip and sample results as

$$F_{\mathrm{vdW}} = -\frac{\partial U_{\mathrm{vdW}}}{\partial d} = -\frac{HR_{\mathrm{tip}}}{6\mathrm{d}^2}. \tag{10.4}$$

For tip-sample distances larger than 1 nm the van der Waals force is the largest force. Apart from the van der Waals force, short-range forces arise from the overlap of the electron wave functions of the outermost shell (chemical bond). These short-range forces have a range of less than a nanometer and can be attractive or repulsive. If the overlap of the electron wave functions of the outer shell reduces the total energy, these chemical bond forces are attractive. We shall not elaborate on the nature of chemical bonds further here, as this topic is treated in detail in textbooks on chemistry and physics.

If we consider a metal tip and a metal surface, an attractive interaction (some kind of metallic bonding) can be expected if tip and sample approach closely. One effect which does not actually occur is that the nuclei repel each other, as they are well shielded by the inner electron shells. When the tip and the sample atoms approach each other at distances closer than those in a chemical bond, the repulsion between the inner electron shells becomes important. The repulsive interaction due to the overlap of inner closed shell orbitals is not just the electrostatic repulsion of the electrons of the closed shells. There is also a quantum mechanical component called Pauli repulsion. In a simple form, the Pauli exclusion principle states that no two electrons can occupy the same state. In the overlapping region between the atoms the states of each atom are not only occupied by "their own electrons" but also partially by electrons of the other atom. Since the low-lying states are all filled (closed shell) these additional electrons from the other atom have to deviate to higher-lying states, leading effectively to a repulsive interaction if the electron wave functions of two neutral atoms with closed shells intrude into each other. The Pauli repulsion is introduced here in simple terms but in a more complete treatment the general form of the Pauli exclusion principle has to be applied. The multi electron wave function must be anti-symmetric under the exchange of two electrons.

All these short-range interactions are included in a quantum mechanical treatment by the Schrödinger equation. However, the (exact) solution of the Schrödinger equation of a system with several electrons is very difficult except for very simple cases. Therefore, model potentials are often used for the qualitative discussion of tip-sample interactions.

A frequently used model potential is the Lennard-Jones potential. This potential describes the interaction between two neutral atoms and consists of a term describing the attractive part of the interaction (van der Waals interaction) and a part describing the repulsive interactions, assumed to be proportional to $1/r^{12}$, as

$$U_{\mathrm{LJ}}(r) = 4U_0 \left[\left(\frac{R_a}{r} \right)^{12} - \left(\frac{R_a}{r} \right)^{6} \right], \tag{10.5}$$

where U_0 is the depth of the potential well, r is the distance between the atoms, and R_a is the distance at which $U_{\mathrm{LJ}}(r)$ is zero. In Fig. 10.1a the Lennard-Jones potential is shown as a red line, as well as the two contributions, the attractive $-1/r^6$ contribution (green) and the repulsive $1/r^{12}$ contribution (blue). While the Lennard-Jones potential is intended to model the interaction between neutral atoms, it also captures the basic features of the tip-sample interaction: attractive interaction for large distances, a potential minimum, and a strong repulsive interaction at short distances. Therefore, we will often use this model potential to describe tip-sample interactions. The Lennard-Jones potential and the corresponding force $F = -\frac{\partial U}{\partial r}$ as well as the force gradient (which will be important in the dynamic mode of AFM) are shown in Fig. 10.1. The shape of the curves is roughly similar, but shifted to the right, as the zero of the potential gradient (force) is at the minimum of the potential, and the zero of the force gradient is at the minimum of the force. The boundary between the attractive regime (negative force) and the repulsive regime (positive force) is indicated as a dashed line in Fig. 10.1 and occurs where the force changes its sign, or correspondingly at the minimum of the potential. If we use the Lennard-Jones potential in the following as tip-sample model potential, we replace r in (10.5) by the tip-sample distance d.

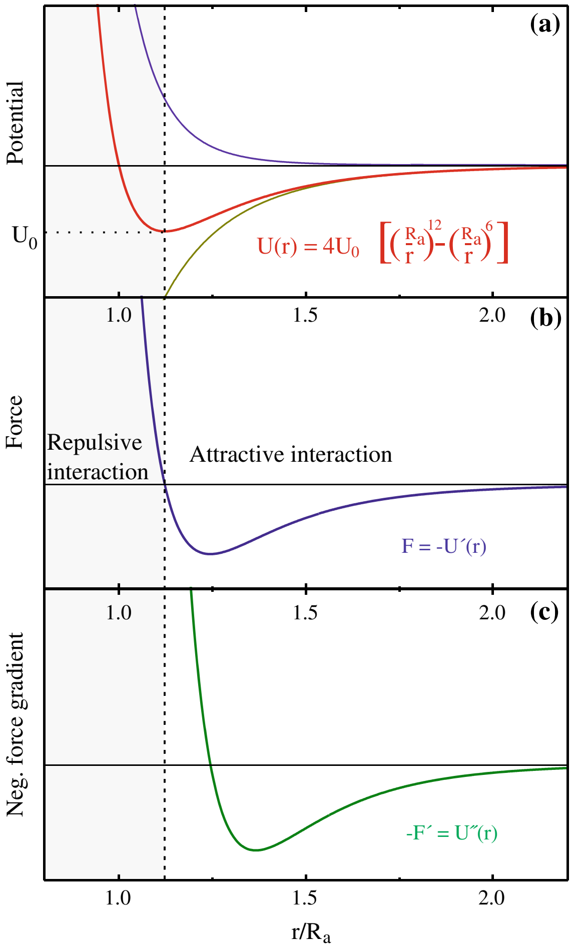


Fig. 10.1 **a** The Lennard-Jones potential will be used in the following as a model potential for a tip-sample interaction. The *green* and the *blue lines* show the attractive and the repulsive parts of the potential, respectively. The corresponding force is shown in **b** and the (negative) force gradient in **c**. The border between attractive and repulsive forces (interactions) is indicated by the *vertical dashed line*

10.2 Tip-Sample Contact Mechanics

If the tip and sample come into contact, not only the corresponding wave functions intrude into each other (as considered using the Lennard-Jones potential), but the positions of the atoms inside the solid change due to the elasticity of the tip and sample materials. This effect is described by the Hertzian theory of the elastic contact between two bodies [4]. The Hertzian theory was formulated for two elastic spheres coming into contact. If the radius of one sphere approaches infinity the situation of a spherical tip coming into contact with a sample surface is described as shown in Fig. 10.2.

The Hertzian theory predicts the elastic force which develops in response to an indentation described by the or tip-sample distance d as

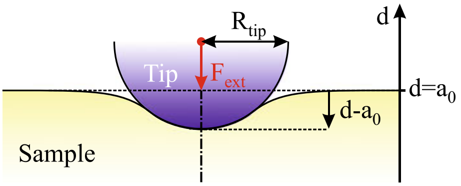


Fig. 10.2 Geometry of a contact between a sphere (tip) and a flat surface

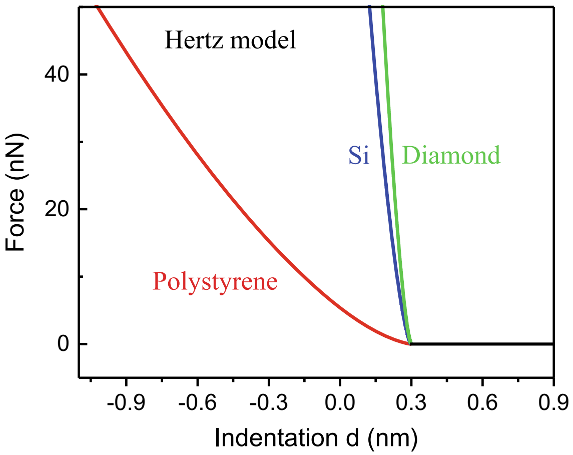


Fig. 10.3 Tip-sample force as function of the indentation according to the Hertzian theory. The curves are for samples of diamond Si, and a soft polymer sample. The following parameters were used: $a_0 = 0.3\,\text{nm}$, $\nu = 0.3$, $R_\text{tip} = 30\,\text{nm}$

$$F_\text{Hertz}(d) = -F_\text{ext}(d) = \frac{4}{3}E^*\sqrt{R_\text{tip}}(a_0 - d)^{3/2} \text{ for } d < a_0, \tag{10.6}$$

with the tip radius R_tip and the effective elastic modulus E^*

$$\frac{1}{E^*} = \frac{1-\nu_\text{tip}^2}{E_\text{tip}} + \frac{1-\nu_\text{sample}^2}{E_\text{sample}}. \tag{10.7}$$

The constants E and ν are the Young's modulus and the Poisson's ratio, respectively. The offset distance a_0 is used in order to bring the continuum approach of the Hertzian theory in accord with tip-sample distances d on the atomic scale. The distance a_0 corresponds to a typical inter-atomic distance and thus a contact is just established at the tip-sample distance $d = a_0$. For tip-sample distances $d < a_0$ (including negative values for d) a contact is formed. For distances $d > a_0$ no elastic contact is present and thus the force $F_\text{Hertz}(d)$ vanishes.

The force-distance dependence according to (10.6) is shown in Fig. 10.3 for three different surfaces: two hard samples (diamond and silicon) with $E_\text{diamond} = 1000\,\text{GPa}$ and $E_\text{Si} = 130\,\text{GPa}$, as well as a soft polymer sample (assumed as completely elastic) with $E_\text{poly} = 1.3\,\text{GPa}$. If the elastic modulus of the sample material (e.g. polymer) is much smaller than the elastic modulus of the tip material (e.g. silicon), this results in a less steep repulsive distance dependence.

The modeling of the tip-sample contact can be extended beyond the Hertzian theory by including attractive forces. In the DMT model (Derjaguin, Muller, and Toporov) a van der Waals type long range attractive force acting outside of the contact area between a spherical tip and the sample is added to the Hertzian force [5–9] as

$$F_{\mathrm{DMT}}(d) = F_{\mathrm{Hertz}}(d) + F_{\mathrm{vdW}}(a_0) = -F_{\mathrm{ext}}(d) \text{ for } d < a_0, \tag{10.8}$$

i.e., if the contact is established ($d < a_0$) a constant adhesion force of $F_{\mathrm{DMT}}(a_0) = F_{\mathrm{vdW}}(a_0) = -HR_{\mathrm{tip}}/(6a_0^2)$ according to (10.4) is added to the Hertzian force, while only $F_{\mathrm{vdW}}(d)$ is assumed for $d \geq a_0$. This gives rise to a DMT tip-sample force of

$$F_{\mathrm{DMT}}(d) = \begin{cases} F_{\mathrm{vdW}} = -\dfrac{HR_{\mathrm{tip}}}{6d^2}. & \text{for } d \geq a_0 \\ \frac{4}{3}E^*\sqrt{R_{\mathrm{tip}}}(a_0 - d)^{3/2} - \dfrac{HR_{\mathrm{tip}}}{6a_0^2} & \text{for } d < a_0. \end{cases} \tag{10.9}$$

The Derjaguin approximation relates the force law, F(d), between two curved surfaces to the interaction free energy per unit area, W(d), between two planar surfaces. This makes this approximation a very useful tool, since it is easier to derive the interaction energy for two planar surfaces rather than for curved surfaces.

If the attractive force between tip and sample is due to a van der Waals type force (i.e. no chemical bonding), the Derjaguin approximation [10] can be used to relate this adhesive force between a sphere and a plane to the surface energies per area γ as

$$F_{\mathrm{vdW}}(a_0) = -\frac{HR_{\mathrm{tip}}}{6a_0^2} = 2\pi R_{\mathrm{tip}}(\gamma_{\mathrm{tip}} + \gamma_{\mathrm{sample}} - \gamma_{\mathrm{tip-sample}}) = 2\pi R_{\mathrm{tip}}\Delta\gamma. \tag{10.10}$$

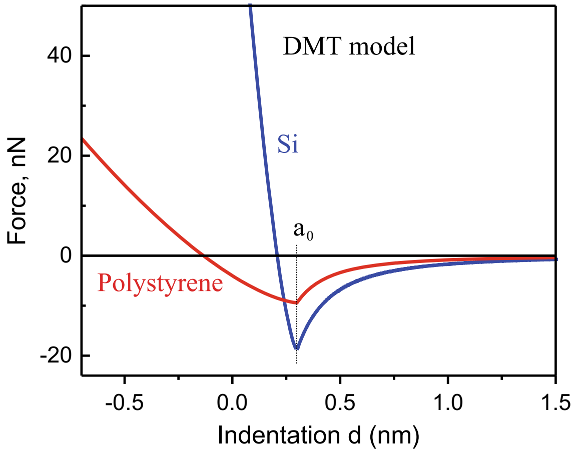


Fig. 10.4 DMT force $F_{\mathrm{DMT}}(d)$ as function of the indentation d according to (10.9). The two curves are for a hard Si sample and a soft polymer sample

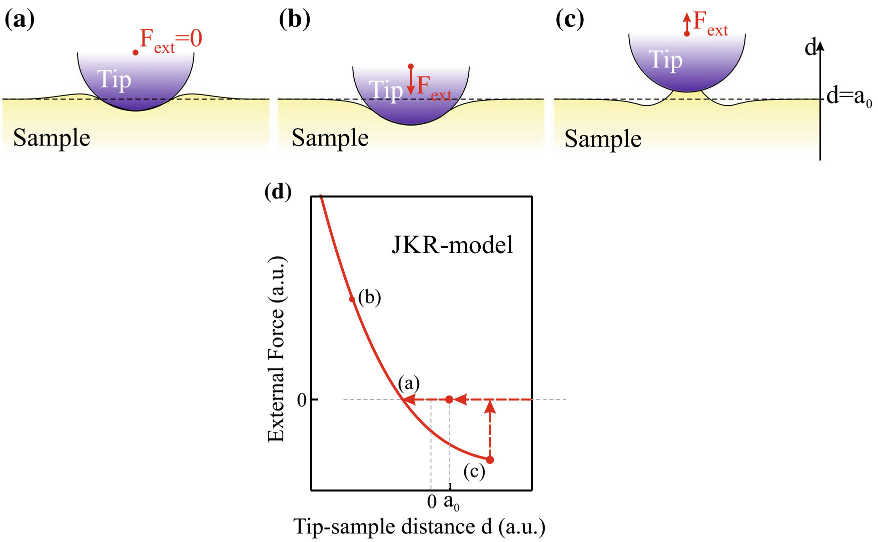


Fig. 10.5 JKR contact for the situations in which **a** the contact is just established and $d < a_0$ due to the adhesion force in the contact area, **b** an external applied force pushing the tip deeper into the sample, and **c** for an external force of opposite direction. In this case a tip-sample force for $d > a_0$ develops due to the adhesive tip-sample contact. **d** shows a plot of the hysteretic behavior of the external applied force as function of the tip-sample distance

The DMT force according to (10.9) and (10.10) is plotted in Fig. 10.4 for the two cases of a silicon sample ($\Delta\gamma = 0.1\,\text{J/m}^2$ assumed) and a soft polystyrene sample with the parameters used in Fig. 10.3 and an assumed surface energy of $\Delta\gamma = 0.05\,\text{J/m}^2$.

The limit in which the DMT theory is valid are hard materials with low surfaces energies and small tip radii [11] and a constant adhesion force independent of the indentation depth is assumed.

While the DMT theory considers long range attractive van der Waals-type tip-sample forces, the model of Johnson, Kendall, and Roberts (JKR model) [12] considers the opposite limit of a short range adhesive force acting in the contact area. The short range adhesive force is modeled by a delta function of a certain strength. Opposite to the case of the DMT force the JKR attractive force acts only inside the contact area. As can be seen in Fig. 10.5 the JKR contact is of hysteric nature. When the contact is established at $d = a_0$ the attractive adhesion force pulls the tip towards the sample $d < a_0$ (Fig. 10.5a) until an equilibrium with the repulsive elastic force is established. With an external applied force the tip indents further into the sample (Fig. 10.5b). If the tip is retracted by an externally applied force of opposite direction, a bridge remains for $d > a_0$, until a maximum (separation) force is reached, beyond which tip suddenly breaks free from the sample. This whole hysteretic process is also shown in a qualitative plot of the external applied force as function of the tip-sample distance d, with the situations shown in (a)–(c) indicated. The equations describing the JKR model can be found in the following references [3, 12–14]. The JKR model describes best contacts of soft materials with high surface energies and tips with large radii [11, 14].

The DMT and the JKR models describe the limiting cases of long range forces acting solely outside of the contact area, and forces acting inside the contact area, respectively. A model covering the intermediate regime was developed by Maugis using the Dugdale approximation (MD model) [9, 15]. In this model a square shaped model force for the tip-sample force is assumed having a certain width and force magnitude, i.e. the adhesion force remains constant up to a certain tip-sample distance and vanishes beyond that distance within the MD model. The JKR case can be modeled by a very narrow and deep force (close to a delta function), while the DMT force can be approximated by a square shaped force with the range of the van der Waals force and a depth corresponding to an average attractive force. Additionally, any intermediate square force shapes can be considered. The Maugis Dugdale model (MD model) is considered in detail in the following references [3, 9, 14].

10.3 Capillary Tip-Sample Forces

Capillary forces are an important issue when performing atomic force microscopy in air. It is known that under ambient conditions a thin water film can exist on the sample and the tip [16]. The thickness of this water layer depends on the relative humidity and can range from below 1 nm to several nanometers [16]. When tip and sample come so close that both water layers touch, a meniscus forms between tip and sample, as shown in Fig. 10.6 for the case of a hydrophilic tip and sample. This capillary force is of hysteretic nature. When the tip and sample approach, there is initially no water meniscus present. If the water films of tip and sample touch, a meniscus forms between tip and sample, as shown in Fig. 10.6b. If the tip-sample gap is increased subsequently, (Fig. 10.6c) the meniscus breaks at a much larger tip-sample distance than it formed. This gives rise to the hysteretic nature of the capillary force. The capillary force is an attractive tip-sample force: If tip and sample separate, the surface area of the meniscus increases. This corresponds to a higher surface energy and thus to an attractive tip-sample force.

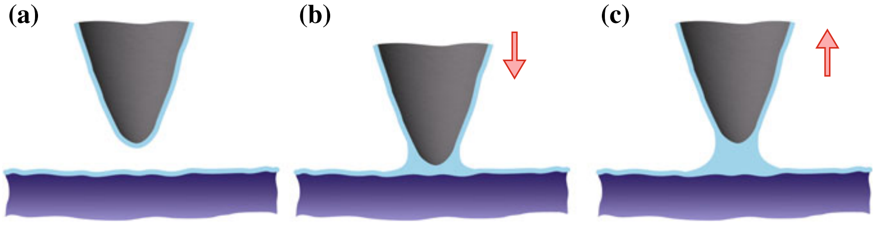


Fig. 10.6 **a** At ambient conditions tip and sample are covered by a thin water layer. **b** If tip and sample touch, the tip-sample gap fills with water, either due to the water films on both or due to capillary condensation **c** The water meniscus between tip and sample remains also if tip and sample disengage, leading to the hysteretic nature of the capillary force

In contact mode AFM the capillary force is an additional attractive force on the order of several nN. This attractive force component modifies the force-distance curves, discussed in Sect. 12.5. Capillary forces modify also the dynamic behavior of the cantilever oscillation, as discussed in [17].

Additionally to the presence of a thin water layer on tip and sample, a liquid meniscus can also form due to capillary condensation at the tip sample contact [18]. Capillary condensation is an effect where in a small confined volume water condenses at a lower vapor pressure than in a large volume. The gap between the tip and the sample acts as a small confined volume (like a capillary) in which water can condense at ambient conditions.

There are models which calculate the capillary force [16–18]. However, since most of the involved parameters (tip radius, tip shape, thickness and composition of the liquid contamination layer, surface energy of the liquid layer and surface energies and of tip and sample) are unknown, a reliable calculation of the capillary forces is difficult. Moreover in the dynamic mode the time dependence of the formation of the meniscus adds another complexity.

10.4 Electrostatic Tip-Sample Force

A further kind of tip-sample interaction is the electrostatic interaction, which is quite long-range. It appears if there are static electric charges trapped on the tip or sample, or if the tip and sample are conductive and are at different potentials. When we consider the tip-sample system as a capacitor with distance dependent capacitance $C(z)$, the energy change of a capacitor induced by a voltage difference of ΔV is given by $E_{\mathrm{el}}(z, \Delta V) = -1/2\, C(z) \Delta V^2$. The electrostatic force is then given by

$$F_{\mathrm{el}}(z, \Delta V) = -\frac{\partial E_{\mathrm{el}}(z)}{\partial z} = \frac{1}{2} \frac{\partial C(z)}{\partial z} \Delta V^2. \tag{10.11}$$

Using this equation, we will evaluate the approximate size of the electrostatic tip-sample force. If we model the capacity between tip and sample by a plate capacitor (plate area A) with capacitance

$$C_{\mathrm{plate}}(z) = \epsilon_0 \epsilon_r \frac{A}{z}, \tag{10.12}$$

the $1/z$ tip-sample distance dependence of the capacity results in a force proportional to $1/z^2$. If the tip is modeled more realistically by a sphere on a cone [19], and the sample by a semi-infinite conductive solid, the electrostatic force between tip and sample results as

$$F_{\mathrm{el}} \approx -\pi \epsilon_0 \epsilon_r \frac{R_{\mathrm{tip}}}{z} \Delta V^2. \tag{10.13}$$

For a tip radius $R_{\text{tip}} = 50\,\text{nm}$, a tip-sample distance of $z = 1\,\text{nm}$, and a voltage of $V = 1\,\text{V}$ a force of about $F_{\text{el}} \approx 1\,\text{nN}$ results. This value is similar to short-range forces occurring between individual atoms. The force can be even larger, since due to the long-range of the electrostatic force also the interactions between the sample and more distant parts of the cantilever might be important.

While the electrostatic force can have considerable values, it vanishes according to (10.11) if $\Delta V = 0$. The potential difference ΔV is determined by two aspects, the bias voltage applied between tip and sample V_{bias} as well as the difference of the work functions between tip and sample (local contact potential difference) as $\Delta V = V_{\text{bias}} - \Delta\Phi/e$, as we have seen in Chap. 9. Due to the work function difference, zero bias voltage does not correspond to a vanishing electrostatic force. The force as a function of the applied bias voltage is, according to (10.13), a (negative) parabola (Kelvin parabola). If tip and sample are both conducting, measuring the tip-sample force as a function of the applied bias voltage, can be used in order to determine the work function between tip and sample as the voltage at which the maximum of the parabola is reached. As long-range electrostatic forces are undesirable in atomic force microscopy the bias voltage is chosen for which ΔV and therefore the electrostatic force vanishes.

The models for the tip-sample interaction discussed above are frequently used for the interpretation of AFM force-distance curves (see Sect. 12.5). In the following text we will however use the Lennard-Jones potential as a model potential, because of its simple analytic form.

10.5 Snap-to-Contact

For a soft cantilever, atomic force microscopy is accompanied by the so-called “snap-to-contact”. To introduce this effect let us discuss a macroscopic example. In the case of a magnet attached to a spring, the magnet will have a stable position in the gravitational field of the earth. If you bring the magnet close to an iron containing plate, the attractive magnetic force will stretch the spring further. The system goes to a new equilibrium position; an equilibrium position can be verified by exciting small oscillations of the magnet around its equilibrium position. However, if the magnet is brought too close to the iron plate, the magnet will snap onto the metal plate. The spring can no longer keep the magnet in a stable position. This snap-to-contact effect in which the system changes its state instantaneously is also observed in AFM. Control over the position of the tip is lost so that certain tip-sample positions cannot be realized.

Now that you have some idea of what snap-to-contact means, we will analyze the stability of a (cantilever) spring system if an outer (tip-sample) potential is added. The total potential energy of the cantilever system consists of two contributions, as shown in Fig. 10.7: (a) The potential between tip and sample U_{ts}, which we model here as a Lennard-Jones potential (with the parameters U_0 and z_a corresponding to the depth of the potential and the distance for which the potential is zero, respectively),

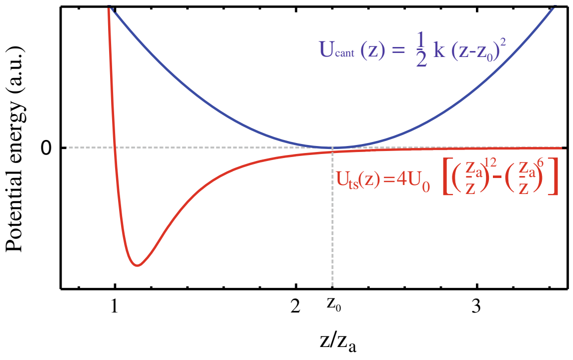


Fig. 10.7 Graphic representation of the two potentials acting on the cantilever: the tip-sample potential modeled by a Lennard-Jones potential and the parabolic potential arising due to the cantilever spring constant

and (b) the parabolic potential U_{cant} arising due to the spring constant k of the AFM cantilever.[1]

The total potential energy of the cantilever-tip-sample system can be written as the sum of both contributions

$$U_{\mathrm{tot}}(z) = U_{\mathrm{ts}}(z) + U_{\mathrm{cant}}(z) = 4U_0 \left[\left(\frac{z_a}{z} \right)^{12} - \left(\frac{z_a}{z} \right)^{6} \right] + \frac{1}{2} k \left(z - z_0 \right)^2 . \quad (10.14)$$

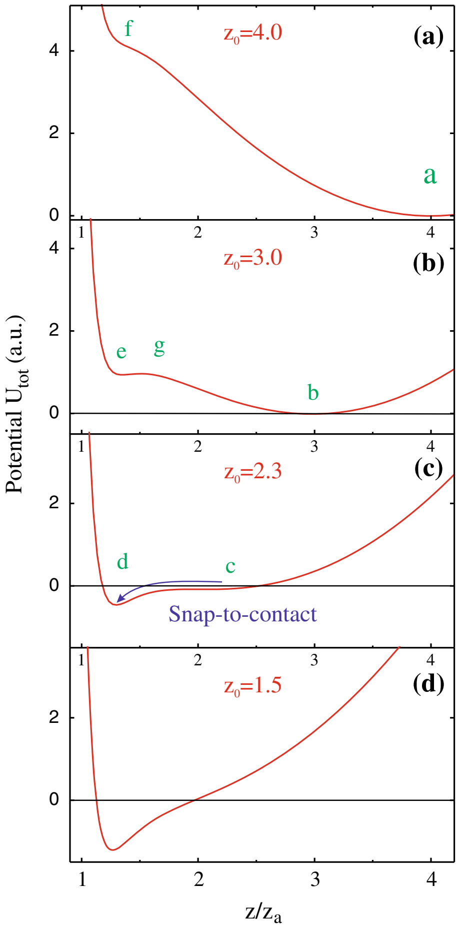
The variable z is the distance between the origin of the Lennard-Jones tip-sample potential (i.e. sample surface) and of the tip position. The parameter z_0 is the distance from the origin to the equilibrium position of the cantilever tip without any influence from the tip-sample potential (tip-sample potential switched off). The distance z_0 can be varied via the piezo element controlling the tip-sample distance, while the actual tip-sample distance at which the potentials are evaluated is z. The bending of the cantilever due to the tip-sample interaction force is $z - z_0$.

Since the interactions are modeled by potentials, they are considered as conservative interactions, i.e. without dissipative interactions. Generally the system "tries" to minimize the total potential energy by realizing the tip-sample distance z, which corresponds to the lowest U_{tot}. If the tip is oscillating, the oscillation occurs around this equilibrium position.

The lowest potential (global minimum) may not be reached due to a barrier present between the nearest local minimum and the global minimum of the total potential of the system. A graphic representation of the total potential of the cantilever (sum of the tip-sample potential and cantilever potential) is given in Fig. 10.8 for different values of the parameter z_0. If the cantilever tip is far from the surface (corresponding to large values of z_0), the spring potential provides a stable potential minimum at $z \approx z_0$ (Fig. 10.8a, b). In fact, the minimum is at slightly smaller z values than z_0 due to the non-zero attractive interaction potential between tip and sample. If the cantilever comes closer to the surface (smaller values of z_0), the potential minimum close to z_0

[1] We use here the coordinate z for the distance between the tip and sample instead of r previously used for the Lennard-Jones potential between two atoms in (10.5).

Fig. 10.8 Graphic representation of the total potential (tip-sample plus cantilever spring potential according to (10.14)) as a function of the tip-sample distance z. The potential is shown for different values of z_0, decreasing from **a** to **d**, as the tip approaches the surface. The parameter z_0 is the equilibrium position of the cantilever tip without any influence from the tip-sample potential. For large distances of the tip from the surface, the tip is in a stable potential minimum close to z_0 as shown in **a** and **b**. As the tip approaches the sample the potential minimum close to z_0 converts to a saddle point (**c**). Below a critical distance between tip and sample the tip snaps to a new minimum close to the sample, dominated by the tip-sample interaction (**d**)



vanishes (converts to a saddle point) due to the increased interaction strength of the tip-sample potential for smaller tip-sample distances (Fig. 10.8c). Correspondingly, the cantilever tip will find a new stable minimum not close to z_0 but closer to the sample surface (Fig. 10.8d). This abrupt jump of the cantilever equilibrium position to a position much closer to the surface is called snap-to-contact.

In the contact mode of AFM, the measurements are performed with the tip snapped into contact, i.e. in a regime in which the repulsive tip-sample interaction prevents any further approach toward the surface. In dynamic AFM measurements (with an oscillating cantilever) snap-to-contact would stop the oscillation due to the very narrow potential minimum close to the surface. Thus, in the dynamic mode the snap-to-contact has to be prevented and in the following we will analyze the conditions under which the snap-to-contact can be prevented.

We will determine at which tip-sample distance(s) z the total potential $U_{\mathrm{tot}}(z)$ has minima (for a given value of the parameter z_0). Specifically it is important to know under which conditions a minimum vanishes.[2] A necessary condition for a minimum of $U_{\mathrm{tot}}(z)$ is that the first derivative of the potential with respect to z has to be zero ($\frac{\partial U_{\mathrm{tot}}}{\partial z} = 0$), which means that

$$\frac{\partial U_{\mathrm{ts}}}{\partial z} + k(z - z_0) = 0. \tag{10.15}$$

Since $-\frac{\partial U_{\mathrm{ts}}}{\partial z} = F_{\mathrm{ts}}$, the above condition is actually a condition of force balance

$$F_{\mathrm{ts}}(z) = -F_{\mathrm{cant}}(z, z_0). \tag{10.16}$$

This balance of forces is graphically represented in Fig. 10.9, with the force due to the cantilever bending F_{cant} represented by straight lines (Hooke's law: $F_{\mathrm{cant}} = -k(z - z_0)$) for different positions of the free cantilever zero point z_0. The slope of the cantilever force lines corresponds to the spring constant k. In this graph, a force equilibrium ($F_{\mathrm{ts}}(z) = -F_{\mathrm{cant}}(z)$) occurs if the red line corresponding to the Lennard-Jones force crosses one of the straight lines representing the (negative) cantilever spring force. It can be seen from Fig. 10.9 that for each position of z_0 one (or more) distances z can be found for which the force balance (10.16) holds.

The force equilibrium (the first derivative of the potential vanishes) identifies only the critical points (minima, maxima, and saddle points). The second (sufficient) condition for stability of the cantilever (potential minimum) is that the second derivative of the total potential with respect to z has to be larger than zero ($\frac{\partial^2 U_{\mathrm{tot}}}{\partial z^2} > 0$, positive curvature). This second condition can be written as

$$\frac{\partial^2 U_{\mathrm{ts}}}{\partial z^2} + k > 0. \tag{10.17}$$

Since $F_{\mathrm{ts}} = -\frac{\partial U_{\mathrm{ts}}}{\partial z}$, this condition can be expressed in terms of the force gradient as

$$k > \frac{\partial F_{\mathrm{ts}}}{\partial z}. \tag{10.18}$$

If tip and sample are still far from each other, the minimum of the potential is at a cantilever position z close to z_0 (Fig. 10.8) and the condition (10.18) is fulfilled at the position of force equilibrium, since the force gradient is very small for large z. If the tip and sample approach each other, this condition of stability holds (at the position of equilibrium) until the force gradient becomes larger than the spring constant (which is the negative gradient of the cantilever force $k = -\frac{\partial F_{\mathrm{cant}}}{\partial z}$). If (10.18) is no longer fulfilled, the potential minimum vanishes, and the spring system becomes

[2] In our analysis we treat the spring constant k and the parameters of the Lennard-Jones potential (U_0 and z_a) as constants.

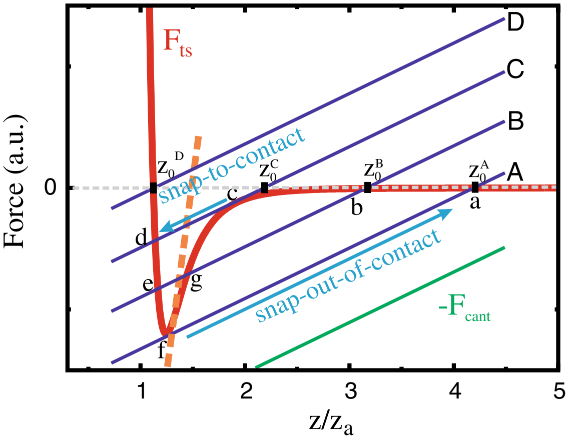


Fig. 10.9 Comparison of the tip-sample force (approximated by a Lennard-Jones type force) to the negative cantilever spring force (straight lines for different z_0, i.e. for different externally set tip-sample distances). If the two forces are the same (point(s) of intersection), a minimum, maximum or saddle point is present in the potential curve (compare Fig. 10.8). The cantilever spring constant k corresponds to the slope of the straight lines. When tip and sample approach each other (decreasing z_0), the gradient of the tip-sample force (slope of the red curve) exceeds k (slope of the blue lines) and a transition from stability (potential minimum) to instability occurs (point c). The tip jumps from (point c) to the stable minimum at point d (snap-to-contact). Correspondingly, snap-out-of-contact occurs at point f where the slope of the Lennard-Jones potential becomes larger than the slope k of the cantilever spring force

instable and snaps to contact (Fig. 10.8c). In the graphic representation in Fig. 10.9 this stability condition holds, if the slope of the tip-sample force F_{ts} (red curve in Fig. 10.9) is smaller than the slope (gradient) of the cantilever force (straight blue lines in Fig. 10.9).

After considering the equations governing snap-to-contact, we will now follow the snap-to-contact effect step by step using Figs. 10.8 and 10.9. For large values of z_0 the tip-sample force can be neglected at the point of equilibrium, which is very close to z_0 (point a in Figs. 10.8a and 10.9). When the cantilever approaches the surface (line B in Fig. 10.9), the cantilever spring force compensates the tip-sample force at the three intersection points b, g, and e in Fig. 10.9. The points b and e correspond to the two minima indicated in Fig. 10.8b while g corresponds to the potential maximum in between. Since the tip started in the right potential minimum it will stay there, even if the minimum close to the surface becomes lower, as there is a potential barrier in between. However, if the tip moves further towards the surface, minimum b and maximum g approach each other and eventually form the saddle point c (line C in Fig. 10.9, compare also Fig. 10.8c). Now the position of the cantilever becomes instable and the cantilever moves to the other minimum d closer to the surface. This is the snap-to-contact. A further shift of the zero position of the tip z_0 towards the surface will change the position of the minimum only slightly due to the large slope of the tip-sample potential. The intersection with line D occurs almost at the same z-position as the intersection with line C in Fig. 10.9.

When the tip is subsequently retracted from the sample, it remains in the potential minimum close to the surface even when the other potential minimum is re-established (point *b* in Fig. 10.8b). Finally, minimum *e* and maximum *g* develop into a saddle point *f* and the tip snaps out of contact into the minimum at point *a* (line A in Fig. 10.9, compare also Fig. 10.8a). This instantaneous jump is called snap-out-of-contact.

Since the snap-to-contact effect is undesirable in dynamic atomic force microscopy, we will now discuss the conditions under which it can be prevented. One strategy is to avoid the snap-to-contact effect by using cantilevers with a large spring constant. If k is larger than the maximal value of the gradient (slope) of the tip-sample force, (10.18) is always fulfilled, i.e. for any value of z_0. This corresponds in Fig. 10.9 to the orange line which has a larger slope than the maximum of the slope of the tip-sample force and thus snap-to-contact is avoided.

Apart from using cantilevers with a high force constant there is another experimental condition under which snap-to-contact can be avoided. This condition can be realized if the cantilever is oscillated around its equilibrium position, i.e. in the dynamic mode of AFM operation. First, the equilibrium tip-sample distance z_0 should be large, which corresponds for instance to the green curve in Fig. 10.9. As a second condition, the oscillation amplitude should be large in order to reach the region very close to the sample (where the tip-sample interaction is different from zero) at least at the turnaround point of the oscillation closest to the sample. The green line will never cross the red line in Fig. 10.9 (apart from a point very close to z_0). Due to the large deflections for tip positions close to the surface, the cantilever force is always larger than the attractive tip-sample force and thus snap-to-contact is prevented. In summary, the conditions of a large oscillation amplitude and simultaneously a large z_0 prevent snap-to-contact and maintain the condition of stability, also for the case of small cantilever force constants k.

10.6 Summary

- The long-range attractive van der Waals force and the short-range forces, such as chemical bonding forces and the Pauli repulsion, contribute to the tip-sample interaction.
- In order to represent the different forces in a simple analytic form the Lennard-Jones potential is used as a model potential comprising an attractive part $\propto -1/r^6$ and a repulsive part $\propto 1/r^{12}$.
- The Hertz model describes the elastic interaction of the tip with the sample. Additional attractive (van der Waals or adhesion) forces are considered in the JKR, DMT, and MD models.
- The electrostatic forces and capillary forces are other important tip-sample forces.
- If the cantilever tip is brought towards the sample an instability can occur if the force gradient of the tip-sample interaction becomes larger than the spring constant of the cantilever $\frac{\partial F_{\mathrm{ts}}}{\partial z} > k$. In this case snap-to-contact occurs and the tip jumps toward the surface.

- Snap-to-contact can be prevented by (a) stiff cantilevers or (b) in the dynamic mode by large oscillation amplitudes keeping the cantilever force larger than the tip-sample force.

References

1. J. Cugnon, The Casimir effect and the vacuum energy: duality in the physical interpretation. Few-Body Syst. **53**, 181 (2012). https://doi.org/10.1007/s00601-011-0250-9
2. J. Israelachvili, *Intermolecular and Surface Forces*, 3rd edn. (Academic Press, London, 2011). ISBN 9780123919274
3. R. Reifenberger, *Fundamentals of Atomic Force Microscopy: Part I: Foundations* (World Scientific, Singapore, 2015). https://doi.org/10.1142/9343
4. H. Hertz, Über die Berührung fester elastischer Körper (On the contact of elastic solids). J. Reine Angew. Math. **92**, 156 (1881)
5. B.V. Derjaguin, V.M. Muller, Y.P. Toporov, Effect of contact deformations on the adhesion of particles. J. Colloid Interface Sci. **53**, 314 (1975). https://doi.org/10.1016/0021-9797(75)90018-1
6. B.V. Derjaguin, V.M. Muller, Y.P. Toporov, On the role of molecular forces in contact deformations (critical remarks concerning Dr. Tabor's report). J. Colloid Interface Sci. **67**, 378 (1978). https://doi.org/10.1016/0021-9797(78)90021-8
7. B.V. Derjaguin, V.M. Muller, Y.P. Toporov, On different approaches to the contact mechanics. J. Colloid Interface Sci. **73**, 293 (1980). https://doi.org/10.1016/0021-9797(80)90157-5
8. V.M. Muller, B.V. Derjaguin, Y.P. Toporov, On two methods of calculation of the force of sticking of an elastic sphere to a rigid plane. Colloids Surf. **7**, 251 (1983). https://doi.org/10.1016/0166-6622(83)80051-1
9. D. Maugis, Adhesion of spheres: the JKR-DMT transition using a dugdale model. J. Colloid Interface Sci. **150**, 243 (1992). https://doi.org/10.1016/0021-9797(92)90285-T
10. B. Derjaguin, Untersuchungen über die Reibung und Adhäsion, IV (Investigations on friction and adhesion). Kolloid Z. **69**, 155 (1934). https://doi.org/10.1007/BF01433225
11. D. Tabor, Surface forces and surface interactions. J. Colloid Interface Sci. **58**, 2 (1977). https://doi.org/10.1016/0021-9797(77)90366-6
12. K.L. Johnson, K. Kendall, A.D. Roberts, Surface energy and the contact of elastic solids. Proc. R. Soc. Lond. Ser. A **324**, 301 (1971). https://doi.org/10.1098/rspa.1971.0141
13. V.M. Muller, V.S. Yuschenko, B.V. Derjaguin, On the influence of molecular forces on the deformation of an elastic sphere and its sticking to a rigid plane. J. Colloid Interface Sci. **77**, 91 (1980). https://doi.org/10.1016/0021-9797(80)90419-1
14. U.D. Schwarz, A generalized analytical model for the elastic deformation of an adhesive contact between a sphere and a flat surface. J. Colloid Interface Sci. **261**, 99 (2003). https://doi.org/10.1016/S0021-9797(03)00049-3
15. D.S. Dugdale, Yielding of steel sheets containing slits. J. Mech. Phys. Sol. **8**, 100 (1960). https://doi.org/10.1016/0022-5096(60)90013-2
16. D.B. Asay, S.H. Kima, Effects of adsorbed water layer structure on adhesion force of silicon oxide nanoasperity contact in humid ambient. J. Chem. Phys. **124**, 174712 (2006). https://doi.org/10.1063/1.2192510
17. L. Zitzler, S. Herminghaus, F. Mugele, Capillary forces in tapping mode atomic force microscopy. Phys. Rev. B **66**, 155436 (2002). https://doi.org/10.1103/PhysRevB.66.155436
18. T. Ondarcuhu, L. Fabiein, in *Surface Tension in Microsystems, Microtechnology and MEMS*, ed. by P. Lambert (Springer, Berlin, 2013). https://doi.org/10.1007/978-3-642-37552-1_14
19. M. Saint Jean, S. Hudlet, C. Guthmann, J. Berger, Van der Waals and capacitive forces in atomic force microscopies. J. Appl. Phys. **86**, 5245 (1999). https://doi.org/10.1063/1.371506

Chapter 11
Cantilevers and Detection Methods in Atomic Force Microscopy

We consider basic requirements for force sensors and introduce a fabrication process for cantilevers. Subsequently, the most common detection method for measuring the cantilever deflection, the beam deflection method, is discussed in detail. Other detection methods are presented only briefly, before calibration measurements for AFM are described. First the sensitivity factor has to be determined. This gives the conversion from the measured sensor voltage (at the output of the deflection measurement electronics) to the actual deflection of the cantilever tip in nanometers. Subsequently, several methods for the determination of the spring constant of the cantilever are discussed.

11.1 Requirements for Force Sensors

When we discuss the requirements for force sensors, the first question is: How strong are the forces we would like to measure? The forces between atoms in solids can be used as a first estimate for the expected tip-sample forces. Typical vibration frequencies of atoms in a solid are $\omega_{\mathrm{vib}} = 10^{13}\,\mathrm{Hz}$ and typical atom masses are of the order of $m = 10^{-25}\,\mathrm{kg}$. Considering the vibrations of the atoms in the model of a harmonic oscillator the well-known relation

$$\omega_{\mathrm{vib}} = \sqrt{\frac{k}{m}} \tag{11.1}$$

can be applied. Thus, the spring constant for the bonds of atoms in a solid results as

$$k = \omega_{\mathrm{vib}}^2 m \approx 10\,\mathrm{N/m} \tag{11.2}$$

© Springer Nature Switzerland AG 2019
B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3_11

With this force constant of 10 N/m and distances between the atoms in the ångström range (10^{-10} m) forces between atoms in the nanonewton regime can be expected following Hooke's law. Another crude way to estimate the forces on the atomic scale is to divide typical bond energies of the order of electron volt by distances of the order of ångströms, resulting in forces between atoms of the order of nanonewtons as well. This sets a limit for the maximum force which should be exerted by the tip on the surface atoms. Much larger forces than nanonewtons can lead to the breaking of the bonds of the surface atoms, which leads to undesired damage to the surface structure, which should be measured nondestructively.

If a cantilever with a spring constant of 1 N/m is used to measure forces in the nanonewton regime, the bending of the cantilever due to a nN force will be in the nanometer regime, which is still detectable as we will see later. For a given detection limit of the cantilever deflection measurement Δz a desirable high force sensitivity ΔF calls, due to Hook's law, for a small force constant in static AFM as $\Delta F = k\Delta z$. Thus, for a high force sensitivity and in order to avoid too high tip-sample forces, cantilevers with a small force constant should be used. In summary, a first condition for a cantilever in static atomic force microscopy is that it should have a small spring constant.

A second requirement for the cantilever is that it should have a high resonance frequency, preferably $\gg$10 kHz. This condition results from the need to realize a high scan speed. If the surface topography is scanned, at every new image pixel the cantilever tip should move to a new height corresponding to the surface topography at this position. When analyzing the harmonic oscillator in Chap. 2 we have seen that the harmonic oscillator can only follow an external motion with gain one and without a phase shift, if the excitation frequency is (much) smaller than the resonance frequency of the harmonic oscillator. When scanning with a pixel frequency of 1 or 10 kHz, the cantilever can be excited with Fourier components up to this frequency. Thus, in order to follow the surface topography that fast, the cantilever (considered as a harmonic oscillator) should have a high resonance frequency, preferably $\gg$10 kHz.

While the above stated requirements for the cantilever were obtained for the static mode the same requirements also apply for the case of the dynamic AFM mode. As we will see in Chap. 13, the measured signal in the dynamic mode is proportional to ω_0/k. Thus, in order to obtain a large signal, a high resonance frequency and a small force constant are required as well.

Another argument for a high resonance frequency of the cantilever arises from the requirement of immunity to external vibrations for an atomic force microscope. We have seen in Sect. 3.6 that a high resonance frequency of the microscope construction is the key to immunity to external vibrations. Since the cantilever is part of the mechanical structure of the microscope also its resonance frequencies should be as high as possible, preferably $\gg$10 kHz.

Altogether we have two requirements for an AFM cantilever: high resonance frequency and small spring constant. Considering the basic equation for a harmonic oscillator

$$\omega_{\text{cant}} = \sqrt{\frac{k}{m}}, \tag{11.3}$$

the two requirements are in opposition: A small spring constant k leads to a small resonance frequency and, vice versa, a high resonance frequency leads to a high spring constant. However, both requirements can be fulfilled if the mass of the cantilever is small. We see from (11.3) that for a frequency of $\omega_{\text{cant}} = 100\,\text{kHz}$ and a force constant of $k = 10\,\text{N/m}$, a cantilever mass of $1\,\mu\text{g}$ results. Therefore, small cantilevers must be used in order to have simultaneously small spring constants (for high force sensitivity) and high resonance frequencies of the cantilever (fast scanning and good stability with respect to vibrations).

11.2 Fabrication of Cantilevers

Cantilevers are produced by semiconductor microfabrication processes using silicon or silicon nitride as materials. Silicon nitride cantilevers consist of a thin silicon nitride film deposited by chemical vapor deposition on a silicon wafer. Subsequently, the silicon is etched away in a certain region in order to expose the cantilever. The thickness of the film determines the thickness of the finished cantilever, which (for silicon nitride cantilevers) usually has a triangular shape. The triangular form of the cantilevers was chosen to prevent torsional motion due to frictional forces in contact mode AFM. Silicon nitride cantilevers have a small spring constant and are often used in contact mode atomic force microscopy. Coating the back side of the cantilevers with gold or aluminum provides high reflectivity for the optical beam deflection detection method.

The most frequently used cantilevers are silicon cantilevers. In the process described in the following, all parts of the cantilever are made of bulk silicon. The key ingredient for the fabrication of these cantilevers is anisotropic wet etching, which means that different crystal directions are etched at different rates using anisotropic etchants like KOH. The (100) direction of Si is etched much faster than the (111) direction. A simplified sketch of the fabrication process of a Si cantilever is shown in Fig. 11.1. The starting point is a Si(100)-oriented wafer on which a structured SiO_2 layer is formed as shown in Fig. 11.1a. This structured SiO_2 layer is formed by standard lithography methods used in semiconductor microelectronics, defining the cantilever shape and the tip position. A subsequent wet etching step leads to a preferred etching in the (100) direction in those areas where no SiO_2 layer is present, while the SiO_2 capped areas are not etched. Furthermore, at the edges of the SiO_2 film Si(111) facets form due to the anisotropically very slow etching speed in this direction, as shown in Fig. 11.1b. The formation of the tip is finished when the small oxide pad on top of the tip is underetched completely. Subsequently, the top of the wafer and on the bottom the handle part (cantilever base) are covered by Si_3N_4 in order to protect the tip structure (Fig. 11.1c). A further wet etching step thins the back of the cantilever beam down to the desired thickness and separates it from the Si wafer (Fig. 11.1d). Tip, cantilever and cantilever base are finished after removal of the protective silicon nitride film by a wet etching step (Fig. 11.1e). Electron microscopy images of a finished cantilever of this type are shown in Fig. 11.2. As the cantilever

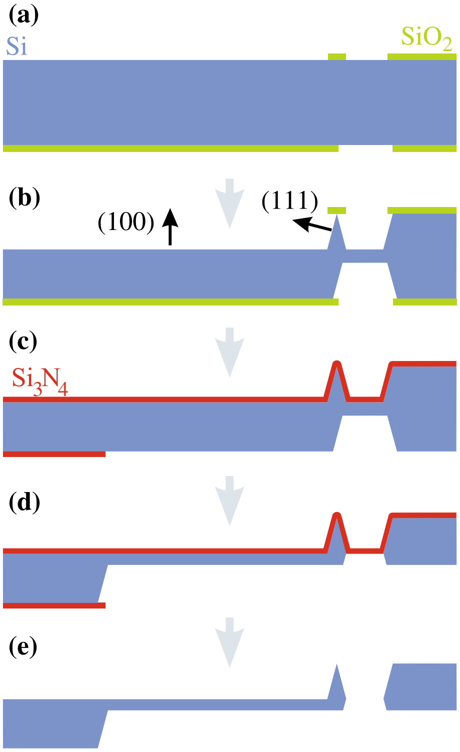


Fig. 11.1 Fabrication of a Si cantilever using alternating lithographic patterning and wet chemical etching as described in the text

beam itself is too small to handle, it is connected to a solid silicon base (cantilever chip) with dimensions of several millimeters, seen partly in the left of the images in Fig. 11.2.

The dimensions of the cantilevers can vary over large ranges, depend on the application. Typical length are 100–400 µm, typical width 20–80 µm and typical thicknesses 1–3 µm. The resonance frequencies range from a few kHz to 300 kHz. The spring constants range from very low values as 0.01 N/m for cantilevers used in static mode to about 50 N/m for cantilevers used in the dynamic mode. Recently, there is a trend to enable fast scanning in the dynamic AFM mode by cantilevers with higher frequencies beyond 1 MHz, which have short length and high spring constants.

At the tip apex, radii down to 10 nm and below can be realized for Si cantilevers. In order to realize even smaller apex radii, carbon nanotubes can be fixed to the end of the tips. Another technique to produce sharp microtips on top of Si tips is electron beam induced deposition. Here a carbon containing gas is injected into an electron microscope chamber and an electron beam is focused onto the tip. As a result of this concentrated bombardment with electrons, the gas decomposes at the tip and a sharp carbon asperity, which can have very high aspect ratio, forms on the tip. A metal containing carbonyl gas can also be used, which is decomposed by an electron beam.

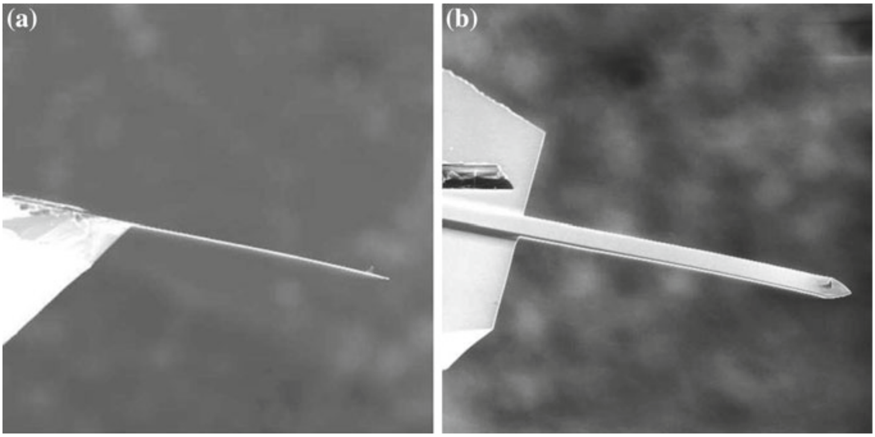


Fig. 11.2 Scanning electron microscopy images of a Si cantilever (length 450 μm) with a Si tip integrated at its end. A *side view* of the cantilever is shown in (**a**) and a *tilted view* in (**b**)

This can lead to the formation of a sharp metal whisker at the end of the tip [1]. Another technique to fabricate ultra-sharp AFM tips is sharpening using a focused ion beam (FIB).

11.3 Beam Deflection Atomic Force Microscopy

Different kinds of atomic force microscopes are characterized by the different techniques used to detect the bending of the cantilever. For most of all atomic force microscopes the beam deflection method is used. The basic setup of the beam deflection method is shown in Fig. 11.3. A laser beam from a laser diode is focused on the end of the back side of the cantilever where it is reflected into a photodiode.

The bending of the cantilever is detected by a split photodiode, i.e. two photodiodes which are separated by a small slit. The difference in the optical signals of the two parts of the split photodiode $S_A - S_B$ is proportional to the angular deflection of the laser beam and therefore proportional to the cantilever deflection (bending). The absolute intensity detected by the photodiode can vary due to fluctuations of the laser intensity and depends on the focusing of the laser beam onto the cantilever. In order to be independent of the absolute intensity of the signal the normalized intensity is used $(S_A - S_B)/(S_A + S_B)$. The beam deflection method requires a mirror-like surface at the back of the cantilever. Additionally, the width of the focused laser beam on the cantilever must be wide enough to reflect the light without too much diffraction. This is necessary since the diameter of the beam on the photodiode should be smaller than the active diameter of the photodiode. In atomic force microscopy setups with beam deflection detection, it is usually the sample that is scanned and not

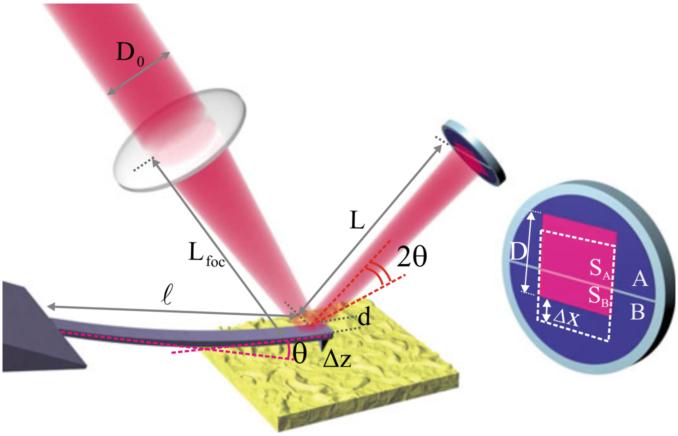


Fig. 11.3 Schematic of the beam deflection AFM method including the relevant dimensions necessary to calculate the sensitivity. The bending of the cantilever end by the angle θ results in an angle of 2θ for the reflected beam. The inset on the right shows the split photodiode (blue) and the laser beam on the photodiode (magenta), assumed as a square for the sake of simplicity

the tip. This is done because when scanning the cantilever (without moving the laser beam simultaneously) the laser spot would (in part) no longer focus on the cantilever.

11.3.1 Sensitivity of the Beam Deflection Method

In the following, the sensitivity of the optical beam deflection method is analyzed, i.e. the relation between the deflection of the cantilever Δz and the output signal of the photodetector electronics. Primarily the output signal of a photodiode is a current I, which is converted to a (proportional) voltage at the output of the photodiode preamplifier electronics using a transimpedance amplifier.

In the following, we will estimate the signal (current I) in the photodiode. The difference of electric currents of the photodiode segments A and B at the output of the photodiode I is proportional to the optical signal $S_A - S_B$ as $I = R(S_A - S_B)$, with R being the sensitivity (response) of the photodiode: output current divided by input optical power in ampere per watt. We assume a total optical power of the laser diode of $S_0 = S_A + S_B$ and estimate (following Fig. 11.3) how the reflected beam moves on the photodiode for a certain deflection of the cantilever Δz.

Analyzing the mechanics of the bending of beams it can be shown that the height change Δz and the deflection angle θ at the free end of the beam with length l are related by [2]

$$\theta = \frac{3}{2}\frac{\Delta z}{l}. \tag{11.4}$$

This angle is a factor $3/2$ larger than that obtained for the rotation of a stiff beam. The laser beam of diameter D_0 is focused by a lens on the end of the back side of the cantilever. The size of this focal spot d is considered to be smaller than the cantilever width.

If we consider ray optics to determine the size of the laser spot on the photodiode D, the intercept theorem states that the laser beam diameter at the lens D_0, the focal length of the lens focusing the laser beam L_{foc}, and the length L (end of the cantilever to photodiode) are related by

$$D = D_0 \frac{L}{L_{\text{foc}}}. \tag{11.5}$$

Following this, D can be made arbitrarily small using a large focal length. However, there is a fundamental limit: D cannot be smaller than the diffraction limit. The reflected beam is actually also a diffracted beam. The spot size of the diffracted/ reflected laser beam at the photodiode D is given by diffraction ($\lambda = d \sin \alpha \approx d \cdot D/L$) as

$$D \approx \frac{\lambda L}{d}, \tag{11.6}$$

where λ is the wavelength of the laser beam and d the focused beam size on the cantilever.

In principle, the largest value for D has to be used, either limited by diffraction or from the ray optics. However, since the diffraction limit is the more fundamental limit, we will use (11.6) in the following for D.

For the sake of simplicity, we assume that the reflected laser spot on the photodiode is uniformly irradiated over a square area of dimension D with an irradiation power per area of S_{area}. We also assume that the whole diffracted beam fits in the active area of the photodiode. Then the total optical laser intensity S_0 can be written as $S_0 = S_{\text{area}} D^2$. If, more realistically, Gaussian beams are considered the numerical factors in the results change slightly.

We will not go into the details of the operation of the photodiode and merely assume that the signal current I of the photodiode is proportional to the difference of the light intensities on both parts A and B of the split photodiode $S_A - S_B$. The difference of the optical signals on both areas of the photodiode can be written according to the inset in Fig. 11.3 and using $\Delta x = 2\theta L$[1] as

$$S_A - S_B = S_{\text{area}} 2 \Delta x D = \frac{S_0}{D^2} 4 \theta L D. \tag{11.7}$$

If we insert now θ and D according to (11.4) and (11.6), the difference of the optical intensities at the photodiode results in

[1] The bending of the cantilever end by the angle θ results in an angle of 2θ for the reflected beam. Thus, the linear deflection of the reflected laser beam on the photodiode results (for small angles) as $\Delta x = 2\theta L$.

$$S_A - S_B = 6S_0 \frac{\Delta z}{l} \frac{d}{\lambda}. \tag{11.8}$$

The output current of the photodiode as a function of deflection Δz can be written using $I = R(S_A - S_B)$ as

$$I = \frac{6RS_0 d}{\lambda l} \Delta z. \tag{11.9}$$

The ratio of Δz and I is also called the detection sensitivity and is (due to the diffraction) independent of the distance between the cantilever and the photodiode. An additional factor arises if the voltage output of the preamplifier converting the photodiode current to a voltage is considered.

11.3.2 Detection Limit of the Beam Deflection Method

Up to now we have analyzed the magnitude of the photocurrent I as a function of the external conditions such as deflection of the tip, laser power, wavelength, and the geometrical parameters of the setup. In the following, the detection limit for the optical beam detection, i.e. the minimum detectable deflection Δz_{min} of the cantilever, will be analyzed. The fundamental source of noise in the beam deflection scheme is shot noise, which arises due to the discrete arrival of the photons at the photodiode. Correspondingly, the noise of the electric current in the photodiode is induced by discrete number of electrons, each generated by a photon with a probability given by the quantum efficiency (generated electrons per photon at the respective wavelength). Here we use the sensitivity of the photodiode R defined as I/S as an equivalent quantity. If we consider the optical power S_A irradiating segment A of the photodiode, the corresponding generated current is $I_A = RS_A$.

In the following, we estimate the fundamental limit in the noise of the photo current imposed by the discrete number of electrons (shot noise). An expression of this shot noise can be derived if one considers an electrical current occurring due to a discrete number of charges, n, flowing per time of measurement, Δt. If we allow for a long measurement time (averaging), say a second or so, the current will be measured with low noise, but this also means that for instance the AFM feedback can only run at this slow speed. Usually the speed of the measurement is expressed by the bandwidth, which is roughly the maximum frequency at which a signal can be detected properly, i.e. without too much loss of signal. If the duration of the measurement of the current is one second, the bandwidth is about one Hertz. If the measurement bandwidth is defined as $B = 1/\Delta t$, the measured current generated by segment A of the photo diode I_A can be written as

$$I_A = en/\Delta t = eBn. \tag{11.10}$$

If the current corresponds to n charges flowing by in the time Δt, the number of these charges will fluctuate on average by $\sqrt{n}$, leading to a current fluctuation of

$$\Delta I_{\mathrm{shot},A} = eB\sqrt{n} = eB\sqrt{\frac{I_A}{eB}} = \sqrt{eBI_A}. \tag{11.11}$$

In our simplified explanation, a numerical factor of $\sqrt{2}$ is missing. In a more rigorous derivation [3] the following equation for the shot noise of segment A results

$$\Delta I_{\mathrm{shot},A} = \sqrt{2\,e\,I_A\,B}. \tag{11.12}$$

As the noise components from the two segments of the photo diode are independent, the combined current noise of the difference signal results as

$$\Delta I_{\mathrm{shot}} = \sqrt{\Delta I^2_{\mathrm{shot},A} + \Delta I^2_{\mathrm{shot},B}} = \sqrt{2\,e(I_A + I_B)\,B} = \sqrt{2\,e\,R\,S_0\,B}, \tag{11.13}$$

as $I_A + I_B = R\,(S_A + S_B) = R\,S_0$.

Identifying the photocurrent estimated above in (11.9) as signal S and the shot noise from (11.13) as the corresponding noise N, the signal-to-noise ratio is given by

$$\frac{S}{N} = \frac{I}{\Delta I_{\mathrm{shot}}} = \frac{6dS_0R\Delta z_{\mathrm{min}}}{l\lambda\sqrt{2eS_0RB}}. \tag{11.14}$$

The smallest detectable cantilever displacement results as

$$\Delta z_{\mathrm{min}} = \frac{l\lambda}{6d}\frac{S}{N}\sqrt{\frac{2eB}{S_0R}}. \tag{11.15}$$

Now we discuss the dependence of the smallest detectable cantilever displacement Δz_{min} on the different quantities involved. A laser beam with higher intensity S_0 will improve the detection sensitivity towards smaller Δz_{min}, however this will also pump more energy into the system which can lead to thermal drift and is especially undesirable in low temperature applications. With a larger measurement bandwidth B, i.e. a shorter averaging time for the measurement, the smallest measurable deflection Δz becomes larger. S/N is the signal-to-noise ratio at which a certain feature (for instance an atomic protrusion) can be just identified. If a signal strength of one, two, or three times the noise signal is required to distinguish a signal feature from noise, the smallest detectable height of that feature Δz_{min} will increase by one, two, or three times. In this sense, the smallest detectable cantilever displacement is proportional to the signal-to-noise ratio desired in order to resolve a feature. With a larger width d of the reflected spot on the back of the cantilever, the diffraction becomes less pronounced and therefore the sensitivity increases. However, the size of the deflected beam is limited by the cantilever width. With a smaller wavelength of the laser beam, the width of the diffracted beam becomes narrower and the sensitivity increases.

For a measurement bandwidth of 1 kHz, using a red light of $\lambda = 0.7\,\mu\text{m}$ with power $S_0 = 2\,\text{mW}$, $R = 0.4\,\text{mA/mW}$ and $l/d = 10$, at a signal-to-noise ratio $S/N = 1$, the detection limit Δz_{min} of about 0.2 pm results.[2] This shows that the detection limit is quite small. The simple beam deflection technique has a very high detection sensitivity.

In Chap. 17 we will also discuss other sources of noise in the measurement of the cantilever defection, such as the amplitude of the cantilever due to thermal excitation.

11.4 Other Detection Methods

Besides the beam deflection method discussed above, also several other methods can be used to detect the deflection of the AFM cantilever, or more generally the AFM sensor. The general requirements for AFM detection methods are as follows:

- High sensitivity of the deflection measurement in the sub-ångström regime
- The measurement technique should not influence the deflection itself and should not disturb the system, for instance by heating
- The technique should be easy to operate, i.e. with a minimal amount of adjusting.

In Fig. 11.4 different methods used to measure the cantilever deflection are displayed. The most widely used technique is the beam deflection method discussed in detail in the previous section [4]. An advantage of this method is that it is easy to implement technically. A disadvantage is the need for the optical adjustment of the focused laser spot onto the backside of the cantilever and of the deflected beam onto the split photodiode.

Another optical detection scheme is interferometry [5]. Here the backside of the cantilever is used as a mirror of an optical laser interferometer. While this technique has high sensitivity it is also the experimentally most complicated. Reasonably simple setups were only implemented using fiber interferometers [6]. One advantage of this technique is the easy absolute calibration of the cantilever deflection by the wavelength of the light.

The piezoresistive detection method operates completely electrically and requires minimal experimental effort for the detection [7]. They are realized by producing a piezoresistive layer on a cantilever and are commercially available. The resistance of this layer changes when stress is applied onto the cantilever. The basic working principle of a piezoresistive sensor is as follows. When the cantilever is bent by a force acting on the tip, a mechanical stress occurs in the cantilever volume. When a resistor formed by a stripe of piezoresistive layer on the cantilever is one of the resistors in a Wheatstone bridge, the resistance of the layer on the cantilever is measured which is proportional to the stress, which is in turn proportional to the

[2] In dynamic AFM the primary bandwidth detecting the oscillatory motion of the cantilever is much larger than 1 kHz used in this quantitative example, however, in this case also the oscillation amplitudes to be detected exceed the pm range by far.

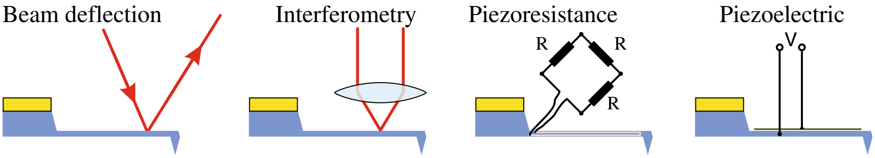


Fig. 11.4 Different kinds of deflection sensors in AFM

deflection of the cantilever. The optimal conditions for maximal device sensitivity are obtained when the Wheatstone bridge is located directly on the support wafer of the sensor. Although the signal-to-noise ratio is slightly worse than in the optical detection schemes, this is still an attractive detection scheme due to its ease of use.

Piezoelectric sensors made of quartz have recently come into use and have the specific advantage that they can be used as sensor and actuator simultaneously. They are used in dynamic AFM measurements where the sensor oscillates at the resonance frequency. The piezoelectric sensor has two electrodes. One electrode can be used to excite the sensor via the converse piezoelectric effect. The actual mechanic oscillation amplitude of the quartz sensor induces via the piezoelectric effect a voltage which is detected on the other electrode. This voltage is proportional to the deflection of the tip which is attached to the quartz sensor. We will discuss this detection scheme using quartz tuning forks and needle sensors in more detail in Chap. 18.

11.5 Cantilever Excitation in Dynamic AFM

Cantilevers have resonance frequencies of up to several hundred kHz. In order to excite such cantilevers close to their resonance frequency the piezoelectric actuator must have an even higher resonance frequency. Often this cannot be realized using the tube piezo element used for scanning, since this has too low resonance frequencies. Therefore, an additional piezo plate with a high resonance frequency is used to oscillate the cantilever base and is frequently called the dither piezo element. This type of cantilever excitation (piezoacoustic excitation) results in a cantilever oscillation amplitude A, which is, since it is close to resonance, much larger than the excitation amplitude.

While the piezoacoustic cantilever excitation is straightforward, in practice some problems can occur. The motion of the dither piezo may not only excite the cantilever, but also mechanical resonances of the AFM structure. This results in distortions and additional peaks in the resonance curve for amplitude and phase. These problems are strongest for operation with small Q-factors, i.e. in air and even more severe during operation in liquids, since in these case the resonance enhancement of the cantilever oscillation is small.

An alternative to the piezoacoustic cantilever excitation via a dither piezo element, avoiding the above mentioned problems, is the direct excitation of the cantilever. For cantilevers coated with magnetic material, an electromagnet can be used to exert an force directly to the cantilever. Another method to excite the cantilever directly is photothermal excitation [8].

In photothermal excitation, a second laser beam in addition to the one used to detect the cantilever deflection (operating at a different wavelength), is used in order to excite the cantilever closer to the base of the cantilever via the photothermal effect. The laser power heats the cantilever locally, leading to thermally induced strains which bend the cantilever. The excitation laser spot is focused close to the cantilever base in order to avoid heating of the tip. If the laser power is modulated, the cantilever will oscillate sinusoidally at this modulation frequency. Since exclusively the cantilever is excited, avoiding the excitation of other parts of the AFM structure, undistorted clean resonance curves are obtained even for small Q-factors.

Another issue is that most of the detection methods mentioned in in the previous section do actually not measure the absolute z-position of the tip, but, due to the detection of the cantilever bending, only the z-position of the tip relative to the cantilever base. This is shown schematically in Fig. 11.5, showing the reference situation with $z = 0$ and $z_{\text{drive}} = 0$ in blue. An upward motion of the cantilever base by z_{drive} will almost not be detected by the beam deflection method, as the cantilever remains unbent $z = z_{\text{drive}}$ and moves only linearly upwards.

A linear movement of the cantilever by z_{drive} leads to a movement of the laser beam on the photodiode by the same amount $\Delta x = z_{\text{drive}}$. Whereas, the same deflection z_{drive} at the end of the cantilever induced by a bending of the cantilever leads to a much larger shift of the laser beam on the photo detector Δx, as outlined in the following. The bending of the cantilever end by the angle θ results in an angle of 2θ for the reflected beam. The shift of the laser beam on the photodiode results (for small angles) as $\Delta x = 2\theta L$. Inserting θ from (11.4), results in

$$\Delta x = \frac{3L}{l}\Delta z. \tag{11.16}$$

With usual dimensions of $L \gg l$ the deflection on the photodiode Δx is more than 300 times larger than the tip deflection Δz. This means that the beam deflection method measures the component of the deflection Δz arising from a cantilever *bending*. A component of Δz arising from a linear (e.g. upwards) motion of the cantilever (without bending) is suppressed by a factor $3L/l$, which is usually greater than 300 and can be neglected.

Thus, the beam defection method detects the part of the cantilever total z-displacement that leads to a bending of the cantilever with a much higher sensitivity than a linear motion. According to Fig. 11.5 the beam deflection method detects the effective z-displacement $z_{\text{eff}} = z - z_{\text{drive}}$.

In dynamic AFM the influence of the (linear) driving motion of the cantilever is even further suppressed. For sensors with a high quality factor the oscillation amplitude A is much larger (Q-times) than the driving amplitude A_{drive}. Thus, A_{drive} can be safely neglected. However, for operation in liquids, the Q-factor is very low and the driving amplitude is comparable to the oscillation amplitude. If a detection

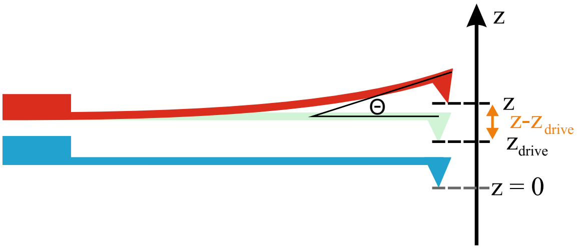


Fig. 11.5 The beam deflection method senses the cantilever bending. Therefore, a pure vertical motion of the cantilever base not leading to a bending, as indicated by the light green cantilever position relative to the blue one, is largely suppressed in the signal of the photodiode. Only the component of the vertical motion of the tip arising from the cantilever bending is sensed by the beam deflection mode. Thus, $z_{\text{eff}} = z - z_{\text{drive}}$ is sensed by the beam deflection method

method is used, which is relying on the bending of the cantilever, in a very accurate treatment the quantity z in the equation of motion (2.25) or (13.4) has to be expressed by $z_{\text{eff}} + z_{\text{drive}}$ and the equation of motion has to be solved for the measured quantity z_{eff}.

11.6 Calibration of AFM Measurements

While the relation between the cantilever deflection Δz and the tip-sample force is easily given by Hooke's law[3] as $F = k\Delta z$, there are still two calibration steps to be done. First, the signal actually measured is not the deflection Δz itself, but the sensor voltage ΔV_{sensor}, which is as a very good approximation proportional to the deflection. The constant of proportionality is called sensitivity S_{sensor} with $S_{\text{sensor}} = \Delta z / \Delta V_{\text{sensor}}$. Furthermore Hooke's law contains the spring constant k, which has to be determined in a second step. Both of these calibration steps are described in the following sections.

The above-mentioned calibration steps lead in static AFM to a calibration of the force which is important in static AFM. However, these calibration steps are also important in dynamic AFM. The spring constant of the cantilever sensor is a fundamental quantity also in the dynamic mode and the sensitivity of the sensor is needed in order to determine the oscillation amplitude in a unit of length, not just as sensor voltage.

[3] If the cantilever is tilted with respect to the surface by an angle α, the relation between the force perpendicular to the surface and the deflection perpendicular to the surface is modified [9] to $F = k\Delta z / \cos^2 \alpha$. Since α is usually small (in the range between 10° and 15°), this correction is small and will be neglected it in the following.

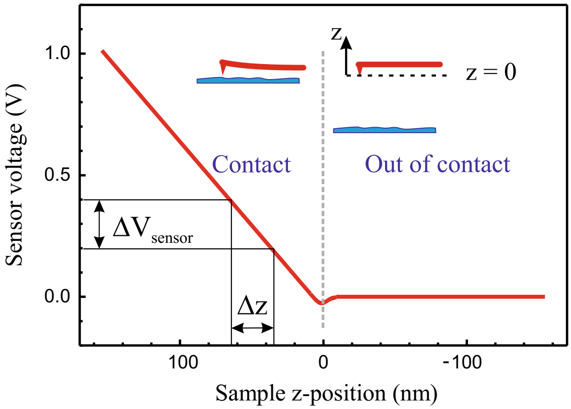


Fig. 11.6 Schematic of a typical sensor voltage versus sample z-position curve used to determine the sensitivity in AFM. The inverse slope measured in the contact regime gives the sensitivity as $S_{\text{sensor}} = \Delta z / \Delta V_{\text{sensor}}$

11.6.1 Experimental Determination of the Sensitivity Factor in AFM

The output of the detection system in atomic force microscopy is usually a voltage (sensor voltage). In the case of optical beam deflection, it is initially a current signal from the photodiodes, which is converted to a voltage by a transimpedance amplifier. Also for other detection methods, the detection system delivers a sensor voltage signal ΔV_{sensor}, which is proportional to the cantilever deflection Δz. Calibration of the sensitivity means finding this proportionality factor $S_{\text{sensor}} = \Delta z / \Delta V_{\text{sensor}}$. For the case of the beam deflection method, we found the approximate analytical expression for the detection sensitivity (11.9). However, due to the multitude of (partly unknown) parameters involved and due to the approximations made, the detection sensitivity is usually determined experimentally.

For this purpose, sensor voltage versus position curves are measured, where the sensor voltage is acquired as a function of the varying sample z-position. By applying a voltage to the z-piezo element, the sample moves relative to the tip. The z-position corresponding to a specific voltage at the z-piezo element is obtained by multiplying this voltage with the corresponding piezo constant. Such a sensor voltage versus position curve is shown schematically in Fig. 11.6 and can be roughly divided into two regions. If the tip-sample distance is large (out of contact), a negligible force acts between the tip and sample and the measured sensor voltage is independent of the sample z-position. Here the sensor voltage is set to zero. If the tip comes into contact with a hard sample, the sample bends the cantilever upwards. This upward cantilever deflection means that the corresponding sensor voltage increases linearly with the sample position.[4] From the corresponding (inverse) slope, the detection sensitivity can be obtained as $S_{\text{sensor}} = \Delta z / \Delta V_{\text{sensor}}$ in nm/V. The calibration should be performed on a hard sample with negligible elasticity, e.g. a silicon wafer. If one

[4] Specific effects occurring close to the kink between the two regions are discussed in Sect. 12.5.

is concerned about the integrity of the tip, this calibration procedure should be done after the actual measurements have been completed. The sensitivity depends on the adjustment of the optical system (laser diode, cantilever, photo diode) and should thus be repeated if something on the optical system has changed. In Sect. 11.6.5 we will introduce another method for sensor calibration in which no contact between tip and sample is made in order to obtain the sensitivity. The method of sensitivity determination outlined above applies to cantilever-type force sensors. A procedure for the determination of the sensitivity for the much stiffer quartz sensors is outlined in Chap. 18.

11.6.2 Calculation of the Spring Constant from the Geometrical Data of the Cantilever

The easiest way is to take the spring constant from the specifications of the manufacturer of the cantilever. However, often this information is not accurate enough. If the shape of the sensor is sufficiently well known, the spring constant can be calculated from the geometry of the cantilever and the elastic constants of the cantilever material. The geometric dimensions of a rectangular cantilever are introduced in Fig. 11.7. The bending of the cantilever is out of the plane of the paper in Fig. 11.7a, while in Fig. 11.7b a side view is shown. The spring constant of a rectangular cantilever beam for the bending direction used in AFM is given by [2]

$$k = \frac{Ewt^3}{4L^3}, \tag{11.17}$$

with E being Young's modulus.

While the width and length of a cantilever can be determined using a plan view optical microscope, the thickness t of the cantilever is usually much smaller and thus

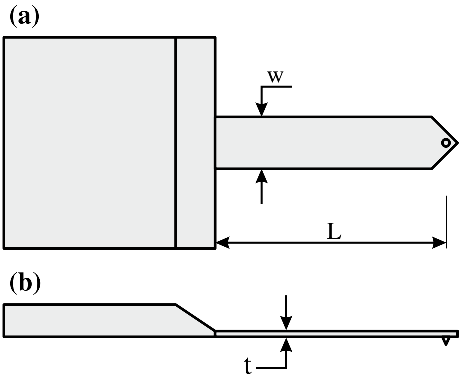


Fig. 11.7 Sketch of a rectangular cantilever together with the carrier chip on the left. **a** *Top view*, **b** *side view* including the dimensions of the cantilever

not easily measured. Unfortunately, this parameter enters with the third power into expression (11.17) for the spring constant. The thickness of the cantilever can be taken from the manufacturers specifications, or from a measurement performed with a scanning electron microscope. However, if no precise information on the thickness t of the cantilever is available, the more easily measurable resonance frequency of the cantilever can be used in order to replace t in (11.17). Considering the effective mass of the rectangular cantilever $m_{\mathrm{eff}} = 0.2357\, m_{\mathrm{spring}}$ from (2.52), the resonance frequency is written as

$$\omega_0 = \sqrt{\frac{k}{m_{\mathrm{eff}}}} = \sqrt{\frac{k}{0.2357\rho L w t}}. \tag{11.18}$$

Combining (11.18) and (11.17), t can be eliminated and the following expression for the spring constant is obtained

$$k = 0.239 w L^3 \omega_0^3 \sqrt{\frac{\rho^3}{E}}. \tag{11.19}$$

This approach to eliminate quantities which are not precisely known by other given or measured quantities can be extended as done in the next section for Young's modulus E. It is useful to replace Young's modulus because it can vary from cantilever to cantilever. For silicon nitride as a compound material, Young's modulus varies depending on the material composition, i.e. on the parameters used during the chemical vapor deposition process. Also the metallic coating on the back side of the cantilever, used for better reflection of the laser beam modifies the Young's modulus of the cantilever.

One reason why the calculation of the spring constant from the geometrical data is not so precise is that it is based on the equation of a rectangular beam (11.17), while in reality the cantilever narrows towards the end, and also the cross section of real cantilevers are not always rectangular. Moreover, the mass of the tip at the end and the not completely rigid fixing of the cantilever at the base make the application of the ideal equation (11.17) imprecise. Therefore, other methods for the determination of the spring constant of the cantilever will be discussed in the following.

11.6.3 Sader Method for the Determination of the Spring Constant of a Cantilever

If the damping of the cantilever in the fluid surrounding the cantilever during its oscillation is considered, the spring constant for a rectangular cantilever can be calculated including the (easily measurable) parameters[5] ω_0 and Q, while excluding t and E [10]. The spring constant results as

$$k = 0.19\rho_f w^2 L Q_f \Gamma_i(Re)\omega_{0,f}^2. \tag{11.20}$$

Here ρ_f is the density of the fluid surrounding the cantilever (usually air), while $\omega_{0,f}$ and Q_f are the resonance frequency and the quality factor of the free cantilever in the presence of the fluid. This equation assumes that the quality factor is much larger than one. The quantity $\Gamma_i(Re)$ is the imaginary part of the hydrodynamic function, as described and shown in Fig. 1 of [10]. The hydrodynamic function is a function of the Reynolds number, which is defined as $Re = \rho_f w^2 \omega_{0,f}/(4\eta)$, with η being the viscosity of the fluid.[6] There is also a relevant web app to calculate the spring constant using the Sader method. The spring constant of triangular cantilevers is related to the spring constant of rectangular cantilevers as described in [11, 13].

11.6.4 Thermal Method for the Determination of the Spring Constant of a Cantilever

Hutter and Bechhoefer proposed another method for the determination of the spring constant of a cantilever [14]. Unlike the Sader method, it is not named after the inventors, but rather called the “thermal method” for the determination of the spring constant and relies on the measurement of the thermal noise of the cantilever. The principle of this method is based on the equipartition theorem. According to this, the thermal noise of the amplitude Δz_{th} of an ideal harmonic oscillator is related to its static spring constant k by [3]

$$\frac{1}{2}k\left\langle \Delta z_{\mathrm{th}}^2 \right\rangle = \frac{1}{2}k_{\mathrm{B}}T, \tag{11.21}$$

with $\left\langle \Delta z_{\mathrm{th}}^2 \right\rangle$ being the mean square of the thermal amplitude fluctuations of the oscillator. In applying this to the case of a cantilever, the mean-square displacement of the free cantilever has to be measured in order to determine the spring constant. In principle, this can be done by monitoring the time behavior of the deflection (squared) for a free cantilever, i.e. far from the surface. However, such measurements are performed in practice using the power spectral density of the cantilever, as shown below. An advantage of the thermal method is that it can be applied to a free cantilever not in contact with the sample.

[5] The parameters ω_0 and Q can be obtained by measuring a resonance curve of the cantilever in response to an external excitation (frequency sweep over the resonance). Alternatively, the thermal noise spectrum can be measured, as described in the next section.

[6] The density and viscosity for the most frequently used fluids (air and water) are: $\rho_{\mathrm{air}} = 1.2\,\mathrm{kg/m^3}$, $\eta_{\mathrm{air}} = 1.85 \times 10^{-5}\,\mathrm{kg/(m\,s)}$, and $\rho_{\mathrm{water}} = 1 \times 10^3\,\mathrm{kg/m^3}$, $\eta_{\mathrm{water}} = 8.9 \times 10^{-4}\,\mathrm{kg/(m\,s)}$, respectively, under ambient conditions and at sea level [12].

In the following, we discuss how the spring constant k can be obtained from the thermal deflection noise density of the cantilever considered as simple harmonic oscillator, i.e. only the fundamental mode is considered. In the time domain the thermal cantilever noise is described by the deflection $\Delta z(t)$, while in the frequency domain the corresponding quantity is the power spectral density (PSD) of the cantilever noise $N^2_{\mathrm{z,th}}(f)$, as introduced in Chap. 5. The noise spectral density is $N_{\mathrm{z,th}}(f) = \sqrt{N^2_{\mathrm{z,th}}(f)}$. In the following, we assume that the noise power spectral density has been measured (by Fourier transformation of the time signal) using a spectrum analyzer.[7] The power spectral density of the thermal noise has as function of frequency a resonance peak behavior known as thermal peak. In Sect. 17.1 it is shown that the thermal noise spectral density of a harmonic oscillator can be written (after the subtraction of a constant background, arising from sensor displacement noise) as

$$N^2_{\mathrm{z,th}}(f) = N^2_{\mathrm{z,th,exc}} G^2(f) = \frac{N^2_{\mathrm{z,th,exc}}}{\left(1 - \frac{f^2}{f_0^2}\right)^2 + \frac{1}{Q^2}\frac{f^2}{f_0^2}}, \tag{11.22}$$

with $N^2_{\mathrm{z,th,exc}}$ being the (frequency-independent) white noise arising from the thermal excitation. From a fit of this function to the experimentally measured noise density, the parameters $N^2_{\mathrm{z,th,exc}}$, Q, and f_0 can be determined. The integral over $G^2(f)$ can be calculated and results as $\pi Q f_0/2$ (compare Sect. 17.1). Thus, using (5.13) and (11.21) the following relation results

$$\left\langle \Delta z^2 \right\rangle = \int_0^\infty N^2_{\mathrm{z,th}}(f)\mathrm{d}f = N^2_{\mathrm{z,th,exc}} \frac{\pi Q f_0}{2} = \frac{k_{\mathrm{B}} T}{k}. \tag{11.23}$$

With this, the spring constant of the simple harmonic oscillator considered here results as

$$k = \frac{2 k_{\mathrm{B}} T}{\pi N^2_{\mathrm{z,th,exc}} Q f_0}. \tag{11.24}$$

Importantly, this thermal method for the determination of the spring constant of the sensor can also be used for other types of sensors than cantilever beams, for instance quartz sensors, which will be discussed in Chap. 18. If the cantilever spring constant is known from other sources, (11.23) can be used to determine the thermal oscillation amplitude $\left\langle \Delta z^2 \right\rangle$.

In Appendix C we present several corrections (going beyond the approximation of the cantilever as a simple harmonic oscillator) which have to be applied for a more exact determination of the force constant by the thermal method.

[7] Details of how to extract the noise power spectral density from the time signal without using a spectrum analyzer are given in Appendix B or [15].

11.6.5 Experimental Determination of the Sensitivity and Spring Constant in AFM Without Tip-Sample Contact

In the preceding sections, we described two methods for the measurement of the spring constant (the Sader method and the thermal method), as well as the standard method for obtaining the sensitivity factor of the cantilever S_{sensor}. This standard method using a sensor voltage versus position curve on a hard sample for the determination of the sensitivity factor has the disadvantage that a hard contact between tip and sample occurs. This can in principle lead to tip damage or a contamination of the tip. Therefore, a calibration of the sensitivity factor without tip-sample contact is desirable.

In the following, we describe how the two non-contact methods for the determination of the cantilever spring constant k can be combined in order to obtain the sensitivity factor as well as the spring constant of the cantilever without any contact between tip and sample [16]. In a first step the Sader method is used, as described above, in order to determine the spring constant of the cantilever k. In the following, the thermal method is used, however, this time in order to obtain the sensitivity factor S_{sensor}. In the thermal method the primary measured quantity is the voltage corresponding to the thermal fluctuations of the AFM sensor (e.g. cantilever) displacement $\Delta V_{\text{sensor,th}}$. This quantity is converted to $\Delta z_{\text{sensor,th}}$ using the sensitivity factor S_{sensor} as $\Delta z_{\text{sensor,th}} = \Delta V_{\text{sensor,th}} S_{\text{sensor}}$. Inserting this in (11.21) results in

$$\frac{1}{2} k S_{\text{sensor}}^2 \left\langle \Delta V_{\text{sensor,th}}^2 \right\rangle = \frac{1}{2} k_{\text{B}} T. \tag{11.25}$$

If the spring constant k is known from the Sader method, S_{sensor} can be determined from the time average of the measured thermal fluctuations of the sensor voltage $\left\langle \Delta V_{\text{sensor,th}}^2 \right\rangle$.

While the previous consideration shows conceptually how to combine the Sader method and the thermal method in order to determine both, the spring constant and the sensitivity factor without tip-sample contact, in practice the spectral density is used, as outlined in the following.

The deflection noise density $N_{\text{z,th}}(f)$ given in (11.22) is related to the actually measured deflection *voltage* noise density $N_{\text{V,th}}(f)$ by $N_{\text{z,th}}(f) = N_{\text{V,th}}(f) S_{\text{sensor}}$. The thermal noise power spectral density $N_{\text{V,th}}^2(f)$ of the cantilever deflection signal can be measured using a spectrum analyzer. This experimentally measured noise density can be fitted (similar to (11.22)) by the function

$$N_{\text{V,th}}^2(f) = \frac{N_{\text{V,th,exc}}^2}{\left(1 - \frac{f^2}{f_0^2}\right)^2 + \frac{1}{Q^2} \frac{f^2}{f_0^2}}, \tag{11.26}$$

in order to determine f_0, Q, and $N_{\mathrm{V,th,exc}}$. Analogous to (11.23) the following equation holds

$$\langle \Delta z^2 \rangle = S_{\mathrm{sensor}}^2 \int_0^\infty N_{\mathrm{V,th}}^2(f)\mathrm{d}f = S_{\mathrm{sensor}}^2 N_{\mathrm{V,th,exc}}^2 \frac{\pi Q f_0}{2} = \frac{k_{\mathrm{B}}T}{k}. \tag{11.27}$$

From this equation the sensitivity factor S_{sensor} can be determined as

$$S_{\mathrm{sensor}} = \sqrt{\frac{2k_{\mathrm{B}}T}{\pi N_{\mathrm{V,th,exc}}^2 k Q f_0}}, \tag{11.28}$$

and thus S_{sensor} can be determined without tip-sample contact. In total with this combination of the Sader method and the thermal method k and S_{sensor} can be determined without tip-sample contact.

11.7 Summary

- Cantilever force sensors for atomic force microscopy should have a small spring constant in order to obtain a high force sensitivity and avoid large tip-sample forces. As a second condition the cantilevers should have a high resonance frequency in order to obtain a fast scanning as well as immunity to external vibrations. Both requirements can be fulfilled by sensors with a small mass, i.e. micrometer dimensions.
- Cantilevers for atomic force microscopy are fabricated from silicon using lithography and wet etching technologies from microelectronics.
- In the beam deflection method, a laser beam is reflected from the back of the cantilever and the angular deflection of the beam is detected by a split photodiode. This signal is proportional to the deflection of the cantilever Δz.
- The optical beam deflection method is a very sensitive method ($\Delta z \sim$ pm) for measuring the cantilever deflection.
- Other AFM detection methods are interferometry, piezoresistive detection, and piezoelectric detection.
- Sensor voltage versus distance curves are used to convert the measured sensor voltage $\Delta V_{\mathrm{sensor}}$ to a cantilever deflection Δz, i.e. determining the sensor sensitivity factor S_{sensor}.
- The cantilever spring constant can be obtained (a) by the material constants and dimensions, (b) by considering damping in a fluid (Sader method), or (c) via the deflection amplitude of the thermal noise signal (thermal method).
- With a combination of the Sader method and the thermal method, k and S_{sensor} both can be determined without tip-sample contact.

References

1. C.H. Oon, J.T.L. Thong, Y. Lei, W.K. Chim, High-resolution atomic force microscope nanotip grown by self-field emission. Appl. Phys. Lett. **81**, 3037 (2002). https://doi.org/10.1063/1.1515120
2. W.D. Pilkey, *Formulas for Stress, Strain and Structural Matrices*, 2nd edn. (Wiley, New York, 2005). https://doi.org/10.1002/9780470172681
3. F. Reif, *Fundamentals of Statistical and Thermal Physics*, (Waveland Pr. Inc., 1965) ISBN 1577666127
4. G. Meyer, N.M. Amer, Novel optical approach to atomic force microscopy. Appl. Phys. Lett. **53**, 1045 (1989). https://doi.org/10.1063/1.100061
5. Y. Martin, C.C. Williams, H.K. Wickramasinghe, Atomic force microscope force mapping and profiling on a sub 100 Å scale. J. Appl. Phys. **61**, 4723 (1987). https://doi.org/10.1063/1.338807
6. D. Rugar, H.J. Mamin, P. Guethner, Improved fiberoptic interferometer for atomic force microscopy. Appl. Phys. Lett. **55**, 2588 (1989). https://doi.org/10.1063/1.101987
7. M. Tortonese, R.C. Barrett, C.F. Quate, Atomic resolution with an atomic force microscope using piezoresistive detection. Appl. Phys. Lett. **62**, 834 (1993). https://doi.org/10.1063/1.108593
8. D. Kiracofe, K. Kobayashi, A. Labuda, A. Raman, H. Yamada, High efficiency laser photothermal excitation of microcantilever vibrations in air and liquids. Rev. Sci. Instrum. **82**, 013702 (2011). https://doi.org/10.1063/1.3518965
9. J.L. Hutter, Comment on tilt of atomic force microscope cantilevers: effect on spring constant and adhesion measurements. Langmuir **21**, 2630 (2005). https://doi.org/10.1021/la047670t
10. J.E. Sader, J.W.M. Chon, P. Mulvaney, Calibration of rectangular atomic force microscope cantilevers. Rev. Sci. Instrum. **70**, 3967 (1999). https://doi.org/10.1063/1.1150021
11. J.E. Sader, J.A. Sanelli, B.D. Adamson, J.P. Monty, X. Wei, S.A. Crawford, J.R. Friend, I. Marusic, P. Mulvaney, E.J. Bieske, Spring constant calibration of atomic force microscope cantilevers of arbitrary shape. Rev. Sci. Instrum. **83**, 103705 (2012). https://doi.org/10.1063/1.4757398
12. S.M. Cook, T.E. Schäffer, K.M. Chynoweth, M. Wigton, R.W. Simmonds, K.M. Lang, Practical implementation of dynamic methods for measuring atomic force microscope cantilever spring constants. Nanotechnology **17**, 2135 (2006). https://doi.org/10.1088/0957-4484/17/9/010
13. M. Godin, V. Tabard-Cossa, P. Grütter, Quantitative surface stress measurements using a microcantilever. Appl. Phys. Lett. **79**, 551 (2001). https://doi.org/10.1063/1.1387262
14. J.L. Hutter, H. Bechhoefer, Calibration of atomicforce microscope tips. Rev. Sci. Instrum. **64**, 1868 (1993). https://doi.org/10.1063/1.1143970
15. G. Heinzel, A. Rüdiger, R. Schilling, Spectrum and spectral density estimation by the Discrete Fourier Transform (DFT), including a comprehensive list of window functions and some new at-top windows (2002). http://pubman.mpdl.mpg.de/pubman/item/escidoc:152164:1/component/escidoc:152163/395068.pdf
16. M.J. Higgins, R. Proksch, J.E. Sader, M. Polcik, S. Mc Endoo, J.P. Cleveland, S.P. Jarvis, Noninvasive determination of optical lever sensitivity in atomic force microscopy. Rev. Sci. Instrum. **77**, 013701 (2006). https://doi.org/10.1063/1.2162455

Chapter 12
Static Atomic Force Microscopy

In static atomic force microscopy the force between the tip and the sample leads to a deflection of the cantilever according to Hooke's law. This cantilever bending is measured, for instance, by the beam deflection method. The name static comes from the fact that the cantilever is not excited to oscillate, as in the dynamic modes of AFM. In the following, we will discuss the static mode, while the dynamic variants are considered in the subsequent chapters. At the end of this chapter, we discuss how force-distance curves can be used to identify the tip-sample interaction regime in which subsequent imaging is performed.

12.1 Principles of Static Atomic Force Microscopy

In static atomic force microscopy, the sample is scanned in the xy-direction while the tip-sample distance is so small that the cantilever sensor can sense the tip-sample force. In the constant force mode of static atomic force microscopy, a certain setpoint value of the tip-sample force is selected via a certain deflection of the cantilever Δz, which is in turn realized by a corresponding sensor signal ΔV_{sensor}. The sensor signal is kept close to the setpoint value via a feedback loop as shown already in Fig. 1.8. When scanning, for example, over a step edge, the tip-sample force changes and thus the corresponding deflection Δz deviates from its setpoint value. The feedback electronics adjusts the z-signal controlling the tip-sample z-distance in order to restore the setpoint value of the cantilever deflection Δz. For ideal feedback, the deflection of the cantilever should always stay very close to its setpoint value. Topographic images are recorded by scanning the tip over the sample surface, while the feedback maintains constant cantilever deflection. The z-height contour corresponds to a contour of constant tip-sample force. For the setpoint value of the force, either a repulsive force or an attractive force can be selected.

Static atomic force microscopy often operates in the repulsive regime of the force-distance curve. In this case, static atomic force microscopy is also known as contact mode atomic force microscopy. The terms static mode and contact mode (repulsive

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3_12

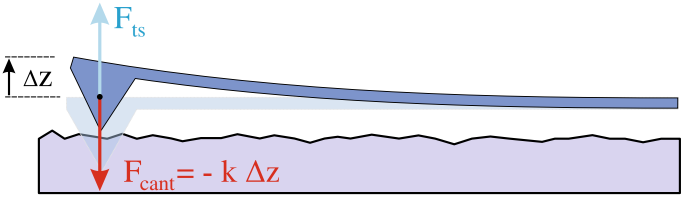


Fig. 12.1 Force equilibrium in static mode. The tip-sample force F_{ts} and the spring force of the cantilever compensate to a net vanishing force

force regime) are often misleadingly used as synonyms. However, it is also possible to operate the static atomic force microscopy in the attractive (non-contact) regime. We will distinguish between the mode of operation: static (non-oscillating cantilever) or dynamic (oscillating cantilever), on the one hand, and the type of interaction probed: repulsive (contact) or attractive (non-contact), on the other hand.

In static atomic force microscopy, the z-position of the tip, i.e. the deflection of the cantilever, is given by a balance of forces. If the tip comes close to the sample, a force F_{ts} acts on the tip. This force leads to a deflection of the cantilever by Δz relative to the equilibrium of the free cantilever and to a corresponding force F_{cant}, as shown in Fig. 12.1. In equilibrium, the total force on the cantilever has to vanish as

$$F_{\mathrm{tot}} = 0 = F_{\mathrm{ts}} + F_{\mathrm{cant}}, \tag{12.1}$$

with $F_{\mathrm{cant}} = -k\Delta z$.

If we take a closer look at the force between tip and sample, F_{ts}, this force comprises several forces: the long-range attractive van der Waals force and the short-range repulsive forces as well as the Hertzian contact force. For the force between individual pairs of tip and sample atoms, we consider as a model potential the Lennard-Jones potential plotted once more in Fig. 12.2b. The direction of the force on individual tip atoms resulting from the interaction with the sample is shown by arrows in Fig. 12.2a. For different atoms of the tip, forces with different strength and direction act depending on the distance to the sample. Tip atoms closer to the sample experience a net repulsive force (red in Fig. 12.2), while the atoms slightly farther from the sample experience an attractive interaction (blue in Fig. 12.2). The total tip-sample force is obtained by summation.

Considering that the force between the tip and sample arises due to summation (integration) over billions of atoms in the tip (and in the sample) it might be feared that nanometer or even atomic resolution might never be reached. In this regard two things are helpful: (a) the long-range (attractive) interactions are much weaker than the short-range repulsive forces and (b) the distance dependence of the long-range forces is much weaker than that of the short-range forces. Thus, the long-range forces result in a background force which is almost independent of the tip-sample distance, if, for instance, the tip-sample distance changes by 1 Å. However, 1 Å change in tip-

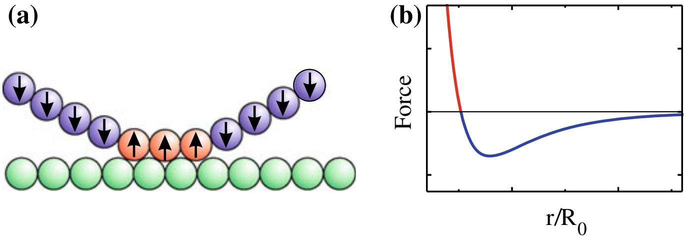


Fig. 12.2 Forces on the tip atoms due to interaction with the sample. For the atoms close to the surface the net interaction with the sample is repulsive (indicated in *red*). For larger distances to the sample, the interaction is attractive (indicated in *blue*)

sample distance changes the short-range forces significantly, enabling nanometer or even atomic resolution, as we will see later.

There are cases in which the total interaction between tip and sample can still be attractive due to the long-range attractive forces, while it is already repulsive for the atoms at the tip apex. Since the tip-sample forces also act on the sample, the sample (and tip) can be deformed if these forces become strong. This deformation of tip and sample in the area of the repulsive interaction can establish a contact area consisting of several atoms and therefore inhibit true atomic resolution of single defects in the atomic lattice. This effect in contact atomic force microscopy is called the "egg carton effect", since the atomic corrugations of tip and sample slide along each other like two egg cartons. Since the repulsive forces increase very strongly with decreasing tip-sample distance, images of constant repulsive force are often identified with the topography of the surface.

The non-monotonous distance dependence of the tip-sample force leads to the fact that for some forces (negative forces in Fig. 12.2b) two tip-sample distances exist for a certain force. Depending on the polarity (direction) of the feedback one or the other of the two branches with different slope can be stabilized. As discussed in Sect. 16.3 in detail, this can lead to instabilities in feedback behavior if the tip unintentionally switches from one branch to the other branch with the opposite slope as a function of distance. Only for net repulsive forces there is a single branch present and instabilities are avoided.

12.2 Properties of Static AFM Imaging

If static atomic force microscopy is operated in the contact mode, the tip is in direct contact with the sample and strong repulsive forces act between tip and sample. To avoid damaging the probed surface, the cantilever should be soft, i.e. the cantilever spring constant should be lower than the effective spring constant of the sample atomic bonds. As discussed in Sect. 10.5, under this condition snap-to-contact occurs, which is actually desired in the contact mode in order to maintain tip-sample contact during scanning.

The standard application of contact AFM is imaging the surface topography with a resolution in the nanometer range. Especially the direct determination of the height of image features is an advantage of AFM measurements. In other microscopy techniques such as optical microscopy or scanning electron microscopy, the lateral feature size is easily measured, but using these techniques do not give easy access to the true height of the imaged features.

Atomically "resolved" images using the contact mode AFM technique were first obtained on layered materials like graphite, boron nitride, mica, molybdenum selenide and others [1, 2]. These materials have the advantage that clean surfaces can be prepared under ambient conditions. While a corrugation with a periodicity of the atomic lattice is observed, defects of atomic size are not observed. This led to the conclusion that small flakes of the layered material are probably attached at the tip apex and that an "egg carton effect" prevents the detection of atomic size defects.

After the first successful applications of contact AFM to layered materials, it was natural to extend the investigations to non-layered materials. For these cases, the effect of dragging flakes of the layered materials over the surface does not occur. Inorganic crystals like NaCl [3] or LiF were prepared in ultrahigh vacuum and imaged with contact AFM. Typical forces between the sample and the tip during imaging are set to approximately 10^{-8} N. The measured step heights range down to single atomic steps and atomic corrugation was observed.

The contact zone between tip and sample in contact mode AFM is assumed not to be a single atom but consisting of many atoms. The tip is usually of a different material than the sample surface and therefore, the tip atoms are not in registry with the sample surface structure. The usual understanding for the observation of atomic corrugation is that the atoms of the tip lock into the atomic lattice of the sample, so the atomic lattice of the sample is imaged. Thus, also on salt crystals like NaCl or LiF no single atomic defects were observed in contact mode AFM. Due to an "egg carton effect" between the sample and the contact area of the tip, it is possible to observe *atomic corrugation*, while no atomic scale defects are seen and correspondingly no true *atomic resolution* is possible.

Typical problems with contact mode AFM are that contact diameters lie in the range of 1–10 nm, limiting the lateral resolution. Moreover, the relatively high forces can lead to a wear of soft (organic or biological) samples.

12.3 Constant Height Mode in Static AFM

Up to now we have considered the constant force mode of static AFM, the tip-sample force is controlled to a certain value given by the setpoint for the cantilever deflection. For the constant height mode we assume for the moment that the sample surface is perfectly aligned to the scanner, i.e. no scanning slope is present (cf. Chap. 7). In this case an xy-scan can be performed (starting with an initially preset tip-sample

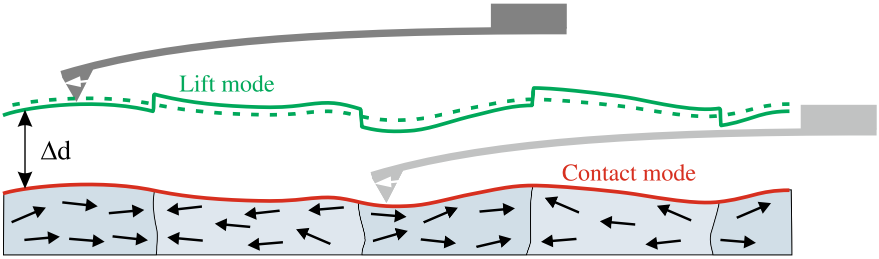


Fig. 12.3 Principle of the lift mode. In a first scan line, the topography is measured (contact mode). In a second scan line, the topography is retraced with an offset Δd (*dashed line*). The deflection due to the long-range magnetic interaction is measured relative to this retraced height (*solid line*)

distance) and the change of the cantilever deflection is measured. In this case, no feedback is involved and the scan can be performed fast. The constant height mode is mostly applied for long-range forces, i.e. electrostatic or magnetic forces.

Since it is difficult in practice to get rid of the sample tilt the actual experimental procedure is different from the principle described above. We consider here as an example a magnetic interaction sensed with a ferromagnetic tip, as sketched in Fig. 12.3. In order to be independent of variations in the topography every scan line is scanned twice. First the topography is measured using the contact mode, and in a second scan of the same line the measured topography is followed with an offset Δd relative to the previous scan line as shown in Fig. 12.3 by the dashed line. In this second line, the long-range magnetic interactions are detected by a corresponding deflection of the cantilever shown as a solid green line in Fig. 12.3. The difference between the two signals (the dashed and solid line) corresponds to the magnetic signal. This kind of constant height mode is also called the lift mode.

12.4 Friction Force Microscopy (FFM)

Due to the relative motion of tip and sample in contact mode, friction in the tip-sample contact will lead to a lateral force on the tip apex. If the scanning direction is sidewise to the cantilever length, this lateral force causes a torsional bending of the cantilever, which can be recorded in beam deflection microscopes as shown in Fig. 12.4a. Quadrant photo-diodes are used in the optical beam deflection method anyway in order to guide the beam reflected from the cantilever to the center of the photo diode. If the four quadrant cells are labeled as shown in Fig. 12.4a, the topography is contained in the signal $(A + B) - (C + D)$. The torsional bending of the cantilever (friction signal) is contained in the signal $(A + C) - (B + D)$. In this way the local variation of friction can be studied with high resolution and for various values of external parameters like the load force or the scanning velocity. One great benefit of friction force microscopy (FFM) is that it is possible to measure whether

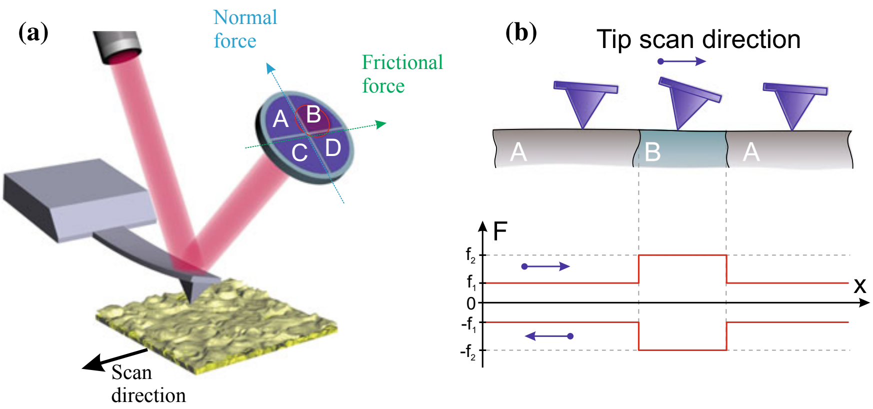


Fig. 12.4 **a** Principle of the detection of friction forces by the beam deflection method using a quadrant photodiode. **b** Difference in the friction signals in trace and retrace direction on materials with two different friction coefficients. The friction signal changes the sign between the trace and retrace directions of the fast scan direction

wear has taken place in the course of the experiment by subsequent imaging of the relevant area.

The friction signal can also be used in order to obtain material contrast. In this case the method is sometimes called also lateral force microscopy (LFM). The principle is shown in Fig. 12.4b for a sample which consists of two different materials A and B with different friction properties this leads to two different friction signals f_1 and f_2, respectively (same height of materials A and B) resulting in a material contrast in the friction signal. A specific signature for a friction signal is the following: If the same scan line is scanned in the opposite direction (retrace) the friction signals reverse as shown in Fig. 12.4b.

12.5 Force-Distance Curves

Force-distance curves are measured by bringing the sample towards the cantilever tip and measuring the cantilever deflection which is proportional to the tip-sample force. These force-distance curves contain the following useful information: (a) The sensitivity of the detection method can be determined as described in Sect. 11.6. (b) Properties like the sample elasticity or the maximum tip-sample adhesion force can be accessed. (c) The working point (setpoint for the cantilever deflection signal) for subsequent AFM imaging can be characterized and chosen properly. For instance, when imaging is performed in the attractive force regime it can be determined how far the working point is located from the point of snap-to-contact. (d) A force-distance curve can be used to determine the tip-sample force-distance dependence, at least partly.

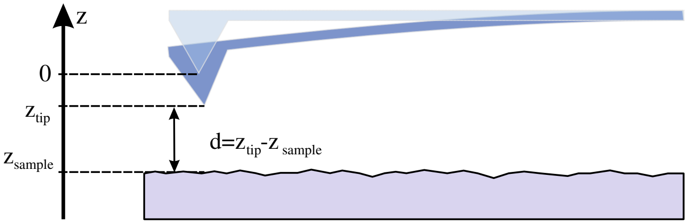


Fig. 12.5 Nomenclature for the coordinates used in force-distance curves

The aim is to obtain the tip-sample force $F_{ts}(d)$ as a function of the tip-sample distance d, as indicated in Fig. 12.5. What is actually measured when acquiring a force-distance curve is the deflection of the cantilever z_{tip} (which is proportional to the tip-sample force) as function of the z-position of the sample z_{sample} as $z_{tip}(z_{sample})$. This has the disadvantage that the tip-sample distance d is not only given by the (intended) z-motion of the sample (induced by a voltage at the z-piezo element) but also by an additional distance change due to the deflection of the cantilever z_{tip} (force measurement) as shown in Fig. 12.5. However, d can always be recovered as $d = z_{tip} - z_{sample}$. With the coordinate system in Fig. 12.5, the action (approach of the sample) and the reaction (deflection of the cantilever) are separated into two coordinates. Also experimentally, these two parameters are measured or set independently: z_{tip} is measured via the cantilever deflection, while z_{sample} is set via the applied z-piezo voltage. As the zero point for z_{tip} and z_{sample}, we choose the equilibrium position of the cantilever tip with the sample far away.

Figure 12.6a shows a schematic of a $z_{tip}(z_{sample})$ plot for the model force-distance curve which is shown in the inset. The blue curve corresponds to the approach of the sample towards the tip, while the red curve corresponds to the retraction of the sample. As the sample approaches the tip (increasing z_{sample} from the right to the left) the cantilever bends slightly towards the sample (negative z_{tip} values) due to the attractive force between tip and sample. At point c, the force gradient exceeds the value of the spring constant k (indicated by a dashed line in the inset). This leads to the previously discussed instability and to snap-to-contact (cf. Sect. 10.5). The cantilever jumps to point d. The maximal cantilever deflection at point c multiplied by the spring constant gives the maximum attractive force before snap-to-contact (usually quite small).

If the sample is moved further towards the tip, the point is reached where attractive and repulsive tip-sample interactions compensate each other and the net tip-sample force vanishes. At this position, the cantilever is unbent ($z_{tip} = 0$). The position z_{sample} has in the experiment an unknown offset (the z-position moving the sample is not known absolute, relative to $z_{tip} = 0$) which is fixed by setting $z_{sample} = 0$ when $z_{tip} = 0$, as done in Fig. 12.6a.[1] If the sample is pushed further towards the

[1] This is somewhat different from the tip-sample distance used for the Lennard-Jones force, where a tip-sample distance (or better atom-atom distance) of zero corresponds to a repulsive force approaching infinite strength.

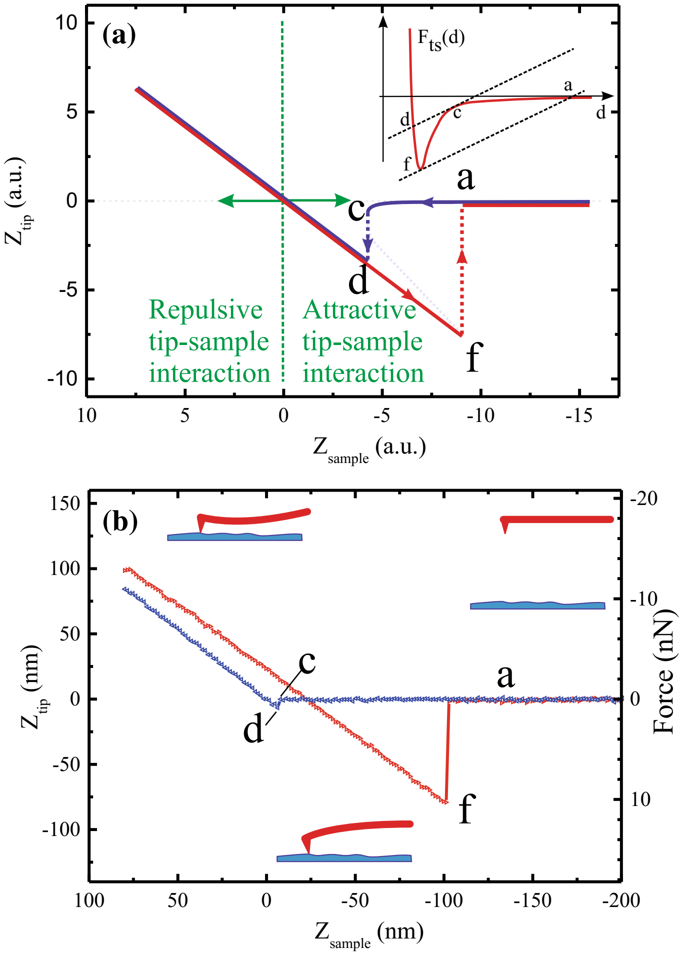


Fig. 12.6 **a** Schematic of a $z_{\mathrm{tip}}(z_{\mathrm{sample}})$ plot with the *blue curve* corresponding to an approach of the sample toward the tip, while the *red curve* corresponds to a retraction of the sample. The nomenclature for the variables is the same as in Fig. 12.5. At points c and f, the tip-sample force gradient becomes equal to the spring constant of the cantilever and leads to an instability associated with snap-to-contact or snap-out-of-contact, respectively. **b** Experimentally measured force-distance curve obtained on a silicon wafer in a lab course at RWTH Aachen University. The cantilever spring constant was 0.13 N/m (The unusual coordinate system has negative z_{sample} values going to the right. This is, however, the way it is normally plotted)

tip, the regime of repulsive tip-sample interaction is entered. In the repulsive regime the sample bends the cantilever upwards. As for hard samples the repulsive force rises very sharply with decreasing tip-sample distance, both tip and sample move together ($\Delta z_{\text{sample}} \approx \Delta z_{\text{tip}}$ and $d \approx 0$) Specifically for a stiff sample with a high elastic modulus, the $z_{\text{tip}}(z_{\text{sample}})$ curve is a straight line with a slope of one, as shown in the left part of Fig. 12.6a. If the sample is soft, the slope can be (initially) smaller than one (due to an indentation of the tip into the sample), resulting in information about the elastic/plastic deformation of the sample (cf. Chap. 15).

If the direction of the sample motion is reversed, the tip motion follows the same straight line in the reverse direction (red line) for stiff samples. The repulsive tip-sample force decreases continuously and finally the attractive regime is entered again, where tip and sample adhere to each other as long as the tip-sample force gradient is smaller than the cantilever spring constant. If the force gradient becomes larger than the cantilever spring constant, the cantilever snaps out of contact (point f). The tip snaps back to a position where the deflection of the cantilever is close to zero (point a). Point f corresponds (approximately) to the position at which the maximum attractive force (adhesion force) between tip and sample acts. Generally, for elastic samples the retraction curve and the approach curve are the same in the repulsive regime, while for a plastic deformation the repulsive force during retraction is smaller than during approach.

In Fig. 12.6b an experimental force-distance curve is shown which in principle resembles the behavior discussed above. The measured tip deflection is converted (via Hooke's law $F_{\text{ts}} = -F_{\text{cant}} = kz_{\text{tip}}$) to a corresponding force F_{cant}, which is shown on the right axis in Fig. 12.6b. In the experimental $z_{\text{tip}}(z_{\text{sample}})$ plot, the jump to contact (from point c to point d) is small. The corresponding force (the attractive force before snap into contact) is less than 1 nN. The maximal attractive force, which is reached at point f just before snap out of contact, can be extracted as 10 nN. Also the width of the attractive force minimum can be read from the difference in z_{tip} between point c and d. This shows that several important parameters can be extracted directly from the measured force-distance curve. In one respect, the measured force-distance curve does not follow the idealized expectation shown in Fig. 12.6a. The approach curve (blue) and the retract curve (red) do not coincide for positive sample distances in Fig. 12.6b. This effect arises due to hysteresis and creep effects of the piezoelectric actuators.

In principle, the measured $z_{\text{tip}}(z_{\text{sample}})$ curve or equivalently the $F_{\text{ts}}(z_{\text{sample}})$ curve (right axis in Fig. 12.6b) can be translated into the more fundamental force-distance curve $F_{\text{ts}}(d) = kz_{\text{tip}}$, with $d = z_{\text{tip}} - z_{\text{sample}}$. However, as can be seen from the inset in Fig. 12.6a, the force-distance curve between points c and f is inaccessible due to snap in and out of contact. Unfortunately, this is one of the interesting regions. For large distances down to point c the tip-sample force is almost negligible, while for distances closer than point f, the force rises very steeply. The range in which the force-distance curve can be measured could be extended by using a cantilever with a larger spring constant. However, this has the drawback of reduced force sensitivity.

The importance of the force-distance curves for subsequent imaging lies in the fact that a particular point on the force-distance curve can be identified and that subsequent imaging of the sample can be performed at a defined position (working point) on this curve. This is important because the imaging in AFM depends critically on the applied force. For instance in imaging soft (biological) samples it is preferable to avoid strong repulsive forces between tip and sample as this leads to wear on soft sample structures. In order to achieve this the force-distance curve can be measured and the working point for imaging is selected close to point f in Fig. 12.6a, i.e. in the regime of attractive tip-sample interaction, thus avoiding large repulsive forces. However, since this condition is close to snap-out-of contact, there is a danger of leaving the desired imaging conditions by snap-out-of-contact.

The use of force-distance curves in order to determine fundamental force-distance dependences is limited. Several fundamental forces act simultaneously and sum up over the tip and sample volume. The measured forces are integrals of several fundamental forces over large volumes of tip and sample. Additional problems such as capillary forces, an unknown tip shape, and piezo creep complicate a more quantitative interpretation of the tip-sample interaction. Due to these limitations, force-distance curves are not used to measure the fundamental forces.

12.6 Summary

- In static AFM, the tip-sample force is measured via the deflection of the cantilever Δz.
- In the constant force mode of static AFM, a certain force setpoint is kept constant by feedback during scanning of the surface. The resulting topography corresponds to a contour at constant tip-sample force.
- In the repulsive interaction regime, the tip-sample contact consists of many atoms and thus no atomic *resolution* is expected, but atomic *corrugation* can be observed.
- The constant height mode is mostly used to image corrugation induced by long-range interactions such as magnetic or electrostatic forces.
- Friction forces can be measured via the torsional bending of the cantilever using a quadrant photodiode.
- Force-distance measurements give access to various parameters of the force-distance curve. The working point for subsequent AFM imaging can be chosen using the information from the force-distance curve.

References

1. G. Binnig, C. Gerber, E. Stoll, T.R. Albrecht, C.F. Quate, Atomic resolution with atomic force microscope. Europhys. Lett. **3**, 1281 (1987). https://doi.org/10.1209/0295-5075/3/12/006
2. T.R. Albrecht, C.F. Quate, Atomic resolution imaging of a nonconductor by atomic force microscopy. J. Appl. Phys. **62**, 2599 (1987). https://doi.org/10.1063/1.339435
3. G. Meyer, N.M. Amer, Optical-beam-deflection atomic force microscopy: the NaCl(001) surface. Appl. Phys. Lett. **56**, 2100 (1998). https://doi.org/10.1063/1.102985

Chapter 13
Amplitude Modulation (AM) Mode in Dynamic Atomic Force Microscopy

In dynamic atomic force microscopy the cantilever is excited at a driving frequency which is close to the resonance frequency of the free cantilever [1–3]. Due to the interaction between tip and the surface, the resonance frequency of the cantilever changes. As shown in this chapter, an attractive force between tip and sample leads to a lower resonance frequency of the cantilever, while for repulsive tip-sample forces the resonance frequency increases.[1]

This change in resonance frequency can be measured directly in the frequency modulation mode (FM) of atomic force microscopy, as described in Chap. 16. In this chapter, we describe the amplitude modulation mode (AM) of AFM. In this mode the cantilever is driven (oscillated) at a fixed frequency with a fixed driving amplitude. The change of the resonance frequency due to the tip-sample interaction leads to a change of the oscillation amplitude and of the phase between excitation and oscillation, which can be measured.

We consider the AM detection mode in this chapter in the small amplitude limit in which the tip-sample force is approximated as linear in the range of the oscillation amplitude. In this case, the AM detection mode can be treated analytically. While in practice the AM detection mode is rarely used in this limit, the basic concepts can be explained more easily using this limit. When in the next chapter the small amplitude limit is lifted, things become somewhat more complicated. However, armed with a basic understanding obtained from the treatment of the small amplitude limit, the more realistic case is then easier to comprehend.

13.1 Parameters of Dynamic Atomic Force Microscopy

Compared to static AFM, there are many more parameters in dynamic AFM.

[1] Actually, this is not strictly true: As shown later it is not the sign of the force, but rather the sign of the *force gradient* that determines the direction of the resonance frequency shift.

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3_13

- The resonance frequency of the free cantilever ω_0
- The force constant of the cantilever k
- The quality factor of the cantilever Q_{cant}
- The driving frequency ω_{drive}
- The driving amplitude of the oscillation A_{drive}
- The oscillation amplitude A
- The phase ϕ between oscillation and driving
- The frequency shift of the resonance frequency $\Delta\omega$ relative to ω_0 due to a tip-sample interaction

The first two parameters are given by the cantilever, while the Q-factor depends on the cantilever and also on the operating environment (ambient or vacuum). Depending on the operating mode, further parameters can be set by the operator or measured:

- In AM detection the amplitude A and phase ϕ of the oscillation are measured, while ω_{drive} is fixed and A_{drive} is set via the setpoint of the feedback loop.
- In FM detection the shift of the resonance frequency $\Delta\omega$ is measured, while the excitation amplitude or the oscillation amplitude are set.

Because this multitude of parameters may seem somewhat discouraging, we will discuss the parameters and the relations among them step by step in the following.

13.2 Principles of Amplitude Modulation Dynamic Atomic Force Microscopy

As the simplest model for the cantilever under the influence of a tip-sample interaction, we consider the driven damped harmonic oscillator as discussed in Sect. 2.4 including the influence of a time-independent external tip-sample force F_{ts}, which depends on the tip-sample distance. In this section, we assume that dissipation enters only via the (air) damping of the cantilever, while the tip-sample interaction is assumed to be conservative.

We assume the limit of small amplitude, which means that F_{ts} can be approximated as linear (as function of the tip-sample distance) in the range of the oscillation amplitude A. We use this limit here because this idealized scenario can be solved analytically. For the usual vibration amplitudes (several tens nanometers) the small amplitude limit does not hold.

We split the tip-sample distance into a constant (offset) distance d, while the oscillatory motion of the tip is described by the coordinate z. The definition of the coordinates of the cantilever-tip-sample system is given in Fig. 13.1.

Before we analyze the oscillating cantilever, we consider the static case (i.e. the oscillatory amplitudes A_{drive} and A in Fig. 13.1 are zero). Without a tip-sample force being present, the tip is at its zero position $z = 0$ and the cantilever is unbent (shown

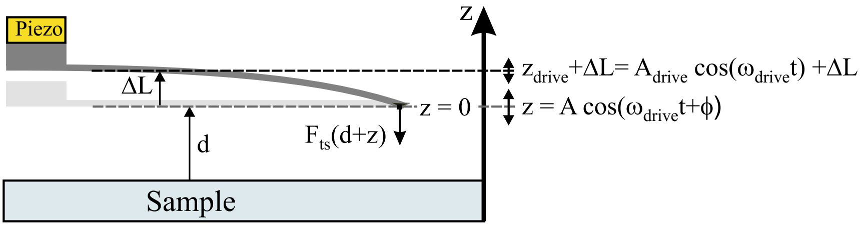


Fig. 13.1 Definition of the coordinates for a driven damped harmonic oscillator under the influence of a tip-sample force

in light gray in Fig. 13.1). In this case the static bending ΔL is zero.[2] If the tip-sample force is now switched on, the tip-sample distance will change. Since we would like to probe the interaction at a tip-sample distance d, the initial zero position of the tip, $z = 0$, is restored by moving the cantilever base in the opposite direction, shown in dark gray in Fig. 13.1. In static equilibrium with the cantilever bent, the tip-sample force and the static bending force balance at $z = 0$ as

$$F_{\mathrm{ts}}(d) = -k\Delta L, \tag{13.1}$$

with ΔL being the static (offset) deflection of the cantilever (Fig. 13.1).

Now we begin to consider an oscillating cantilever. For the tip-sample force $F_{\mathrm{ts}}(d+z)$, we use the coordinate $d+z$ (tip-sample distance), with the offset d being the distance from the sample to the equilibrium position of the tip $z = 0$, relative to which the oscillatory motion occurs. Due to the small amplitude assumption, we can expand the force $F_{\mathrm{ts}}(d+z)$ around the equilibrium position of the tip ($z = 0$, corresponding to a tip-sample distance d) as

$$F_{\mathrm{ts}}(d+z) = F_{\mathrm{ts}}(d) + \left.\frac{\partial F_{\mathrm{ts}}(d+z)}{\partial z}\right|_{z=0} \cdot z + \cdots . \tag{13.2}$$

If we neglect higher order terms, the force changes linearly with z, like it is the case for a spring. Hence the influence of the tip-sample force can be described by a spring with a spring constant k' equal to the negative force gradient, as

$$k' \equiv -\left.\frac{\partial F_{\mathrm{ts}}(d+z)}{\partial z}\right|_{z=0} . \tag{13.3}$$

The tip-sample interaction can be represented by adding a spring with spring constant k' to the cantilever spring with the spring constant k, as shown in Fig. 13.2a. The two spring constants add up[3] to an effective spring constant $k_{\mathrm{eff}} = k + k'$. However, this

[2] The tip length is set to zero in order to avoid an additional offset length.

[3] Since the two springs attach to the tip from above and below one might think that this should lead to a subtraction of the spring constants. Here we show that the spring constants indeed add

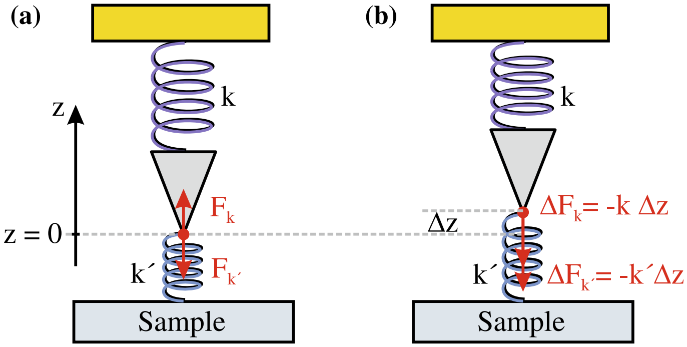


Fig. 13.2 **a** For the case of small amplitudes, the cantilever-tip-sample system can be effectively described by two springs, one representing the cantilever with force constant k and one representing the tip-sample interaction with force constant k'. **b** This system is equivalent to a system with an effective spring constant of $k_{\mathrm{eff}} = k + k'$

analogy (replacing the tip-sample interaction by a spring) should not be stretched too far, since real spring constants of springs are always positive, while a tip-sample interaction can also have a "negative spring constant". Such a negative spring constant k' cannot be realized by a coil spring or a cantilever-shaped spring, but can exist in a more general sense as a potential of negative curvature.

We will now consider a sinusoidal excitation of the cantilever base at the frequency ω_{drive} and amplitude A_{drive} around the position of static equilibrium ($z_{\mathrm{drive}} = \Delta L$) as $z_{\mathrm{drive}} = A_{\mathrm{drive}} \cos(\omega_{\mathrm{drive}} t)$. As a result of this excitation, the tip will oscillate in the steady-state around its equilibrium position as $z = A\cos(\omega_{\mathrm{drive}} t + \phi)$. The equation of motion can be written, according to the treatment in Sect. 2.4 as

$$m\ddot{z} + m\frac{\omega_0}{Q_{\mathrm{cant}}}\dot{z} = -k(z - z_{\mathrm{drive}} - \Delta L) + F_{\mathrm{ts}}(d) + \left.\frac{\partial F_{\mathrm{ts}}(d+z)}{\partial z}\right|_{z=0} \cdot z, \quad (13.4)$$

by including the offset length ΔL and adding the linear approximation for the tip-sample force on the right.[4] As the static force due to the bending of the cantilever cancels out against, the force $F_{\mathrm{ts}}(d)$, according to (13.1), these terms can be removed from the equation of motion, which reads after division by m and using (13.3) as

up. As indicated in Fig. 13.2 the cantilever spring under the influence of a tip-sample force can be approximated by a cantilever effective mass held by two springs (i.e. the cantilever spring k and the spring k' representing the tip-sample interaction). In static equilibrium, $z = 0$, the forces of both springs compensate as $F_k + F_{k'} = 0$. If the cantilever is moved by Δz during the oscillation, Fig. 13.2b shows that the force components relative to the forces in static equilibrium point in the same direction for both springs and $\Delta F = \Delta F_k + \Delta F_{k'} = -(k + k')\Delta z$ results. Thus, the spring constants k and k' combine to $k_{\mathrm{eff}} = k + k'$.

[4] In the spring model the force $F_{\mathrm{ts}}(d)$ can be considered arising form an offset stretch of the tip-sample spring.

$$\ddot{z} + \frac{\omega_0}{Q_{\text{cant}}}\dot{z} + \frac{k}{m}(z - z_{\text{drive}}) = \left.\frac{\partial F_{\text{ts}}(d+z)}{\partial z}\right|_{z=0} \cdot \frac{z}{m} = -\frac{k'}{m}z. \tag{13.5}$$

After replacing $k/m = \omega_0^2$ the equation of motion reads as

$$\ddot{z} + \frac{\omega_0}{Q_{\text{cant}}}\dot{z} + \frac{k+k'}{m}z = \omega_0^2 z_{\text{drive}}. \tag{13.6}$$

If we replace $k + k'$ by the effective spring constant k_{eff}, the equation of motion (13.6) is identical to the equation of motion of the driven damped harmonic oscillator with damping (2.25), with the only replacement $k/m \to k_{\text{eff}}/m \equiv \omega_0'^2$. Thus, we know the solution of (13.6) from (2.31) as

$$A^2 = \frac{\omega_0^4 A_{\text{drive}}^2}{\left(\omega^2 - {\omega_0'}^2\right)^2 + \frac{\omega_0^2\omega^2}{Q_{\text{cant}}^2}}. \tag{13.7}$$

In the following, we assume that $\left|k'\right| \ll k$ and thus ω_0' is very close to ω_0. In this case ω_0 can be replaced by ω_0' in (13.7), which then corresponds to a resonance curve as discussed in Sect. 2.4, with the only difference that now the resonance frequency is $\omega_0' = \sqrt{k_{\text{eff}}/m}$, instead of $\omega_0 = \sqrt{k/m}$. Thus, the linearized tip-sample force shifts the resonance frequency from ω_0 to ω_0'. This shifted resonance frequency can be written as

$$\omega_0' = \sqrt{\frac{k_{\text{eff}}}{m}} = \sqrt{\frac{k+k'}{m}} = \sqrt{\frac{k}{m}\left(1 + \frac{k'}{k}\right)} = \omega_0\sqrt{1 + \frac{k'}{k}}. \tag{13.8}$$

Since for small x the approximation $\sqrt{1+x} \approx 1 + \frac{1}{2}x$ holds, and as $\left|k'\right| \ll k$, the new resonance frequency of the cantilever can be written as

$$\omega_0' \approx \omega_0\left(1 + \frac{k'}{2k}\right). \tag{13.9}$$

The shift of the resonance frequency results in

$$\Delta\omega = \omega_0' - \omega_0 \approx \omega_0\frac{k'}{2k} = -\frac{\omega_0}{2k}\left.\frac{\partial F_{\text{ts}}}{\partial z}\right|_{z=0}. \tag{13.10}$$

This result can be easily related to the experimentally observed frequency shift Δf as

$$\Delta f = \frac{\omega_0' - \omega_0}{2\pi} \approx f_0\frac{k'}{2k} = -\frac{f_0}{2k}\left.\frac{\partial F_{\text{ts}}}{\partial z}\right|_{z=0}. \tag{13.11}$$

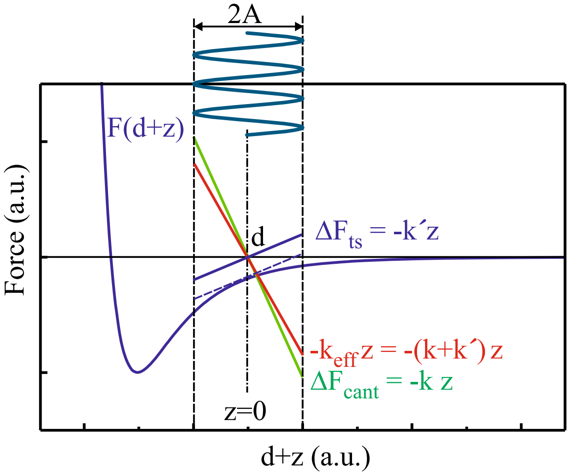


Fig. 13.3 In the small amplitude limit, the tip-sample force is approximated as linear within the range of the oscillation and considered as proportional to $-k'$. In this figure $k' < 0$ at the tip-sample distance d, while the cantilever spring constant k is always positive. Thus, the total effective force constant $k + k'$ is the cantilever spring constant k reduced by $|k'|$

Together with the resonance frequency (approximately maximum of the resonance curve according to (2.42)) also the whole resonance curve shifts by Δf.

In summary, the frequency shift of the resonance curve induced by the tip-sample interaction is proportional to the negative gradient of the tip-sample force ($-F'_{ts}(d) \equiv -\partial F_{ts}(d+z)/\partial z|_{z=0}$) if the following conditions are fulfilled: (a) The tip-sample force can be approximated as linear in the range of the oscillation amplitude, and (b) the absolute value of the tip-sample force gradient is much smaller than the spring constant of the cantilever $|k'| \ll k$ (the spring constant of the cantilever k is always positive). The small amplitude limit and its interpretation in terms of the effective spring constant is also summarized in Fig. 13.3. A Lennard-Jones type force is shown together with the tip oscillation path with amplitude A around the average tip-sample distance d. The cantilever force $\Delta F_{\text{cant}} = -kz$ is shown as a green line. The tip-sample force is approximated locally around $z = 0$ as linear $\Delta F_{ts} \approx -k'z = \partial F_{ts}/\partial z|_{z=0}\, z$, which is indicated by the blue line. The dashed blue line corresponds to the gradient of the tip-sample force at $z = 0$ (tangent to the force curve). The resulting total force is shown as a red line with a slope of $k_{\text{eff}} = k + k'$. Since in the particular case considered here $k' < 0$, the spring constant of the cantilever spring constant k is reduced by $|k'|$ comparing the green and red lines.

For a positive, i.e. negative, negative-tip-sample force gradient $-\partial F_{ts}/\partial z = k' < 0$, the resonance frequency will shift, according to (13.11), to lower values $\Delta f < 0$, while for a negative, i.e. positive, negative force gradient $-\partial F_{ts}/\partial z = k' > 0$ the resonance frequency will shift to higher values $\Delta f > 0$. The frequency shift does not depend on the constant static offset force $F_{ts}(d)$. This offset force results only in a static deflection of the cantilever, which is compensated by an offset shift of the cantilever base by ΔL, according to (13.1).

Often it is stated slightly imprecisely that the frequency shift Δf is positive (towards higher frequencies relative to f_0) for repulsive forces and negative for

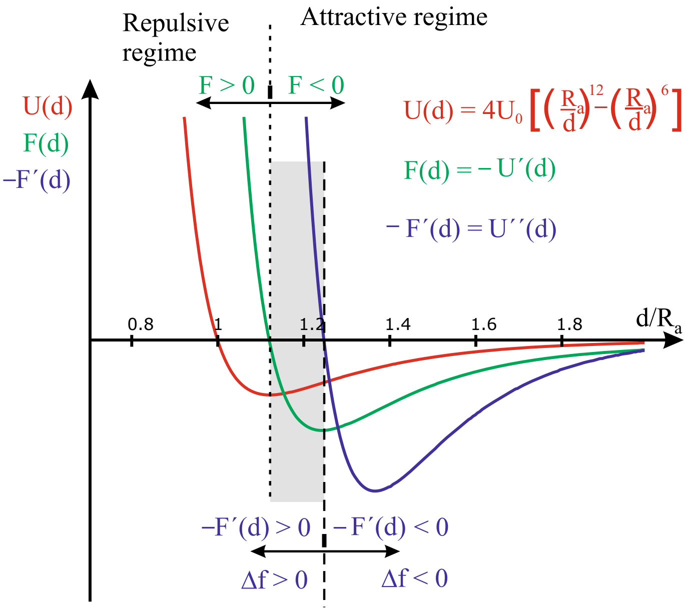


Fig. 13.4 Potential, force and negative force gradient for the Lennard-Jones model potential shown as a function of the average tip-sample distance d. As the frequency shift Δf is proportional to the negative force gradient it can be stated: For distances outside the *shaded region* the frequency shift Δf is positive (towards higher frequencies relative to f_0) for repulsive forces, and negative for attractive forces

attractive forces. We can understand this if we have a closer look at Fig. 13.4, where the potential, the force, and the (negative) force gradient are shown. Here again the Lennard-Jones potential is considered as a model for the tip-sample interaction. The border between the repulsive and attractive regime is located at the zero of the force (dotted line in Fig. 13.4). Correspondingly, the border between the positive and negative force gradient is shown by a dashed line. For the largest range of tip-sample distances, the force and the negative force gradient (green and blue curves in Fig. 13.4, respectively) have the same sign. Only for a small range of distances (shaded gray in Fig. 13.4) do the tip-sample force and the negative force gradient have a different sign. As discussed above, the frequency shift Δf is proportional to the *negative* force gradient (13.11). Correspondingly, attractive forces (negative sign) lead (in the majority of cases—except in the gray-shaded range) to a decrease of the resonance frequency. Thus, the statement that the frequency shift Δf is positive (towards higher frequencies) for repulsive forces and negative for attractive forces is true for most tip-sample distances.

The relative frequency change can be written as

$$\frac{\Delta f}{f_0} = \frac{1}{2}\frac{k' A^2}{k A^2} = \frac{E_{\text{interaction}}}{2E_{\text{cantilever}}}. \tag{13.12}$$

This means that the relative frequency shift is given by the ratio of the energy of the tip-sample interaction (spring constant k') divided by twice the energy stored in the cantilever oscillation (spring constant k).

13.3 Amplitude Modulation (AM) Detection Scheme in Dynamic Atomic Force Microscopy

We have seen that in the small amplitude limit a force gradient of the tip-sample interaction shifts the resonance frequency ω_0 by $\Delta\omega$. Accordingly, the whole resonance curve shifts by $\Delta\omega$ relative to that of the free cantilever, as shown in Fig. 13.5b.

In the amplitude modulation (AM) detection scheme, the cantilever is excited with a fixed driving amplitude A_{drive} at a fixed frequency ω_{drive} close to the resonance frequency. The resulting cantilever oscillation amplitude A is measured. As shown in Fig. 13.5, this amplitude depends indirectly on the tip-sample distance. The amplitude

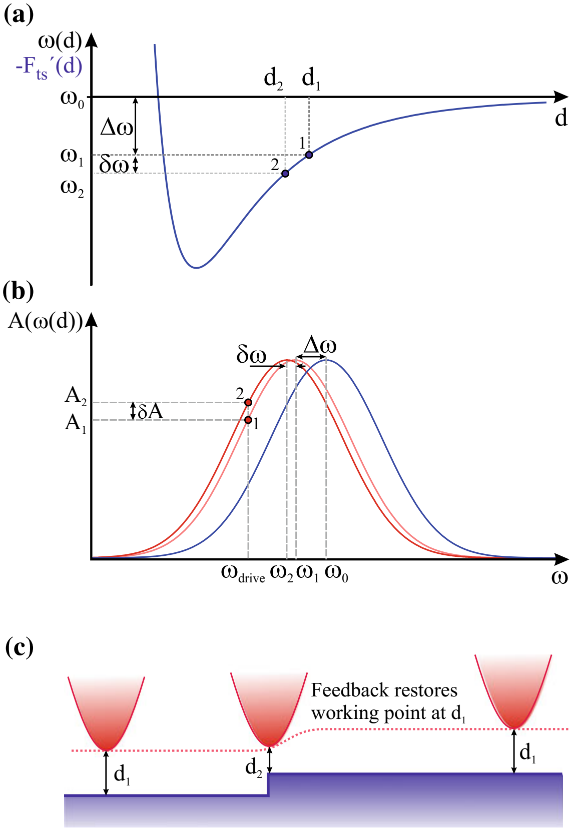


Fig. 13.5 In the AM detection scheme of dynamic AFM the measured signal depends indirectly on the tip-sample distance. **a** Primarily, the force gradient and therefore also the resonance frequency (shift) depend on the tip-sample distance (here a Lennard-Jones potential is assumed). **b** Secondly, the measured amplitude depends on the frequency shift. **c** When scanning over a step edge, the tip-sample distance changes until the feedback restores the original tip-sample distance

depends on the frequency shift of the resonance curve, which depends on the force gradient, which depends in turn on the tip-sample distance as $A(\Delta\omega(F'_{\mathrm{ts}}(d)))$.

In the following, we go through these dependencies step by step. The dependence of the force gradient on the tip-sample distance $F'_{\mathrm{ts}}(d)$ based on the Lennard-Jones model potential is shown in Fig. 13.5a. As discussed in the previous section, the frequency shift is proportional to the force gradient indicated by the double labeling of the ordinate in Fig. 13.5a. In Fig. 13.5b resonance curves $A(\omega)$ are shown which are shifted together with the respective resonance frequency. The actual oscillation amplitude of the cantilever at the driving frequency ω_{drive} is the measurement signal. In the feedback loop for the amplitude signal, a setpoint amplitude is selected, e.g. A_1 in Fig. 13.5b. The feedback loop controls the measured amplitude to the setpoint value by changing the z-position of the tip (or sample). This changes the tip-sample distance, which changes the force gradient, which changes the resonance frequency, and thus indirectly the amplitude is ultimately changed and kept at its setpoint value by the feedback. If the feedback loop maintains a constant oscillation amplitude throughout a scan, this corresponds to a height profile taken at constant force gradient. In order for an amplitude change to be highly sensitive to the corresponding frequency change, the amplitude setpoint should be close to the position of maximum slope of the resonance curve.

In our example, we chose $\omega_{\mathrm{drive}} < \omega_0$, corresponding to a negative negative (i.e. positive) force gradient (attractive tip-sample interaction). If a driving frequency larger than ω_0 is selected, this corresponds to a working point in the regime of a positive negative force gradient (roughly repulsive tip-sample interaction).

Now we discuss the feedback process for the case of the tip scanning over a step edge as shown in Fig. 13.5c. Initially, the amplitude setpoint A_1 stabilizes a frequency shift ω_1 and the corresponding tip-sample distance d_1 (working point 1 in Fig. 13.5a, b). If the tip approaches the step edge, the tip-sample distance decreases to d_2. This brings the tip into a region of larger (more negative) force gradient, shifting the resonance frequency by $\delta\omega$ to ω_2. This shift of the resonance frequency by $\delta\omega$ leads to an increase of the amplitude by δA to A_2 at ω_{drive} (working point 2 in Fig. 13.5a, b). The feedback acts on this deviation from the setpoint value A_1 by increasing the tip-sample distance d until the setpoint amplitude A_1 is restored to d_1.

In summary, a certain amplitude change corresponds to a certain resonance frequency shift, which corresponds to a certain tip-sample force gradient, which corresponds to a certain tip-sample distance $A(\Delta\omega(F'_{\mathrm{ts}}(d)))$. Therefore, keeping the feedback loop at a constant oscillation amplitude corresponds to establishing a constant average tip-sample distance d. An image scanned at constant tip-sample distance is called the topography. However, this assignment is only true if the *same* dependence of the frequency shift on the tip-sample distance (Fig. 13.5a) is present all over the surface.

Let us now consider scanning over a border with two different dependencies of the frequency shift as a function of tip-sample distance as shown in Fig. 13.6. This will lead to an apparent height contrast even if the actual height of the atoms in both

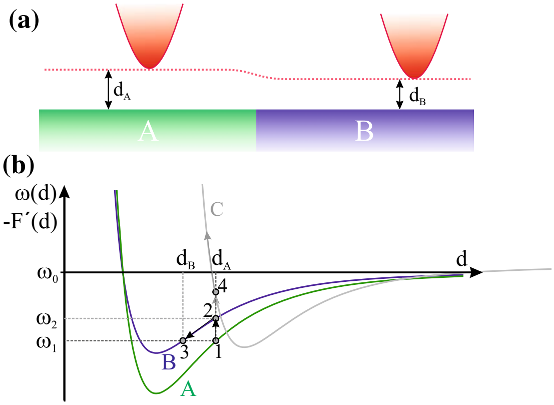


Fig. 13.6 **a** A scan from a region with material A to a region with material B can lead to a different apparent tip height in atomic force microscopy. **b** This arises due to the different force gradient-distance curves present in the two regions. For another force gradient-distance curve (*C*) an instability will occur due to the different sign of the slope of the force gradient, i.e. due to the non-monotonous dependence of the force gradient on the tip-sample distance

areas is the same. Initially the tip is in region A with the corresponding force gradient dependence shown in Fig. 13.6b. The setpoint frequency ω_1 stabilizes the tip-sample distance to d_A (working point 1). If by lateral scanning the tip crosses the border from A to B, the force gradient curve B in Fig. 13.6b applies, resulting in a different frequency shift ω_2 (working point 2). The feedback restores the setpoint frequency ω_1 by reducing the tip-sample distance, however, now the force-distance behavior of material B (blue curve) applies, resulting in a tip-sample distance d_B (working point 3). This leads to a reduced apparent height d_B during imaging and results an a material contrast between materials A and B.

While the assumed force gradient curve B resulted in a different apparent height in region B, more severe cases are also possible. Let us now assume the extreme case of the force gradient curve C in Fig. 13.6b. This case will lead to a jump to the working point 4 when the tip enters region C. At this working point the force gradient-distance curve has a negative slope and thus the feedback works in the wrong direction: The feedback will reduce the tip-sample distance in order to try to restore the larger (more negative) frequency shift setpoint. While this direction of feedback was the right one for a positive slope of the force gradient curve, it is the wrong feedback direction for the opposite slope at working point 4. The feedback will constantly reduce the tip-sample distance, leading to a tip crash. This shows that the non-monotonous dependence of the force gradient on the distance can lead to serious instabilities.

13.4 Experimental Realization of the AM Detection Mode

A scheme of the experimental setup for the amplitude modulation AFM detection is shown in Fig. 13.7. The sinusoidal driving signal at the fixed frequency ω_{drive} is generated by an oscillator. This signal excites the piezoelectric actuator driving the cantilever base, which results in turn in a cantilever oscillation amplitude A, which is, since it is close to resonance, much larger than the excitation amplitude. If tip and sample approach each other, the oscillation amplitude at the fixed excitation frequency ω_{drive} will change due to a shift of the resonance frequency induced by the tip-sample interaction, as discussed in the previous section. The cantilever deflection (sinusoidal signal) is measured, for instance, by the beam deflection method as indicated in Fig. 13.7. The signal from the split photodiode is converted by the preamplifer electronics to a voltage signal proportional to the cantilever deflection. This signal is an AC voltage signal at the frequency ω_{drive} with an amplitude proportional to the cantilever oscillation amplitude A.

Using a lock-in amplifier (described in Chap. 6), the amplitude of the AC signal at frequency ω_{drive} is measured. The lock-in amplifier needs the driving signal as a reference signal. At the output of the lock-in amplifier, a quasi-DC signal of the amplitude is obtained.[5]

This quasi-DC amplitude signal (demodulated from the AC signal at ω_{drive}) is used as the input signal for the z-feedback controller. The measured cantilever amplitude is compared to the setpoint amplitude. The controller determines an appropriate z-signal need to maintain a constant oscillation amplitude. Via the quite indirect relation between oscillation amplitude and tip-sample distance, maintaining a constant oscillation amplitude corresponds to maintaining a constant tip-sample distance. Thus, the z-feedback signal is used as the height signal, mapping the topography during data acquisition.

In the following, we describe the demodulated signal and the reaction of the feedback in more detail by considering the example of a scan over a step edge in the topography, as shown in Fig. 13.8 (see also Fig. 13.5c). As a starting condition, we assume that before scanning over a step edge the amplitude is nicely kept closely to the amplitude setpoint value A_1. When the step is approached laterally, the tip-sample distance will decrease from d_1 to d_2. This leads, as discussed in the last section, to a deviation of the oscillation amplitude from A_1 to A_2 which is measured as the demodulated amplitude signal at the output of the lock-in amplifier. Thus, this quasi-DC amplitude signal contains the deviations from the setpoint amplitude (due to the topography of the surface) before they are compensated by the feedback. Subsequently, this measured amplitude A_2 enters as input signal to the feedback

[5] Technically the driving signal can be considered as a carrier signal which is modulated by a low-frequency (quasi-DC) amplitude signal (deviations from the desired amplitude setpoint due to the sample topography). Then the task of the lock-in amplifier is the demodulation of the low frequency amplitude signal. The term AM demodulation is traditionally used in connection with the audio signal detection/demodulation in AM radio receivers. This is the reason why the term AM detection is used for this detection scheme.

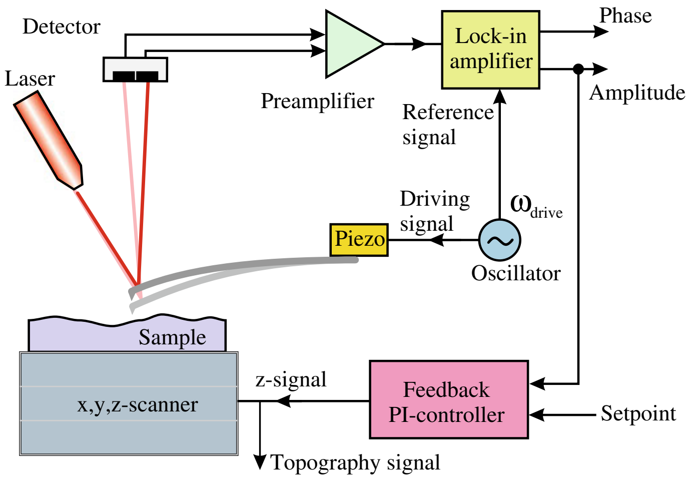
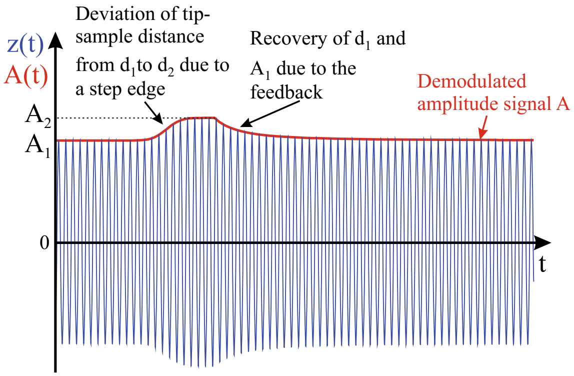


Fig. 13.7 Experimental setup for the AM detection scheme using a lock-in amplifier to detect the deviation of the oscillation amplitude from the setpoint value

Fig. 13.8 Demodulated amplitude signal *A* (red) and tip oscillation signal *z* as function of time when scanning over a topographic feature, e.g a step edge. The build-up of the amplitude signal and the reaction of the feedback are shown as sequential, while in reality they occur simultaneously



controller and deviations from the setpoint amplitude A_1 are compensated by the feedback which adapts the z-signal to a value equivalent to the step height. After this, the setpoint oscillation amplitude A_1 corresponding to the tip-sample distance d_1 is recovered.

A lock-in amplifier can also provide a phase signal, the difference between the phase of the cantilever oscillation and the phase of the driving signal. During a scan of the surface topography the phase signal can be recorded as free signal (i.e. not used for the feedback). This phase signal contains useful information on the tip-sample interaction, as we will discuss in Chap. 14.

The setup shown in Fig. 13.7 can also be used to record the resonance curve of the free cantilever not in contact with the sample. This is done by disabling the feedback and ramping the driving frequency over the resonance frequency, while measuring the oscillation amplitude and the phase. The measurement of the resonance curve allows to determine parameters like the resonance frequency ω_0, the Q-factor, and the amplitude at the resonance frequency $A(\omega_0) = A_{\text{free}}$. The value of ω_0 is needed to chose the driving frequency and A_{free} is needed to choose a proper amplitude setpoint.

A certain minimal detectable amplitude change in AM detection translates via the slope of the resonance curve to a minimal detectable frequency shift and finally to the resolution obtained for the tip-sample distance. The larger the slope of the resonance curve, the smaller the frequency shifts that can be detected for a given minimal detectable amplitude change. The slope of the resonance curve increases with increasing Q-factor. Thus, in AM detection the sensitivity for the detection of a frequency shift increases for higher Q-factors. However, as we will see in the following section, high Q-factors lead in the AM detection scheme to unacceptably long time constants (low bandwidth). Due to this the AM detection scheme is not used for cantilevers with Q-factors larger than about 500.

13.5 Time Constant in AM Detection

The time constant for AM detection can be obtained by analyzing the solution of the equation of motion for the driven damped harmonic oscillator (2.25). The change of the motion $z(t)$ in reaction to a changed tip-sample interaction can be modeled by an (instantaneous) change of the resonance frequency of the harmonic oscillator from ω_0 to ω_0'. Either a numerical solution of the equation of motion or an analytical solution can be analyzed.

According to Sect. 2.5 the analytic solution of the equation of motion of the driven damped harmonic oscillator after a change of the resonance frequency at time $t = 0$ can be written as

$$z(t > 0) = A' \cos(\omega_{\text{drive}} t + \phi') + G e^{-\omega_0' t/(2Q)} \cos(\omega_{\text{hom}} t + \phi). \qquad (13.13)$$

The first term corresponds to the new steady-state oscillation at the driving frequency ω_{drive} under the influence of the shifted resonance frequency ω_0'. The new steady-state amplitude A' and phase ϕ' are given by (2.32) and (2.35), respectively, replacing ω_0 by ω_0'. The second term in (13.13) corresponds to an exponentially decaying transient. G and ϕ are determined by the initial conditions and ω_{hom} is introduced in Sect. 2.2 and Sect. 2.5.

In Fig. 13.9a the envelope of the cantilever deflection $z(t)$ is plotted as a function of time for a Q-factor of 100, a resonance frequency $f_0 = 150\,\text{kHz}$, and an instantaneous increase of the resonance frequency by $\Delta f = f_0 - f_0' = 1319\,\text{Hz}$ at time zero.[6] The envelope of the cantilever deflection $z(t)$ is plotted, since a single oscillation is not

[6] This value for the frequency shift was chosen as it leads to half of the original amplitude in the steady-state.

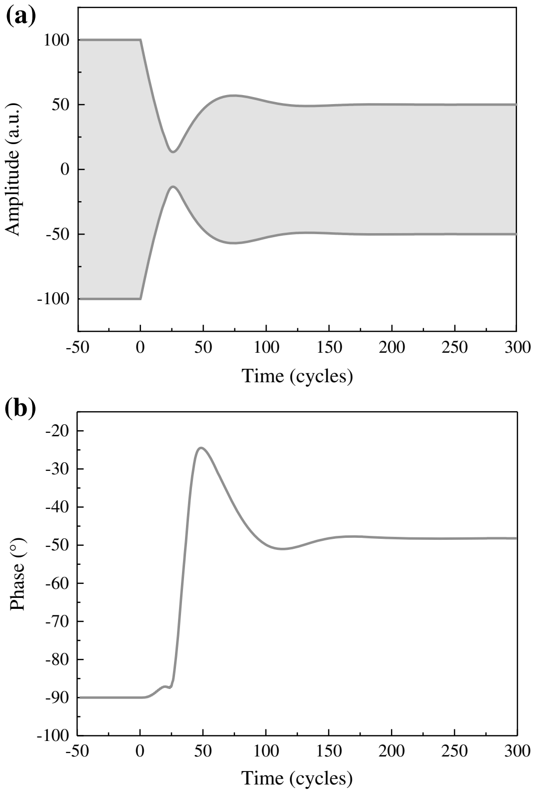


Fig. 13.9 The envelope of the oscillation amplitude (**a**) and the phase (**b**) in reaction to a change of the resonance frequency from ω_0 to ω_0' at time $t = 0$. The amplitude and phase response show that, after a transient, the new steady-state amplitude and phase are reached after about Q oscillations

visible on the time scale shown. The transient to the new steady-state amplitude is characterized by exponential behavior and a strong beat term. The new steady-state amplitude of half of the original amplitude is reached after about Q/π oscillations, corresponding to a time $\tau \approx Q/(f_0'\pi) = 0.2\,\text{ms}$ (cf. (2.43)). This time constant still allows for sufficiently fast scanning speeds in AFM scanning.

In Fig. 13.9b the time dependence of the phase is shown. The phase was determined from the cantilever deflection $z(t)$ numerically simulating a lock-in detection. Similar to the amplitude, also the phase reaches its new steady-state value after a transient of about Q/π oscillations.

For the case of a high Q-factor of 10,000, the time constant τ is 100 times larger, leading to unacceptably long scanning times when using cantilevers with a large Q-factor (i.e. in vacuum) in the AM detection mode. When the tip-sample interaction changes quickly, for instance during a fast scan over a sharp step edge, it takes roughly the time τ before the corresponding tip oscillation amplitude changes to its new steady-state value, corresponding to the new tip-sample distance. In the transient time until the new amplitude has been established a false amplitude enters into the feedback loop, which does not yet correspond to the actual new tip-sample distance. Thus, only after this settling time should the tip be moved on to the next measuring point. For cantilevers with a high Q-factor this results in an unacceptably long scanning time. Therefore, AM detection is not used for high Q cantilevers (i.e. in vacuum). For high Q cantilevers a different detection scheme (FM detection) is used, which will be discussed in Chap. 16. The AM detection scheme is used for cantilevers at ambient conditions, where the quality factor is less than several hundred due to dissipative damping in air.

The same conclusion about the time response of a harmonic oscillator can alternatively to the analysis of the equation of motion also be derived from an energy consideration. The energy which can be supplied by external driving to a harmonic oscillator per cycle can according to (2.44) be expressed as $2\pi E_{\text{osc}}/Q$. Therefore, it takes the time of about Q oscillations to change the oscillation state of the harmonic oscillator substantially (e.g. to decrease the amplitude to one half). The larger the Q-factor is, the smaller is the influence of the driving and the closer becomes the behavior of the driven oscillator to that of a free harmonic oscillator without damping.

13.6 Dissipative Interactions in the Dynamic AM Detection Mode

Up to now we have considered the AM detection method in the limit where the tip-sample interaction is conservative. As discussed, a conservative tip-sample interaction induces a shift of the resonance frequency of the cantilever. In this section, we will consider a model which includes dissipative tip-sample interactions in a crude way. To keep things simple, we will still deal with the small amplitude limit, i.e. an expansion of the tip-sample force up to the linear order is sufficient.

In the treatment of the harmonic oscillator in Chap. 2, dissipation was included by the Q-factor. The types of dissipative forces included via the Q-factor are forces proportional to the velocity, e.g. energy losses (damping) of a cantilever oscillating in air or a liquid. This cantilever dissipation energy $E^{\text{diss}}_{\text{cant}}$ per cycle leads according to (2.44) to a corresponding Q-factor $Q_{\text{cant}} \propto 1/E^{\text{diss}}_{\text{cant}}$. An additional dissipative tip-sample interaction is associated with a dissipated energy of $E^{\text{diss}}_{\text{ts}}$ per cycle and a corresponding Q-factor Q_{ts} can be assigned. As the dissipation energies add up to a total dissipation energy, the inverse Q-factors add up to an effective Q-factor as

$$E^{\text{diss}}_{\text{tot}} = E^{\text{diss}}_{\text{cant}} + E^{\text{diss}}_{\text{ts}} \propto \frac{1}{Q_{\text{cant}}} + \frac{1}{Q_{\text{ts}}} \equiv \frac{1}{Q_{\text{eff}}}. \tag{13.14}$$

To take a dissipative interaction into account via the corresponding Q-factor is a realistic model for a velocity dependent force, like damping in air. For a dissipative tip-sample interaction this treatment becomes questionable. Nevertheless, we will now consider the dissipative tip-sample interaction via the effective Q-factor, since in this case we can still use the previously derived equations for the amplitude and the phase (2.32) and (2.35) of a driven damped harmonic oscillator, respectively.

We use the effective quality factor and replace the resonance frequency of the free cantilever ω_0 by the shifted resonance frequency $\omega_0' \approx \omega_0 + \omega_0 k'/(2k)$, according to (13.9). In order to avoid too many subscripts we identify $\omega \equiv \omega_{\text{drive}}$. With this the amplitude and phase read as a function of the driving frequency ω, according to (2.32) and (2.35) as

$$A^2 = \frac{A_{\text{drive}}^2}{\left[1 - \left(\frac{\omega}{\omega_0'}\right)^2\right]^2 + \frac{1}{Q_{\text{eff}}^2}\left(\frac{\omega}{\omega_0'}\right)^2}, \tag{13.15}$$

and

$$\tan\phi = \frac{-\omega_0'\omega}{Q_{\text{eff}}\left(\omega_0'^2 - \omega^2\right)} = \frac{-\frac{\omega}{\omega_0'}}{Q_{\text{eff}}\left[1 - \left(\frac{\omega}{\omega_0'}\right)^2\right]}, \tag{13.16}$$

respectively.

In the following, we show that in AM detection it cannot be distinguished whether a conservative interaction (leading to a frequency shift) or a dissipative interaction (leading to a different Q-factor) is the reason for a certain measured amplitude change. We consider in the following the two limiting cases of only conservative interaction or only dissipative interaction.

In Fig. 13.10a the amplitude and phase for a free cantilever (blue curve: ω_0, Q_{cant}) are compared to the case in which a conservative tip-sample interaction is included (red curve: ω_0', Q_{cant}). In this case, the conservative tip-sample interaction leads to a shift of the whole resonance curve.[7] Due to the constant quality factor, the amplitude and shape of the resonance curve and phase curve do virtually do not change. This shift of the resonance curve and phase curve leads to a different amplitude and phase measured at the (fixed) driving frequency $\omega = \omega_{\text{drive}}$, as indicated by the vertical line in Fig. 13.10a. In this figure, the driving frequency was selected to be somewhat lower than ω_0.

The opposite assumption is that only the damping changes and the resonance frequency stays constant (ω_0, Q_{eff}). In this case, the frequency at which the maximal amplitude of the resonance curve occurs stays approximately constant very close to ω_0 with and without interaction (Fig. 13.10b), while the resonance curve and the phase as a function of frequency become broader with increasing damping (lower quality factor) as shown by the green curve in Fig. 13.10b. This leads to a reduced amplitude and also to a change of the phase shift at the driving frequency (vertical line in Fig. 13.10b).

As in the AM detection mode only the amplitude is measured, during scanning it is not possible to distinguish whether an amplitude change occurs due to a conservative interaction (resonance frequency shift due to topography) or due to a dissipative interaction (change of the Q-factor due to material properties). Both lead to a change

[7] The curves in Fig. 13.10 are plotted using (13.15) and (13.16). While the resonance curves for two different resonance frequencies do not exactly correspond to a shift of the resonance curve, Fig. 13.10a shows that these curves correspond to a very good approximation to a shift.

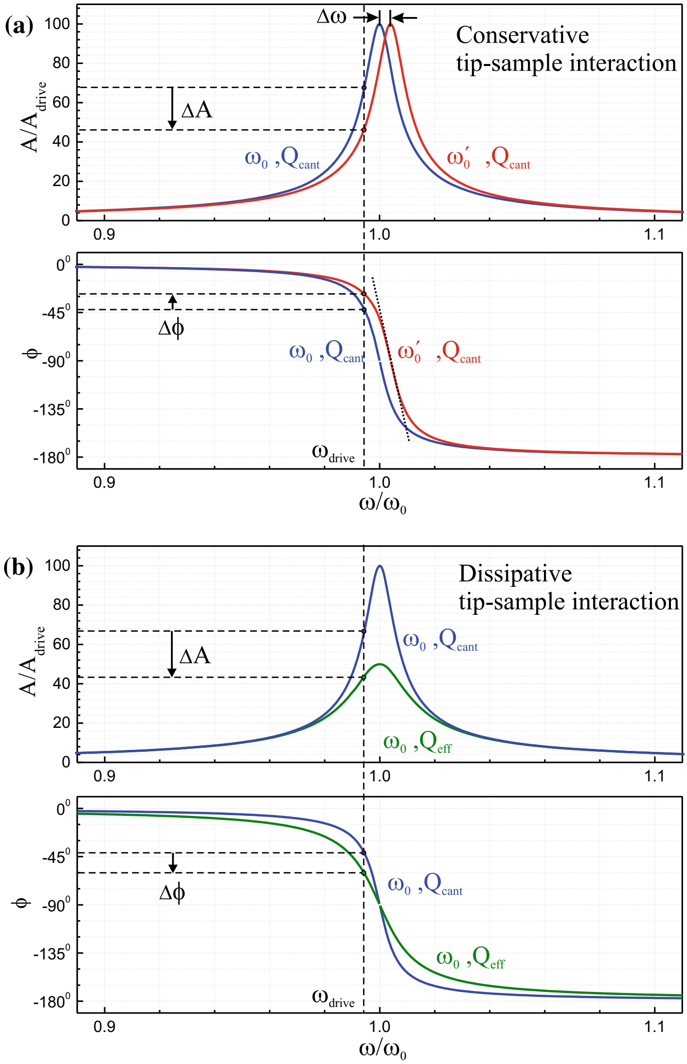


Fig. 13.10 **a** Amplitude and phase for a free cantilever (*blue curve*) compared to the case with a conservative tip-sample interaction included (*red curve*). The two resonance curves as well as the phase curves are shifted with respect to each other by $\Delta\omega$. **b** Amplitude and phase for a free cantilever compared to the case with a dissipative tip-sample interaction included (*green curve*), i.e. the effective quality factor decreases, while the resonance frequency stays constant. In both cases (**a**) and (**b**) the oscillation amplitude at ω_{drive} is reduced by the same amount, which makes it impossible to distinguish between a conservative and a dissipative interaction during scanning in the AM detection mode based on a measurement of the amplitude

of the amplitude at the driving frequency. It is not known whether an initial change of A during a scan (later compensated by the feedback loop) arises due to a change of ω_0 or Q.

The dependence of the amplitude on Q_{eff} can lead to a material contrast. If in two laterally adjacent areas the true height of the two different materials as well as the conservative tip-sample interactions are the same, different damping (different Q_{ts}) occurring due to the two different materials can lead to a different oscillation amplitude, which results, after restoration of the amplitude by the feedback, in an apparent height difference between the two materials due to the different tip-sample dissipation.

If both A and ϕ were measured (during scanning) it is in principle possible to use these two measured values and invert (13.15) and (13.16) for ω_0' and Q_{eff}. Since (13.15) and (13.16) are a rather complicated to solve, alternatively the complete resonance curves of amplitude and phase (like the ones shown in Fig. 13.10) can be measured in a spectroscopic type of measurement. The frequency shift can then be obtained from the position of the maximum in the amplitude or the frequency at which the phase is −90°. The damping Q_{eff} can be determined from the width of the resonance curve in amplitude or phase. All these measurements have to be performed without feedback and therefore require high stability (i.e. low drift). Further, these parameters can be obtained as a function of the tip-sample distance d at a specific location on the surface.

13.7 Dependence of the Phase on the Damping and on the Force Gradient

Generally, the dependence of the phase on the damping and on the force gradient is contained in (13.16). From Fig. 13.10, we can see that the dependence of the phase as function of frequency can be approximated as linear close to the (shifted) resonance at ω_0' or ω_0 at which $\phi = -90°$ (e.g. dotted line in Fig. 13.10a). In the following, we will derive this linear relation between phase and frequency for small frequency deviations $\delta\omega$ relative to ω_0'. Using in the nominator of (13.16), the approximation $\omega_0' \approx \omega_0$ and in the denominator the approximation $\omega_0' + \omega \approx 2\omega$, as well as subsequently the relation $\delta\omega = \omega - \omega_0'$, results in

$$\tan\phi = \frac{-\omega_0'\omega}{Q_{\text{eff}}\left(\omega_0'^2 - \omega^2\right)} \approx \frac{-\omega_0\omega}{Q_{\text{eff}}\left(\omega_0' + \omega\right)\left(\omega_0' - \omega\right)} \approx \frac{\omega_0}{2Q_{\text{eff}}\delta\omega}. \tag{13.17}$$

Close to the resonance, the phase will be close to −90° and the deviation from this value will be termed the phase shift $\delta\phi$ with $\phi = -\pi/2 + \delta\phi$. The arctan can be approximated in this case as $\arctan x \approx -\pi/2 - 1/x$, resulting in

$$\phi = -\frac{\pi}{2} + \delta\phi = \arctan\left(\frac{\omega_0}{2Q_{\text{eff}}\delta\omega}\right) \approx -\frac{\pi}{2} - \frac{2Q_{\text{eff}}}{\omega_0}\delta\omega. \tag{13.18}$$

Thus, the phase shift $\delta\phi$ relative to the phase $-90°$ results as

$$\delta\phi = -\frac{2Q_{\text{eff}}}{\omega_0}\delta\omega. \tag{13.19}$$

This equation can be used for the conversion between the frequency shift and the phase shift close to resonance.

If driving is performed at the free resonance frequency ω_0 (as will be considered in detail in the following chapter) and if we consider the phase shift due to the frequency shift induced by the tip-sample interaction $\Delta\omega$, then $\delta\omega$ becomes $\Delta\omega$ and (13.17) can be written using (13.10) as

$$\tan\phi \approx \frac{\omega_0}{2Q_{\text{eff}}\Delta\omega} \approx \frac{k}{Q_{\text{eff}}k'}. \tag{13.20}$$

The phase shift at ω_0 due to the tip-sample interaction results in the linear approximation according to (13.19) as

$$\Delta\phi = -\frac{2Q_{\text{eff}}}{\omega_0}\Delta\omega = \frac{Q_{\text{eff}}k'}{k} = -\frac{Q_{\text{eff}}}{k}\frac{\partial F_{\text{ts}}}{\partial z}\bigg|_{z=0}. \tag{13.21}$$

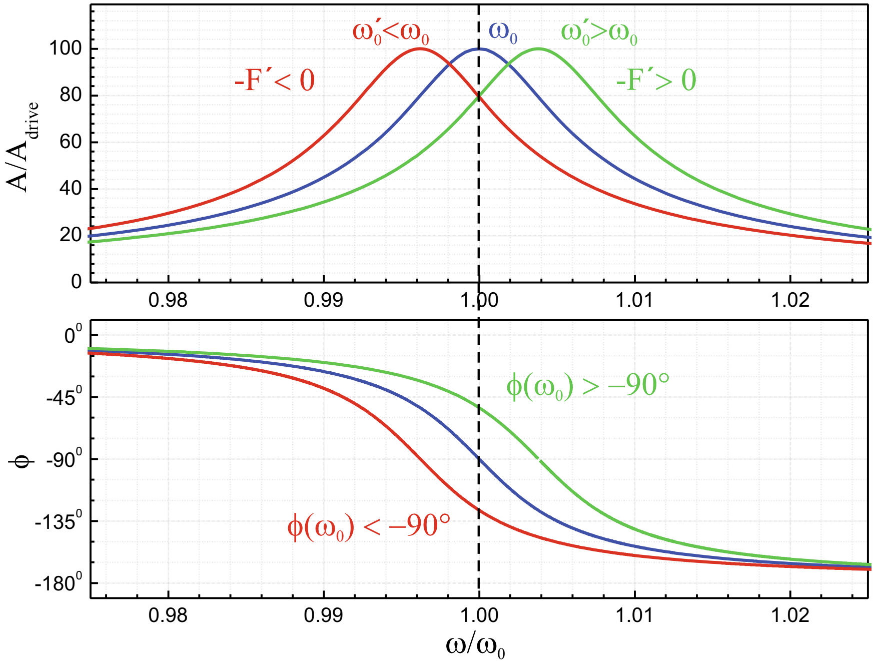


Fig. 13.11 Shift of the resonance curves (amplitude and phase) under the influence of a force gradient F' due to the tip-sample interaction

The phase shift depends linearly on both the effective quality factor and the force gradient of the tip-sample interaction. Since the phase depends on $\Delta\omega$ and Q_{eff} in a different way than the amplitude (cf. the arrows in Fig. 13.10 and (13.15)), the phase recorded as a free signal (not used for the feedback) can result in a different contrast (phase contrast) than the amplitude signal.

According to (13.21), the sign of the force gradient determines the sign of the phase shift at ω_0, since Q_{eff} is always positive. For attractive forces (more precisely, positive force gradients) the phase is more negative than $-90°$ ($\phi < -90°$), and correspondingly for repulsive forces (negative force gradients) the relation $\phi > -90°$ holds for the phase.

As a graphic summary we show in Fig. 13.11 the resonance curves for amplitude and phase (according to (13.15) and (13.16)) for a driven damped harmonic oscillator under the influence of an external force gradient F' which shifts the resonance curve $A(\omega)$ and the phase behavior $\phi(\omega)$.

13.8 Summary

- If the tip oscillation amplitude is small, the tip-sample interaction can be described by a second spring with small spring constant k' acting between tip and sample additionally to the cantilever spring k. The spring constant k' is given by the negative force gradient of the tip-sample interaction.
- The frequency shift of the resonance frequency under the influence of a conservative tip-sample interaction is given by

$$\Delta\omega = \omega_0 \frac{k'}{2k} = -\frac{\omega_0}{2k} \frac{\partial F_{\mathrm{ts}}(d+z)}{\partial z}\bigg|_{z=0} . \tag{13.22}$$

  This equation holds if the tip-sample force can be approximated as linear within the range of the oscillation amplitude and if $|k'| \ll k$.
- Roughly, the frequency shift $\Delta\omega$ is positive (towards higher frequencies) for repulsive forces and negative for attractive forces.
- In the amplitude detection mode (AM), the cantilever is driven at a fixed frequency and amplitude. The oscillation amplitude (and phase) is measured using the lock-in technique and used as the feedback signal.
- The measured oscillation amplitude depends on the frequency shift of the resonance curve induced by the force gradient of the tip-sample interaction, which in turn depends on the tip-sample distance $A(\Delta\omega(F'_{\mathrm{ts}}(d)))$. Feedback on constant oscillation amplitude corresponds to constant frequency shift and finally constant tip-sample distance.
- The non-monotonous dependence of the frequency shift on the tip-sample distance can lead to instabilities in the feedback behavior.

- A measured change of the amplitude (phase) during imaging in the AM mode can be induced by a frequency shift (conservative interaction) as well as by a change in quality factor (dissipative interaction).
- The phase shift close to the resonance is proportional to the frequency shift as $\delta\phi = -\frac{2Q_{\text{eff}}}{\omega_0}\delta\omega$. Thus, the phase shift depends linearly on Q_{eff} and the force gradient.

References

1. G. Binnig, C.F. Quate, Ch. Gerber, Atomic force microscope. Phys. Rev. Lett. **49**, 57 (1982). https://doi.org/10.1103/PhysRevLett.56.930
2. Y. Martin, C.C. Williams, H.K. Wickramasinghe, Atomic force microscope force mapping and profiling on a sub 100 Å scale. J. Appl. Phys. **61**, 4723 (1987). https://doi.org/10.1063/1.338807
3. R. Garcia, *Amplitude Modulation Atomic Force Microscopy*, 1st edn. (WileyVCH, Weinheim, 2010). https://doi.org/10.1002/9783527632183. ISBN:9783527408344

Chapter 14
Intermittent Contact Mode/Tapping Mode

While the previous chapter was aimed at providing a basic understanding of dynamic atomic force microscopy, we turn now to the intermittent contact mode (or tapping mode) which is the mode that is used most frequently at ambient conditions. In the intermittent contact mode the oscillation amplitude is large compared to the range of the tip-sample force and ranges from large distances with negligible tip-sample interactions deep into the repulsive regime. For these large oscillation amplitudes, the linear approximation of the tip-sample force used so far in the AM mode is no longer valid. Due to this, the harmonic oscillator becomes an anharmonic oscillator and an analytical solution of the equation of motion becomes difficult. We will derive general dependencies (for instance via the law of energy conservation) or we use the results from numerical solutions of the equation of motion. We will see that the resonance curve of an anharmonic oscillator is very different from the usual case of a harmonic oscillator. Thus, concepts used for the harmonic oscillator like the frequency shift of the resonance curve cannot be directly applied to the intermittent contact mode.

While operating with much larger amplitudes, the tapping mode has similarities to the AM detection mode discussed in Chap. 13. In both modes the cantilever is excited at a fixed driving frequency and the measured quantity is the oscillation amplitude. In the tapping mode, the amplitude depends monotonously on the tip-sample distance, which avoids instabilities in the feedback control of the tip-sample distance. Finally, we discuss how the dissipative tip-sample interactions are related to the phase of the oscillation in the intermittent contact mode.

14.1 Dynamic Atomic Force Microscopy with Large Oscillation Amplitudes

In the intermittent contact mode, the oscillation amplitude is quite large (typically several tens of nanometer) and cantilever force constants of typically several tens of N/m are used. As the name intermittent contact mode suggests, the tip comes into intermittent contact with the sample, which leads to very strong short-range force

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3_14

contributions close to the sample. In tapping mode, the constant driving frequency is usually selected at or very close to the resonance frequency of the free cantilever (not at maximum slope, as in the AM slope detection mode cf. Sect. 13.3). The measured signal is the amplitude A, which contains information on the average tip-sample distance d. In order to maintain an oscillation of the tip, snap-to-contact has to be prevented, which is possible when using large oscillation amplitudes, as discussed in Sect. 10.5.

First we consider a purely conservative tip-sample interaction, i.e. the only dissipation present is the (air) damping of the cantilever described by the corresponding Q-factor. In a later section also dissipative tip-sample interactions will be included. Figure 14.1 shows the tip-sample force F_{ts} and the cantilever force as a function of the momentary tip-sample distance $d + z$. The average tip-sample distance is d, i.e. $z = 0$. In most of the amplitude range $2A$ the tip-sample force is negligible and the sum of the tip-sample force and the spring force is linear with z. However, close to the lower turnaround point (lowest z-value) strong deviations from linear force-distance behavior occur due to the strong repulsive tip-sample force. Due to this strongly non-linear force-distance behavior we do no longer use the approximation for a harmonic oscillator. Accordingly, we cannot use the concept of the frequency shift of the whole resonance curve introduced in the previous chapter.

At large tip-sample distances the tip oscillates at its free resonance frequency ω_0 with its (large) free resonance amplitude A_{free}. When the tip is brought towards

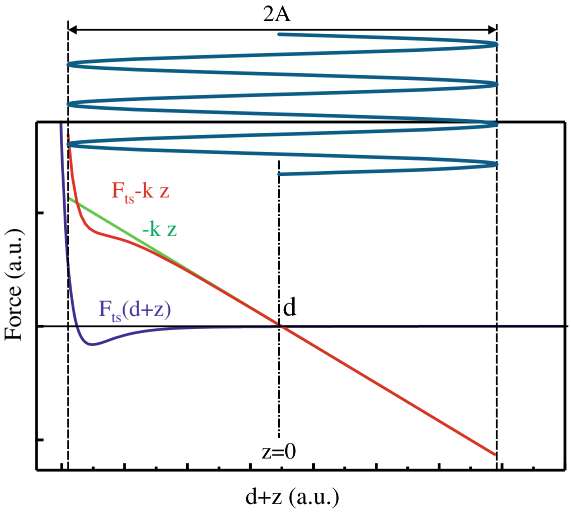


Fig. 14.1 Force-distance dependence of the cantilever force (*straight green line*), the tip-sample force (*blue line*), and the total force (*red line*) as a function of the momentary tip-sample distance $d + z$. In tapping mode, the range of the amplitude $2A$ is so large that it extends from almost zero tip-sample force at the upper turning point to deep in the repulsive regime at the lower turning point. The total force displays non-linear behavior corresponding to an anharmonic oscillator; in spite of this the oscillation path is still very close to sinusoidal

the surface, it will eventually reach the repulsive interaction regime and it might be assumed that the trajectory of the oscillation should deviate strongly from a sinusoidal shape due to the very strong repulsive force. However, it appears (experimentally [1] and from simulations [2]) that the oscillation trajectory can still be approximated with very high precision as a sinusoidal shape. This sinusoidal oscillation is an important fact in understanding the tapping mode. While the form of the oscillation stays sinusoidal even in a strongly anharmonic potential, the amplitude changes due to the strong repulsive interaction.

As an example, the oscillation traces for two different average tip-sample distances d are shown in Fig. 14.2a, when operation is performed in constant height mode, i.e. without feedback, restoring an amplitude setpoint. It was found experimentally [1] and also from simulations [2] that the oscillation amplitude reduces approximately linearly with decreasing average tip-sample distance d, once the oscillation path reaches the repulsive regime, as shown in Fig. 14.2b. Actually, a slope of one (i.e. $A \approx d$) is usually observed due to the short range repulsive force. In tapping mode detection, a certain amplitude A (corresponding to the average tip-sample distance $d \approx A$) is chosen as the amplitude setpoint for the z-feedback. If the tip-sample interaction is approximated as a hard wall, a closed form (linear) expression for the oscillation amplitude as function of the tip-sample distance is obtained [2].

One reason why the tapping mode is so popular is that the dependence between the measured signal (oscillation amplitude) and the tip-sample distance is monotonous.

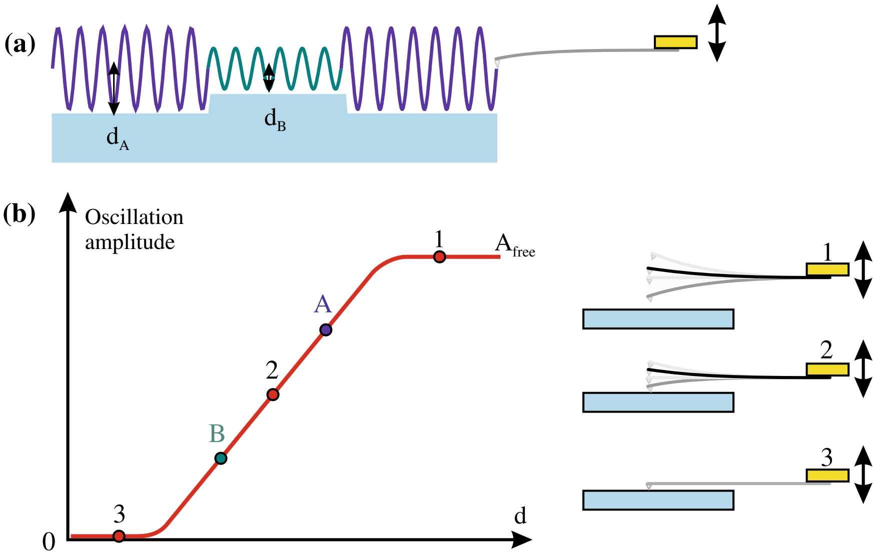


Fig. 14.2 **a** Schematic of the tip oscillation for two different average tip-sample distances d_A and d_B. The oscillation remains sinusoidal also at reduced distances d. **b** The vibration amplitude (being the free amplitude A_{free} for large tip-sample distances) decreases with decreasing tip-sample distance d, once the oscillation path reaches the repulsive range, i.e. $d < A_{\text{free}}$

This allows for a robust feedback signal and avoids the possibility of instabilities which can occur if the measured signal depends on the tip-sample interaction in a non-monotonous way (cf. Sects. 13.3 and 16.3).

In the following, we will provide a semi-quantitative explanation for the amplitude reduction if the oscillation enters the regime of strong (repulsive) interaction. No dissipative tip-sample interaction is needed in order to explain the reduced oscillation amplitude. The amplitude reduction can be understood within the model of a driven oscillator (not harmonic).

We will discuss the energy flow supplied to the oscillator by the driving oscillation with amplitude A_{drive}. Initially, the tip-sample distance is large, and we assume that the oscillator is driven at its free resonance frequency ω_0. This leads to an oscillation with the resonance amplitude $A_{\text{free}} = QA_{\text{drive}}$. At the resonance, the phase between the driving oscillation and oscillator motion is $-90°$ resulting in a maximal energy transfer from the excitation to the oscillation. Due to a tip-sample interaction in the intermittent contact mode (assumed to be conservative) the phase of the oscillation will deviate from its value of $-90°$ for the free cantilever, leading to a reduced amplitude. Off-resonance the energy transfer from the external excitation to the oscillator is (much) less efficient resulting in a reduced oscillation amplitude. Let us consider this idea in a more quantitative manner.

Due to the strong effects of anharmonicity in the tapping mode, we do not use any of the results previously obtained for the harmonic oscillator, e.g. shape of the resonance curve, phase curve, or the concept of frequency shift of the whole resonance curve. The following analysis of the driven anharmonic oscillator is very general, only relying on (a) the (experimentally proven) assumption of a sinusoidal oscillation and (b) on the general law of energy/power conservation. We consider a driven damped oscillator with the cantilever base (or the driving piezo) oscillating as $z_{\text{drive}} = A_{\text{drive}} \cos(\omega t)$. The resulting sinusoidal motion of the tip relative to its equilibrium position d can be written in the steady-state as $z = A\cos(\omega t + \phi)$. The average power supplied by driving the cantilever base can be written as

$$\langle P_{\text{drive}} \rangle = \langle F \cdot \dot{z}_{\text{drive}} \rangle = \frac{1}{T} \int_0^T k\left[z_{\text{drive}}(t) - z(t)\right] \dot{z}_{\text{drive}}(t)\, \mathrm{d}t. \tag{14.1}$$

Since all the functions in the integral are simple harmonic functions, the integral can be solved analytically, resulting in

$$\langle P_{\text{drive}} \rangle = -\frac{1}{2} k A_{\text{drive}} A \omega \sin\phi. \tag{14.2}$$

This expression is valid very generally, it is not necessary to assume that the driving frequency ω is close to the resonance frequency. It can be seen that the maximum power is delivered if the phase is $-90°$. This power supplied will be dissipated by the (air) damping of the cantilever Q_{cant} (since we assumed a purely conservative tip-sample interaction).

If we further consider that the energy stored in the oscillator close to resonance is $E_{\mathrm{osc}} \approx 1/2\, kA^2$, and the energy supplied by the driving and then dissipated during one cycle is $E_{\mathrm{drive}} = \langle P_{\mathrm{drive}} \rangle T$, the quality factor Q_{cant} can (according to (2.44)) be written as

$$Q_{\mathrm{cant}} = 2\pi \frac{E_{\mathrm{osc}}}{E_{\mathrm{drive}}} = \frac{-A}{A_{\mathrm{drive}} \sin\phi}. \tag{14.3}$$

If we identify the oscillation amplitude of the free cantilever (without any tip-sample interaction) as $A_{\mathrm{free}} = Q_{\mathrm{cant}} A_{\mathrm{drive}}$, the oscillation amplitude can be written as

$$A = A_{\mathrm{drive}} Q_{\mathrm{cant}} \sin(-\phi) = A_{\mathrm{free}} \sin(-\phi) \tag{14.4}$$

and thus

$$\frac{A}{A_{\mathrm{free}}} = \sin(-\phi). \tag{14.5}$$

This shows that the amplitude decreases as the phase deviates from the resonance case −90° due to a tip-sample interaction. The energy from the excitation (driving) can no longer be effectively transferred to the oscillating cantilever. A change of the resonance condition due to a conservative tip-sample interaction leads to an excitation (driving) of the oscillator off-resonance and reduces thus the amplitude.

The dependence of the phase ϕ on the amplitude A/A_{free} according to (14.5) is shown in Fig. 14.3. This result, should be consistent with the specific result obtained for the harmonic oscillator. At the first sight it is not obvious that the resonance curve $A(\omega)$ and phase curve $\phi(\omega)$ of the harmonic oscillator lead to (14.5). However, we can derive an expression $\phi(A/A_{\mathrm{free}})$ from (2.32) and (2.35) by eliminating the dependence on ω, and (14.5) results.[1]

From Fig. 14.3 we see that an oscillation with a certain amplitude A can be realized at two different phases, lower and higher than the phase of the free cantilever at resonance, i.e. −90°. In the following we will show that $\phi < -90°$ corresponds to a net attractive tip-sample force, while $\phi > -90°$ corresponds to a net repulsive tip-sample force.

We start from the equation of motion for the driven damped harmonic oscillator (2.25) and include the static deflection introduced in Fig. 13.1, as well as the tip-sample force. The anharmonicity enters by using the full anharmonic tip-sample force F_{ts}, instead of the linear approximation. This results in

$$m\ddot{z} = -\frac{m\omega_0}{Q_{\mathrm{cant}}}\dot{z} - k(z - (z_{\mathrm{drive}} + \Delta L)) + F_{\mathrm{ts}}(d + z). \tag{14.6}$$

[1] The phase $\phi(A/A_{\mathrm{free}})$ can be obtained numerically from (2.32) and (2.35). If this result is plotted in Fig. 14.3 it is indistinguishable on top of the curve obtained from (14.5). Alternatively (2.32) and (2.35) can be rearranged analytically leading to (14.5) in a very good approximation.

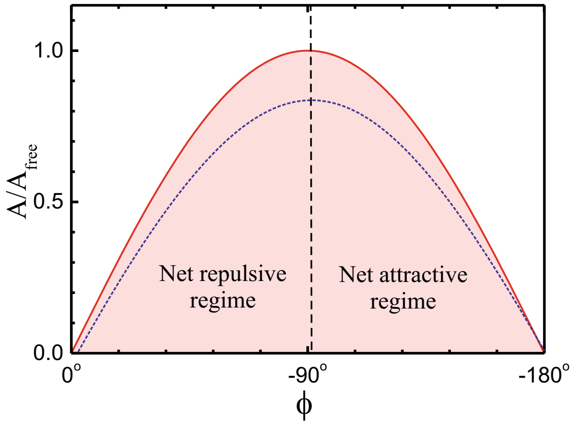


Fig. 14.3 Dependence of the amplitude on the phase according to (14.5). This expression is obtained from energy conservation and the assumption of a sinusoidal oscillation. For a given amplitude A/A_{free}, the phase can have two different values. $\phi < -90°$ corresponds to a net attractive tip-sample force, while $\phi > -90°$ corresponds to a net repulsive tip-sample force. The *dotted curve* results for a dissipative interaction and will be considered later

For simplicity, we consider the driving frequency at the resonance frequency of the free cantilever, $\omega_{\mathrm{drive}} = \omega_0$. Thus $z_{\mathrm{drive}} = A_{\mathrm{drive}} \cos \omega_0 t$. The resulting cantilever oscillation z is assumed to be sinusoidal. In a more general treatment $z(t)$ can be written as a Fourier series including multiples of the oscillation frequency [3, 4]. Throughout this text we consider only the first term of this Fourier expansion and the cantilever oscillation z and its time derivatives can be written as

$$z = A \cos (\omega_0 t + \phi) \ , \tag{14.7}$$

$$\dot{z} = -\omega_0 A \sin (\omega_0 t + \phi) \ , \tag{14.8}$$

$$\ddot{z} = -\omega_0^2 A \cos (\omega_0 t + \phi) = -\omega_0^2 z. \tag{14.9}$$

If we insert this into (14.6), the following equation results

$$-m\omega_0^2 z = \frac{m\omega_0^2 A}{Q_{\mathrm{cant}}} \sin (\omega_0 t + \phi) - k(z - \Delta L) + F_{\mathrm{ts}}(d + z) + kA_{\mathrm{drive}} \cos (\omega_0 t) \, . \tag{14.10}$$

Since $m\omega_0^2 = k$, the term on the left side of (14.10) cancels out the term $-kz$ on the right side. Instead of solving the equation of motion (which is difficult due to the non-linear dependence of $F_{\mathrm{ts}}(d + z)$), we will perform a kind of averaging over (14.10) in order to derive a useful expression between the relevant quantities. We multiply (14.10) by $A \cos (\omega_0 t + \phi)$ and integrate over one period. It can be seen form the symmetry of the expressions that the first and the second term on the right side are zero after multiplication and integration over one period. Thus, the remaining equation reads as

$$A\int_0^T F_{\mathrm{ts}}(d+z)\cos(\omega_0 t+\phi)\,\mathrm{d}t = -kAA_{\mathrm{drive}}\int_0^T \cos(\omega_0 t)\cos(\omega_0 t+\phi)\,\mathrm{d}t. \tag{14.11}$$

The integral on the right side results as $1/2T\cos\phi$. Thus (14.11) can be written as

$$\frac{1}{T}\int_0^T F_{\mathrm{ts}}(d+z)A\cos(\omega_0 t+\phi)\,\mathrm{d}t \equiv \langle F_{\mathrm{ts}}\cdot z\rangle = -\frac{1}{2}kAA_{\mathrm{drive}}\cos\phi. \tag{14.12}$$

If we finally use $A_{\mathrm{free}} = Q_{\mathrm{cant}}A_{\mathrm{drive}}$, the cosine of the phase results as[2]

$$\cos\phi = \frac{-2Q_{\mathrm{cant}}}{kAA_{\mathrm{free}}}\langle F_{\mathrm{ts}}\cdot z\rangle\,. \tag{14.13}$$

When analyzing this equation we have to consider that z is negative in the range where F_{ts} is different from zero (i.e. close to the lower turnaround point), cf. Fig. 14.1. Thus, an attractive (negative) force will lead to a positive $\langle F_{\mathrm{ts}}\cdot z\rangle$ and finally via (14.13) to a phase $\phi < -90°$. Correspondingly, a repulsive force $F_{\mathrm{ts}} > 0$ leads to a phase $\phi > -90°$. If $\langle F_{\mathrm{ts}}\cdot z\rangle = 0$ this leads to the resonance phase of $\phi = -90°$.

Generally, during one oscillation cycle, attractive as well as repulsive interactions will be "visited" by the tip. The terms "net attractive" or "net repulsive" force correspond to $\langle F_{\mathrm{ts}}\cdot z\rangle$ being positive or negative, respectively. If we have our working point in the tapping mode at a certain amplitude, but if we do not know whether this corresponds to the net attractive or repulsive regime, we can use the phase in order to obtain this important information, as also indicated in Fig. 14.3. In this way, the measurement of the phase provides an unambiguous distinction between net attractive and net repulsive interactions.

14.2 Resonance Curve for an Anharmonic Force-Distance Dependence

The results in the previous section were obtained using very general considerations, either energy considerations, or integration over the equation of motion. Alternatively, the equation of motion for an anharmonic oscillator can be solved. This can

[2] If we approximate the tip-sample force by $F_{\mathrm{ts}} = -k'z$ (harmonic oscillator), $\langle F_{\mathrm{ts}}\cdot z\rangle = -1/2\,k'A^2$ results (cf. (16.10)). Inserting this into (14.13) and remembering that according to (14.5) $A/A_{\mathrm{free}} = -\sin\phi$, the following expression for the phase is obtained $\tan\phi = k/(k'Q_{\mathrm{cant}})$, which corresponds to expression (13.20) obtained for the harmonic oscillator.

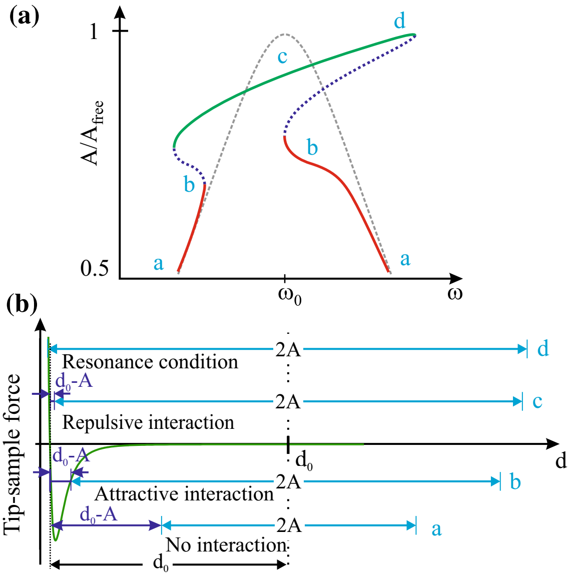


Fig. 14.4 **a** Resonance curve of an anharmonic oscillator (*solid line*) for a fixed average tip-sample distance $d = d_0$, compared to the free oscillation (*dashed gray curve*). For an anharmonic interaction the resonance curve becomes multivalued. The low-amplitude branch is shown in *red* and the high-amplitude branch in *green*. **b** The oscillation ranges corresponding to the regions $a - d$ of the resonance curve (**a**) are indicated in a plot of the force-distance curve

be attempted either analytically [5], or by evaluating the solution of the equation of motion numerically for a particular model of the tip-sample force [6]. If the tip approaches the sample, the anharmonicity increases and the resonance curve evolves from the well-known form, indicated as dotted gray line in Fig. 14.4a, to odd shapes, for instance that shown in color in Fig. 14.4a.

One of the main differences between a harmonic oscillator and an anharmonic oscillator is that the resonance frequency (i.e. the frequency at which the phase is $-90°$) of an anharmonic oscillator changes with the amplitude ($\omega_0' = \omega_0'(A)$), while for a harmonic oscillator the resonance frequency is independent of the oscillation amplitude. Thus, in a simplified reasoning for each segment on the resonance curve (with different amplitude) a different resonance frequency applies for the anharmonic oscillator. This leads to oddly shaped resonance curves, since not the whole resonance curve shifts, but parts of the resonance curve shift differently due to their different amplitudes. In the following we will qualitatively explain the peculiar shape of the resonance curve for an anharmonic oscillator as shown in Fig. 14.4a. In this figure the average tip-sample distance is fixed at d_0 and considered to be so close to the surface that the turnaround point close to the surface reaches into the regime of repulsive interaction at the maximum amplitude.

The numerical solutions of the equation of motion [6] show that the following general rule still holds: An attractive interaction shifts the resonance frequency to lower frequencies, while a repulsive interaction shifts the resonance frequency to higher frequencies. However, in contrast to the case of the harmonic oscillator the resonance curve does not shift homogeneously as a whole. For the anharmonic oscillator we have to apply this shift rule locally, i.e. individually for certain amplitudes of the resonance curve. According to Fig. 14.4b, an increasing oscillation amplitude corresponds to an decreasing distance between the sample surface and the lower turnaround point $d_0 - A$.

For frequencies (much) lower than the resonance frequency the amplitude is small (off-resonance), and does not reach the regime of tip-sample interaction, as shown in Fig. 14.4b. Therefore, the resonance curve is very close to the resonance curve of the free cantilever (region *a* in Fig. 14.4a, no shift of the resonance curve). Closer to the free resonance frequency the oscillation amplitude increases and at the turnaround point close to the surface the tip reaches the regime of attractive tip-sample interaction, as shown in Fig. 14.4b. This results effectively in a local downshift of the resonance frequency explaining the "ear" seen to the left in region *b* in Fig. 14.4a.[3] As explained below, the oscillation state of the harmonic oscillator can swich to a high amplitude branch (green in Fig. 14.4a), resulting in smaller tip-sample distances at the turnaround point close to the surface. The resulting repulsive interaction leads to a local upward shift of the resonance curve (region *c* in Fig. 14.4a). If the tip comes even closer to the surface at the lower turnaround point, this leads to a further upwards shift of the resonance frequency, leading to the "ear" seen to the right in region *d* in Fig. 14.4a and the resonance is reached ($A = A_{\text{free}}$ and $\phi = -90°$).

As seen from Fig. 14.4a, for certain ranges of frequencies the resonance curve of an anharmonic oscillator becomes multivalued. The solutions shown as blue dotted lines are unstable [5], while the low-amplitude branch (red in Fig. 14.4a) and the high-amplitude branch (green) correspond to two stable solutions of the equation of motion for a specific driving frequency ω. This coexistence of two oscillation states (with different amplitudes) for the same external conditions (ω_{drive}, A_{drive}) is a characteristic of the anharmonic oscillator. As we will see in the following, abrupt switches between these branches can occur. While the resonance curve was discussed here as a function of the driving frequency $\omega \equiv \omega_{\text{drive}}$, in tapping mode atomic force microscopy the driving frequency is kept constant and we will discuss this case in the following.

[3] Correspondingly, the left "ear" also occurs on the high-frequency side of the resonance curve.

14.3 Amplitude Instabilities for an Anharmonic Oscillator

In Fig. 14.5a we show the oscillation amplitude as a function of the average tip-sample distance d with the oscillation excited at the free resonance frequency $\omega_{\text{drive}} = \omega_0$. This figure shows the reduction of the amplitude for decreasing tip-sample distance, as already shown in Fig. 14.2b. Additionally, often a switching between the high-amplitude branch and the low-amplitude branch (present due to the anharmonicity) is observed as shown in Fig. 14.5a. The tip-sample approach is shown in red while the retraction is shown in green.

The jumps shown in Fig. 14.5a can be explained considering the resonance curves shown in Fig. 14.6a–c for different average tip-sample distances during approach and retraction (d_1, d_2, and d_3). The excitation is considered to be at the free resonance frequency of the cantilever ω_0.

As discussed above, the anharmonic tip-sample interaction leads to a distortion of the resonance curve with multivalued segments, instead of the simple shape of the resonance curve for a harmonic interaction. For a relatively large average tip-sample distance of d_1, the "ear" on the low-frequency side of the resonance curve visible in Fig. 14.6a arises due to the attractive interaction. This assignment can be made (locally) in the sense that the frequency shift to lower frequencies occurs for an attractive interaction, found in the harmonic case. Thus, a "local shift" of the resonance occurs for amplitudes at which the tip dives into the corresponding interaction zone. Due to this local shift of the resonance curve the amplitude at the free resonance

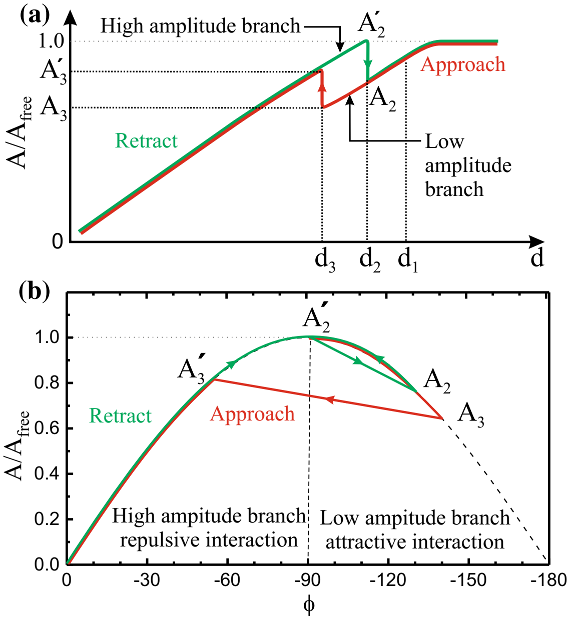


Fig. 14.5 **a** Amplitude-distance curves with jumps between the low-amplitude branch and the high-amplitude branch shown for approach (*red*) and retraction (*green*). **b** Phase as function of the oscillation amplitude during approach (*red*) and retraction (*green*). Phase values below the −90° line correspond to a net attractive interaction (low-amplitude branch), while phase values above the −90° line correspond to a net repulsive interaction (high-amplitude branch)

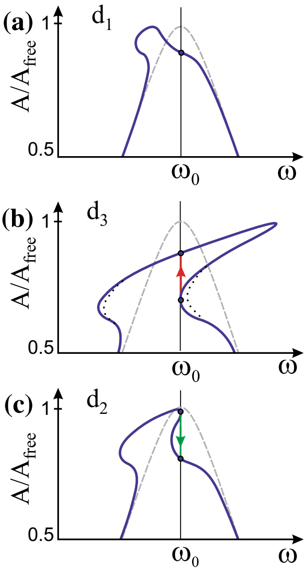


Fig. 14.6 Resonance curves for different average tip-sample distances d. The driving frequency is considered to be at the resonance of the free cantilever ω_0. **a** For large tip-sample distances (around d_1), the tip reaches only the attractive regime at the lower turnaround point, leading to an "ear" on the low-frequency side. **b** At smaller tip-sample distances of about d_3 the lower branch disappears at ω_0 and a jump to the high-amplitude branch occurs (*red arrow*). **c** If the tip-sample distance increase again, the oscillation stays on the high-amplitude branch until the "ear" on the high-frequency side disappears and the jump back to the low-amplitude branch occurs (*green arrow* in (**c**)). This figure is adapted from [6]

frequency is already reduced relative to the free amplitude (formation of the "ear" in Fig. 14.6a).

For smaller tip-sample distances d_3, an "ear" develops on the high-frequency side (Fig. 14.6b) for large amplitudes due to the repulsive tip-sample interaction. Due to this "ear" a low-amplitude branch and a high-amplitude branch develop at ω_0. In Fig. 14.6b the situation is shown in which the low-amplitude branch of oscillation disappears at ω_0. The dotted line in Fig. 14.6b indicates the situation for tip-sample distances slightly smaller than d_3, where no low-amplitude branch exists anymore at ω_0. The oscillation switches abruptly to the high-amplitude branch indicated by the red arrows in Figs. 14.5a and 14.6b. The difference in amplitude between the two branches is (only) about 1 nm. With the tip in the high-amplitude branch the amplitude decreases when the tip approaches closer to the surface, i.e. for smaller d (Fig. 14.5a).

When the tip is subsequently retracted from the sample, the high-amplitude branch disappears at ω_0 for a tip-sample distance larger than d_2 and the oscillation returns abruptly to the low-amplitude branch (green arrows in Figs. 14.5a and 14.6c). Working in the bistable tip-sample distance regime, where the high- and the low-amplitude modes exist, can always lead to the danger of switching between these solutions due to noise or feedback problems at sharp features in the topography. In this case, an amplitude setpoint outside the bistable region should be chosen.

While the switching between the two branches can occur as described above, there are also circumstances in which one branch is stable. For instance for low (free) oscillation amplitudes it was observed that the oscillation remains in the low amplitude branch for all tip-sample distances [7]. Moreover, we considered here only conservative tip-sample interactions for which the maximal amplitude remained constant, as seen in Fig. 14.6. For dissipative interactions the amplitudes can decrease. Hysteretic dissipative interactions can occur due to a water capillary neck which can form under ambient conditions between tip and sample. These interactions can modify the switching between the two branches of oscillation [7].

Since the difference in the oscillation amplitude between the high-amplitude and the low-amplitude branches is small ($\sim$1 nm), a way to identify in which branch the cantilever is oscillating is desired. As we will show in the following, this assignment can be made via the phase ϕ. According to (14.13), $\phi < -90°$ corresponds to a net attractive interaction, while $\phi > -90°$ corresponds to a net repulsive interaction.

In Fig. 14.5b, the double-valued dependence of the phase on the amplitude according to (14.5) is plotted as a dashed gray line. The evolution of the phase in the intermittent contact mode occurs as follows. As the average tip-sample distance d is reduced the tip reaches first the attractive tip-sample region leading to phase shift becoming more negative than $-90°$ according to (14.13). This oscillation state corresponds according to Fig. 14.6a to the low amplitude branch. At amplitude A_3, the previously discussed jump from the low-amplitude branch to the high-amplitude branch, i.e. to $A_3^{'}$, occurs. This results in a jump in the phase above $-90°$ (i.e. repulsive interaction) and the phase approaches zero for smaller tip-sample distances.

During the retraction (increasing d), the green line is followed.[4] Thus according to (14.13) the high-amplitude branch with $\phi > -90°$ corresponds to a net repulsive interaction. At $A_2^{'}$ the high amplitude branch is lost (Fig. 14.6c) and a jump to the low amplitude branch (attractive interaction) occurs, followed by a subsequent increase of the phase to $-90°$ during further retraction towards the free oscillation.

In total via the phase we can obtain the assignment that the low-amplitude branch corresponds to $\phi < -90°$ (net attractive tip-sample interaction), while the high-amplitude branch corresponds to $\phi > -90°$ (net repulsive interaction) and the measurement of the phase gives direct information if imaging is performed in the low or the high-amplitude branch.

When measuring the phase or the amplitude-distance dependence $A(d)$, a working point either in the low-amplitude branch (net attractive) or in the high-amplitude branch (net repulsive interaction) can be selected for subsequent imaging. Depending on the material imaged, different interaction regimes may be desired. For a soft

[4] Here we used the dependence $\phi(A/A_{\text{free}})$ while in an experiment the $\phi(d)$ is obtained. However, the two dependences can be converted into each other using the (measured) $A(d)$ dependence.

delicate sample the attractive interaction regime may be desired in order to minimize the tip-sample interaction, while for a hard sample the repulsive regime may be desired in order to penetrate a contamination layer on top of the hard sample.

Due to the bistable nature of the amplitude-distance behavior, the oscillation state may switch from one to the other state. One reason for a change of the oscillation state is a difference in the material properties (cf. Fig. 13.6). When scanning from material C to material B (same height of the atoms), different material dependent force-distance curves can, for instance, trigger a switch from the high-amplitude state (net repulsive interaction, working point 4) on material C to the low-amplitude state (net attractive interaction, working point 2) on material B, or the other way around (cf. Fig. 13.6, curves C and B). The smaller oscillation amplitude leads to a reduction in the average tip-sample distance d by about $\sim 1\,\mathrm{nm}$, which can be mistaken for a topographic step. However, monitoring additionally the phase can help to distinguish a real step in the topography from a border between different materials. In the low-amplitude branch $\phi < -90°$, while in the high-amplitude branch $\phi > -90°$. A purely topographic step (same material) is not associated with a phase change. In this way, a true height change, e.g. due to a step edge (no phase change), can be distinguished from a switch from the high-amplitude oscillation state to the low-amplitude oscillation state due to different materials. This can lead to material contrast which can be observed during scanning.[5]

Up to now we have considered the excitation frequency to be at the free resonance frequency $\omega_{\mathrm{drive}} = \omega_0$. However, in tapping mode the driving frequency is often chosen to be detuned, i.e. not exactly at but slightly above or below the free resonance frequency. The implications of the detuned driving on the amplitude as a function of tip-sample distance are summarized in the following [6]. If in tapping mode the driving frequency is chosen lower than the free resonance frequency, the bistable region is narrower and in most of the working points (amplitude setpoints) the oscillation is stable in the high-amplitude branch (no instabilities) as shown in Fig. 14.7a. This corresponds to a stable operation with the tip being at the lower turnaround point in the repulsive interaction regime and is desirable for hard samples. If the driving frequency is chosen larger than the free resonance frequency, the oscillation remains, down to very low amplitudes on the low-amplitude branch and the bistable region extends almost over the complete range of tip-sample distances as shown in Fig. 14.7c. This can be a disadvantage in terms of possible instabilities. On the other hand, the low-amplitude branch corresponds to an operation in the range of the attractive tip-sample interactions. This can be desirable for imaging soft samples if repulsive tip-sample interactions are to be minimized.

[5] There are also other reasons for the switch between different oscillation sates. For instance, the presence of a valley in the surface topography can enhance the attractive forces (larger regions of the tip will feel the attractive interaction) and thus change the force-distance behavior locally, resulting in a switch to the other branch of the oscillation state.

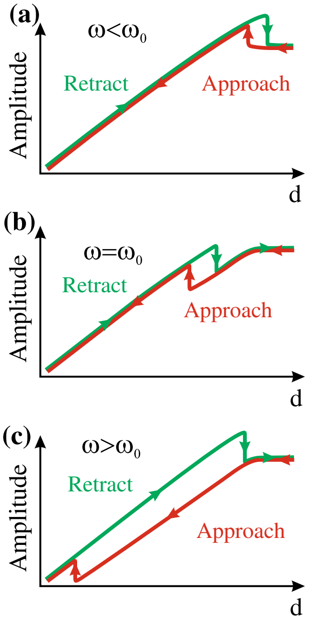


Fig. 14.7 Amplitude as a function of the average tip-sample distance d for driving frequencies **a** below ω_0, **b** at ω_0, and **c** above ω_0. The curves are shown for approach (*red*) and retraction (*green*). As an exercise, the dependencies in **a**–**c** can be deduced from Fig. 14.6a–c This figure is adapted from [6]

14.4 Energy Dissipation in Tapping Mode Atomic Force Microscopy

In our discussion of the tapping mode up to now for simplicity we have considered only conservative tip-sample interactions. When introducing dissipative interactions in dynamic AFM in the small amplitude limit, we subsumed the dissipative part of the tip-sample interactions in one number, the quality factor Q_{ts}, according to (13.14). For the case of large amplitudes used in the tapping mode, the strength of the dissipative tip-sample interaction is different at different distances occurring during one cycle of oscillation. Qualitatively, the dissipative tip-sample interactions should have an appreciable value only close to the lower turnaround point of the oscillation cycle in tapping mode, while the viscous cantilever damping in air is proportional to the velocity, i.e. maximal at the average tip-sample position d_0. Since the conservative and the dissipative part of the tip-sample interaction are a priori unknown, any modeling (e.g. by solving the equation of motion) is difficult from the start. However, no matter how complicated the (conservative and dissipative) interactions are, the law of energy (power) conservation holds.

Therefore, we will now extend our the previous approach, which lead us to (14.5), and use the principle of energy conservation to include also the dissipative tip-sample interaction in the balance of the power. We apply the usual convention that the power entering the system (which is the oscillating cantilever) is positive, while the power leaving the system has a negative sign. Following this convention, the power driving the cantilever P_{drive} is positive, while the power dissipated from the cantilever to the surrounding fluid P_{cant}, as well as the power dissipated to the tip-sample interaction P_{ts} have negative values. In the steady-state, the the sum of these powers averaged over one period vanishes, leading to

$$\langle P_{\text{drive}}\rangle + \langle P_{\text{cant}}\rangle + \langle P_{\text{ts}}\rangle = 0. \tag{14.14}$$

If we would like to avoid negative powers, we can us their absolute values, resulting in the following equation

$$\langle P_{\text{drive}}\rangle = |\langle P_{\text{cant}}\rangle| + |\langle P_{\text{ts}}\rangle|. \tag{14.15}$$

In the following, we analyze this power into and out of the driven cantilever-tip-sample system. No assumptions on the tip-sample force are made, the only assumption made in the following is that the oscillation under the influence of the tip-sample force still remains sinusoidal, which is proven experimentally to be the case [1].

The power pumped into the system by external driving of the cantilever was calculated in (14.2) as[6]

$$\langle P_{\text{drive}}\rangle = -\frac{1}{2}kA_{\text{drive}}A\omega\sin\phi. \tag{14.16}$$

The cantilever damping by the fluid is assumed to be proportional to $\dot{z}$, as $F^{\text{diss}}_{\text{cant}} = -\frac{m\omega_0}{Q_{\text{cant}}}\dot{z}$. Along the same lines as in (14.1), the power dissipated in the cantilever can be calculated as

$$|\langle P_{\text{cant}}\rangle| = \left\langle \frac{m\omega_0}{Q_{\text{cant}}}\dot{z}^2 \right\rangle = \frac{1}{T}\frac{m\omega_0}{Q_{\text{cant}}}\int_0^T A^2\omega^2\sin^2(\omega t + \phi)\mathrm{d}t = \frac{kA^2\omega^2}{2Q_{\text{cant}}\omega_0}. \tag{14.17}$$

Due to (14.15), the power dissipated in the tip-sample interaction can be written as

$$|\langle P_{\text{ts}}\rangle| = \langle P_{\text{drive}}\rangle - |\langle P_{\text{cant}}\rangle| = \frac{kA^2\omega}{2Q_{\text{cant}}}\left(\frac{Q_{\text{cant}}A_{\text{drive}}\sin(-\phi)}{A} - \frac{\omega}{\omega_0}\right). \tag{14.18}$$

This result was obtained using the general law of energy (or power) conservation without any assumptions about the nature of the tip-sample interaction. If the driving frequency ω is chosen at the resonance frequency of the free cantilever ω_0, (14.18) can be written as[7]

[6] Since $0 > \phi > -180°$, $\langle P_{\text{drive}}\rangle$ is positive.

[7] While we used here the principle of energy conservation to derive (14.19), this equation can be obtained alternatively by multiplying (14.10) with $\omega_0 A\sin(\omega_0 t + \phi)$ and integrating over one period, as will be shown in Sect. 14.5.

$$|\langle P_{\mathrm{ts}}\rangle| = \frac{kA^2\omega_0}{2Q_{\mathrm{cant}}}\left(\frac{A_{\mathrm{free}}}{A}\sin(-\phi) - 1\right) , \tag{14.19}$$

with $A_{\mathrm{free}} = Q_{\mathrm{cant}} A_{\mathrm{drive}}$. Correspondingly, the dissipated energy per oscillation period T results as

$$|\langle E_{\mathrm{ts}}\rangle| = |\langle P_{\mathrm{ts}}\rangle| \cdot T = \frac{2\pi E_{\mathrm{osc}}}{Q_{\mathrm{cant}}}\left(\frac{A_{\mathrm{free}}}{A}\sin(-\phi) - 1\right) , \tag{14.20}$$

with $E_{\mathrm{osc}} = 1/2\,kA^2$ being the energy contained in the cantilever oscillation, if ω is close to ω_0. The last term in (14.20) is the power dissipated by the cantilever damping, while the first term in (14.20) is the total dissipated power.

In the case that no dissipative tip-sample interactions are present ($\langle E_{\mathrm{ts}}\rangle = 0$), the simple relation for the phase already obtained in (14.5) results as

$$\sin(-\phi) = \frac{A}{A_{\mathrm{free}}}. \tag{14.21}$$

We can rearrange (14.20) if we remember that $Q_{\mathrm{cant}} = 2\pi E_{\mathrm{osc}}/\langle E_{\mathrm{cant}}\rangle$ and we then obtain the following expression for the phase

$$\sin(-\phi) = \frac{A}{A_{\mathrm{free}}}\left(\frac{\langle E_{\mathrm{ts}}\rangle}{\langle E_{\mathrm{cant}}\rangle} + 1\right). \tag{14.22}$$

The second term in (14.22) is the contribution due to the conservative tip-sample interaction, while the first term includes the contribution due to the dissipative interactions.

In the intermittent contact mode, the amplitude is kept constant by the feedback and thus the phase remains constant during scanning (according to (14.21)) if no dissipative tip-sample interaction is present. A phase change is therefore related to a dissipative tip-sample interaction and maps of the phase recorded as a free signal (not used for feedback) correspond to maps of the dissipative tip-sample interactions. Vice versa: Since A is kept constant by the feedback, a change of the conservative tip-sample interaction does not lead to a phase change.

Now we consider as an approximation that $\langle E_{\mathrm{ts}}\rangle$ is a constant in (14.22), i.e. not dependent on the oscillation amplitude A/A_{free}. This means that at the lower turnaround point always the same energy is dissipated independent of the amplitude. For this case the $\phi(A/A_{\mathrm{free}})$ dependence from (14.22) is displayed in Fig. 14.3 as a dashed curve for $\langle E_{\mathrm{ts}}\rangle \,/\, \langle E_{\mathrm{cant}}\rangle = 0.1$.

Finally, we give a quantitative example of the power dissipated into the tip-sample interaction. All variables in (14.19) are either known or can be measured. In a tapping mode experiment on a silicon wafer in air, a power dissipation $|\langle P_{\mathrm{ts}}\rangle| = 0.3\,\mathrm{pW}$ was obtained independent of the oscillation amplitude [1].

14.5 General Equations for Amplitude and Phase in Dynamic AM Atomic Force Microscopy

In the previous sections we derived equations for the amplitude and the phase in tapping mode, or generally in dynamic amplitude modulation (AM) AFM without the limit of a small oscillation amplitude. We obtained these equations partly from the equation of motion and partly from the principle of energy conservation. In order to simplify the analysis we considered often special cases, like conservative tip-sample interactions, or excitation at the resonance frequency, i.e. $\omega_{\text{drive}} = \omega_0$. Now we will derive from the equation of motion general equations for amplitude and phase without these limits.

We start from the equation of motion (14.6) with F_{ts} not necessarily conservative. We consider a driving oscillation $z_{\text{drive}} = A_{\text{drive}} \cos \omega t$ at a frequency ω which can be different from ω_0, resulting in a cantilever oscillation $z = A \cos(\omega t + \phi)$. Inserting this and the time-derivatives of z into (14.6) results in

$$-m\omega^2 A \cos(\omega t + \phi) = \frac{m\omega_0 \omega A}{Q_{\text{cant}}} \sin(\omega t + \phi) - kA \cos(\omega t + \phi) + k\Delta L + F_{\text{ts}}$$
$$+kA_{\text{drive}} \cos(\omega t) \,. \tag{14.23}$$

If we now multiply this equation by $A \cos(\omega t + \phi)/T$ and integrate over time from zero to T, the following equation results

$$-m\omega^2 A^2 \frac{1}{T} \int_0^T \cos^2(\omega t + \phi) \mathrm{d}t = \frac{m\omega_0 \omega A^2}{Q_{\text{cant}} T} \int_0^T \sin(\omega t + \phi) \cos(\omega t + \phi) \mathrm{d}t$$
$$-kA^2 \frac{1}{T} \int_0^T \cos^2(\omega t + \phi) \mathrm{d}t + k\Delta L \frac{1}{T} \int_0^T \cos(\omega t + \phi) \mathrm{d}t + \frac{1}{T} \int_0^T F_{\text{ts}} z(t) \mathrm{d}t$$
$$+kA_{\text{drive}} A \frac{1}{T} \int_0^T \cos(\omega t) \cos(\omega t + \phi) \mathrm{d}t. \tag{14.24}$$

The first and third integral on the right vanish (due to symmetry) after integration over one period, while the integral over $\cos^2$ over one period results in $T/2$ and the integral over the last term on the right results in $T/2 \cos \phi$. Thus, the above equation simplifies to

$$-\frac{m\omega_0^2 A^2}{2} \frac{\omega^2}{\omega_0^2} = -\frac{1}{2} kA^2 + \frac{1}{T} \int_0^T F_{\text{ts}} z(t) \mathrm{d}t + \frac{1}{2} kAA_{\text{drive}} \cos \phi, \tag{14.25}$$

and finally to

$$\langle F_{\mathrm{ts}}(d + z(t)) \cdot z(t)\rangle = \frac{1}{2}kA^2\left(1 - \frac{\omega^2}{\omega_0^2}\right) - \frac{1}{2}kAA_{\mathrm{drive}}\cos\phi. \tag{14.26}$$

In the limit of $\omega = \omega_0$ this corresponds to (14.13).

A second equation for the amplitude and the phase can be obtained in a similar manor if (14.6) is multiplied by $A\omega \sin(\omega t + \phi)$ and integrated over one period. Due to symmetry the integrals of $\sin x$ and $\sin x \cdot \cos x$ over one period vanish. Moreover, since $\cos(\omega t)\sin(\omega t + \phi)$ integrated over one period results in $T/2 \sin\phi$, the following equation results

$$-\langle F_{\mathrm{ts}}(d + z(t)) \cdot \dot{z}(t)\rangle\, T = |\langle E_{\mathrm{ts}}\rangle| = \pi k A A_{\mathrm{drive}} \sin(-\phi) - \pi k A^2 \frac{\omega}{Q_{\mathrm{cant}}\omega_0}, \tag{14.27}$$

which is equivalent to (14.18).[8] The two independent equations (14.26) and (14.27) can be solved for the amplitude and the phase [8] ($A(\omega, \langle E_{\mathrm{ts}}\rangle, \langle F_{\mathrm{ts}} \cdot z\rangle)$ and $\phi(\omega, \langle E_{\mathrm{ts}}\rangle, \langle F_{\mathrm{ts}} \cdot z\rangle)$).

In the following we consider cases in which the tip-sample force is conservative or dissipative and discuss the consequences on the expressions $\langle F_{\mathrm{ts}} \cdot \dot{z}(t)\rangle$ and $\langle F_{\mathrm{ts}} \cdot z(t)\rangle$ which occur in the equations relating the amplitude and the phase, i.e. (14.26) and (14.27), respectively. In Fig. 14.8a one oscillation cycle of the tip oscillation $z(t)$ is shown with the tip reaching the sample at the lower turnaround point. This curve is symmetric (even) with respect to the time at which the lower turnaround point is reached (dotted vertical line). In Fig. 14.8b an example of a conservative tip-sample force (Lennard–Jones type force) is shown as function of time (red line). For a conservative tip-sample force the force depends only on the tip-sample distance $d + z$. Thus, the tip-sample force is the same for a certain tip-sample distance during approach or retraction. A conservative tip-sample force is even with respect to the time at which the lower turnaround point is reached. The same force is exerted at the same z-positions during approach and retraction of the tip. This also implies that the work done during one cycle of oscillation vanishes.

Now we will show that the expression $\langle F_{\mathrm{ts}}(d + z(t)) \cdot \dot{z}(t)\rangle$ occurring in (14.27) vanishes for a conservative force when averaging is performed over a complete cycle of oscillation. Without loss of generality we can perform the integral

$$\langle F_{\mathrm{ts}}(d + z(t)) \cdot \dot{z}(t)\rangle = \frac{1}{T}\int_0^T F_{\mathrm{ts}}(d + z(t)) \cdot \dot{z}(t)\mathrm{d}t \tag{14.28}$$

over the time-range from $t = 0$ to $t = T$. Since $\dot{z}(t) = -A\omega \sin(\omega t + \phi)$ is odd with respect to the time of the lower turnaround point (green curve in Fig. 14.8a)

[8] $\langle F_{\mathrm{ts}} \cdot \dot{z}(t)\rangle$ is negative, as power is dissipated from the system to the environment.

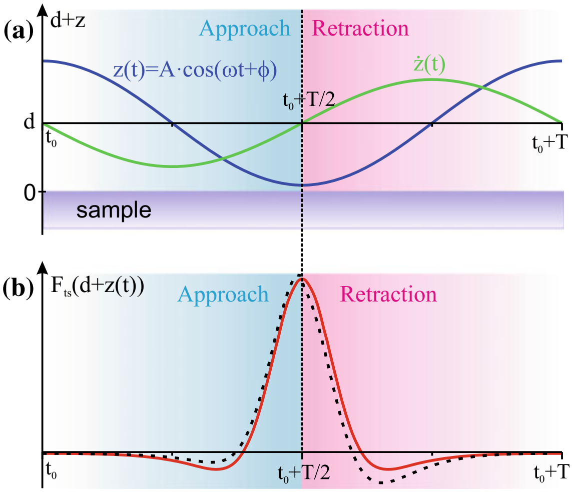


Fig. 14.8 **a** The tip motion $z(t)$ during one oscillation cycle is even with respect to the lower turnaround point of the oscillation, while $\dot{z}(t)$ is odd. **b** A conservative tip-sample force is even with respect to the lower turnaround point (approach and retraction are highlighted by blue and red background colors, respectively), while non-conservative force has different values for approach and retraction at the same tip positions (dashed curve)

while a conservative tip-sample force is even, the product of both is odd and the integral (14.28) vanishes for conservative forces. In this case and for $\omega = \omega_0$ (14.27) simplifies to (14.5).

In the case of a non-conservative tip-sample force the force is no more even with respect to the lower turnaround point, as indicated by the dashed line in Fig. 14.8b. If the total non-conservative tip-sample force can be decomposed into a conservative and a dissipative part as $F_{\text{ts}} = F_{\text{ts}}^{\text{cons}} + F_{\text{ts}}^{\text{diss}}$ the contribution from the conservative force to the integral in (14.28) vanishes and only the contribution due to the dissipative force has to be considered as $\langle F_{\text{ts}} \cdot \dot{z}(t)\rangle = \left\langle F_{\text{ts}}^{\text{diss}} \cdot \dot{z}(t)\right\rangle$ in (14.27).

Now we discuss the expression $\langle F_{\text{ts}} \cdot z(t)\rangle$ occurring in (14.26) with respect to the symmetry of F_{ts}. Any non-conservative force, being different for approach and retraction, can formally be decomposed into an even and an odd contribution[9] as

$$F_{\text{ts}}^{\text{even}}(d+z) = \frac{1}{2}\left(F_{\text{ts}}^{\text{approach}}(d+z) + F_{\text{ts}}^{\text{retract}}(d+z)\right), \tag{14.29}$$

and

$$F_{\text{ts}}^{\text{odd}}(d+z) = \frac{1}{2}\left(F_{\text{ts}}^{\text{approach}}(d+z) - F_{\text{ts}}^{\text{retract}}(d+z)\right), \tag{14.30}$$

[9] The even/odd force contributions with respect to the time $\Delta t > 0$ relative to the time of the lower turnaround point $t_0 + T/2$ are $F_{\text{ts}}^{\text{even/odd}}(d + z(t_0 + T/2 - \Delta t)) = 1/2\,(F_{\text{ts}}(d + z(t_0 + T/2 - \Delta t)) \pm F_{\text{ts}}(d + z(t_0 + T/2 + \Delta t))) = 1/2\left(F_{\text{ts}}^{\text{approach}}(d+z) \pm F_{\text{ts}}^{\text{retract}}(d+z)\right)$. If $\Delta t \to -\Delta t$, $F_{\text{ts}}^{\text{even}} \to F_{\text{ts}}^{\text{even}}$, while $F_{\text{ts}}^{\text{odd}} \to -F_{\text{ts}}^{\text{odd}}$.

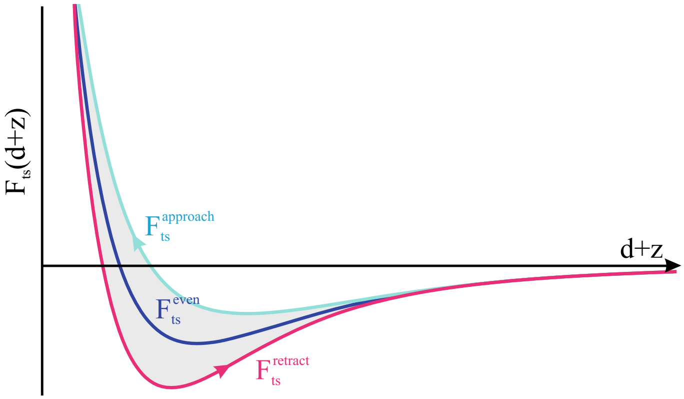


Fig. 14.9 Relation between the forces during approach and retraction and the even force component. The odd force component is half of the difference between the approach and the retraction force

giving rise to

$$F_{\mathrm{ts}}^{\mathrm{approach}}(d+z) = F_{\mathrm{ts}}^{\mathrm{even}}(d+z) + F_{\mathrm{ts}}^{\mathrm{odd}}(d+z), \tag{14.31}$$

and

$$F_{\mathrm{ts}}^{\mathrm{retract}}(d+z) = F_{\mathrm{ts}}^{\mathrm{even}}(d+z) - F_{\mathrm{ts}}^{\mathrm{odd}}(d+z). \tag{14.32}$$

This decomposition is also shown graphically in Fig. 14.9 as function of the tip-sample distance. The force being even with respect to the lower turnaround point is the average of the approach and retraction curves (dark blue line in Fig. 14.9), while the odd contribution is half of the difference between the approach and the retraction curves.

Often the assignment $F_{\mathrm{ts}}^{\mathrm{even}} = F_{\mathrm{ts}}^{\mathrm{cons}}$ and $F_{\mathrm{ts}}^{\mathrm{odd}} = F_{\mathrm{ts}}^{\mathrm{diss}}$ is made. Some caveats about this assignment are discussed in [9]. While a conservative force has to be even and for example a velocity dependent dissipative force is odd, generally a dissipative force is not necessarily odd. If we nevertheless follow the above assignment, the expression $\langle F_{\mathrm{ts}} \cdot z(t)\rangle$ vanishes for an odd force (having opposite sign for approach and retraction), because the tip position $z(t)$ is even with respect to the time of the lower turnaround point. The integral of the odd function $F_{\mathrm{ts}}^{\mathrm{diss}} \cdot z(t)$ over one oscillation cycle vanishes. For an even (conservative) force, both terms in the integral, the tip position $z(t)$ and F_{ts} are even and thus the integral does not vanish and $\langle F_{\mathrm{ts}} \cdot z(t)\rangle = \left\langle F_{\mathrm{ts}}^{\mathrm{cons}} z(t)\right\rangle$ in (14.26).

In total, the contribution to the integral in (14.26) comes only from the even (conservative) force component, while the contribution to the integral in (14.27) comes only from the odd (dissipative) force component (while the integral over the respective other force component vanishes).

14.6 Properties of the Intermittent Contact Mode/Tapping Mode

The intermittent contact mode allows high-resolution topographic imaging even of soft samples. The greatest advantage of the tapping mode is related to the contamination layer present at surfaces under ambient conditions. This thin contamination layer, mostly consisting of water, results in enormous problems when using the non-contact mode. This contamination layer masks the properties of the actual surface under study below the contamination layer. More importantly, if the tip touches this (water) contamination layer, unwanted capillary forces lead to a very strong undesirable force component masking the actual forces from the surface under study. In the case of the tapping mode, the tip passes through this contamination layer and interacts with the actual surface, while in the non-contact mode an unintentional touching of the contamination layer can lead to strong unintended force contributions.

In the contact mode the tip is pressed onto the surface and the contamination layer does not play a significant role. However, here the relatively strong (nN) vertical force leads to strong lateral forces, resulting in wear or sample damage, as the tip scans over the surface. The alternating tapping and motion out of the range of the tip-sample interaction due to the large amplitude in intermittent mode inherently prevents lateral forces causing damage (wear) of tip or sample during scanning. Due to the very short contact to the surface, the surface material is not pulled sideways by shear forces since the applied force is always vertical. The large oscillation amplitudes also allow to use relatively soft cantilevers and nevertheless avoiding snap-to-contact. This shows that the tapping mode has several important advantages over the other modes. The tapping mode thus exploits the advantages of contact mode and non-contact mode and it avoids their disadvantages. The intermittent contact mode has several advantages when imaging a surface, however, a disadvantage is that it gives no easy access to quantities describing the tip-sample interaction like the force or the force gradient, since these quantities are averaged in a non linear manner over the oscillation cycle.

Tapping mode imaging is implemented in ambient air by oscillating the cantilever at or very near the cantilever resonance frequency at typical oscillation frequencies between 50 and 500 kHz. Amplitudes in the range of 10–100 nm are used in this mode, when the tip is not in contact with the surface (free amplitude). Force constants in the range between 10–50 N/m are usually used. The oscillation amplitude of the cantilever tip is measured by a corresponding amplifier and fed to the input of the controller electronics. The feedback loop then adjusts the tip-sample separation to maintain a constant setpoint amplitude for instance 80–90 % of the free amplitude. In order to stabilize the oscillation in the net repulsive interaction regime (high-amplitude branch), the driving frequency is often chosen below (usually about 5 % below) the resonance frequency of the free cantilever, i.e. $\omega < \omega_0$. It is also found that larger oscillation amplitudes A tend to stabilize the high-amplitude branch (repulsive interaction regime), while smaller amplitudes tend to stabilize the low-amplitude branch for usual values of $A/A_{\text{free}} \approx 0.5 - 0.9$.

As we have already seen, the amplitude has a monotonous dependence on the tip-sample distance (Fig. 14.2b). This leads to a clear unambiguous feedback signal, being another very important advantage of the tapping mode. This is different from the frequency shift used as the feedback signal, where the non-monotonous dependence on the tip-sample distance can lead to serious instabilities as discussed in Sect. 16.3.

14.7 Summary

- The intermittent contact mode (tapping mode) is a detection mode which differs from the AM mode in the following ways: (a) The oscillation amplitudes are large (typically 50 nm), reaching deep into the repulsive regime and correspondingly the tip-sample force has a non-linear distance dependence. (b) The driving frequency is at or very close to the free resonance frequency ω_0.
- The oscillation amplitude decreases linearly with decreasing average tip-sample distance d, giving rise to a stable feedback signal without the danger of serious instabilities. This amplitude reduction also occurs without any dissipative tip-sample interaction due to a less efficient energy transfer off-resonance. The resonance condition $\phi = -90°$ applying for the case of the free cantilever is left due to the tip-sample interaction.
- An anharmonic tip-sample force leads to the coexistence of two vibrational modes with a low-amplitude and a high-amplitude (separated by about 1 nm), corresponding to a net attractive and net repulsive interaction, respectively. Transitions between these modes occur at particular tip-sample distances, or when scanning from one material to another. These modes can be distinguished by the phase, $\phi < -90°$ for the low-amplitude (attractive) mode and $\phi > -90°$ for the high-amplitude (repulsive) mode.
- The dissipative tip-sample interaction energy can be calculated via the energy conservation. The power dissipated into the tip-sample interaction can be determined by measuring the oscillation amplitude and the phase.
- Maps of the phase signal in the intermittent mode of atomic force microscopy correspond to maps of tip-sample dissipation.
- In contrast to the contact mode, in the tapping mode no sidewise frictional forces are exerted on the sample minimizing the wear on delicate samples.

References

1. J.P. Cleveland, B. Anczykowski, A.E. Schmid, V.B. Elings, Energy dissipation in tapping-mode atomic force microscopy. Appl. Phys. Lett. **72**, 2613 (1998). https://doi.org/10.1063/1.121434
2. M.V. Salapaka, D.J. Chen, J.P. Cleveland, Linearity of amplitude and phase in tapping-mode atomic force microscopy. Phys. Rev. B **61**, 1106 (2000). https://doi.org/10.1103/PhysRevB.61.1106

3. A.I. Livshits, A.L. Shluger, A.L. Rohl, Contrast mechanism in non-contact SFM imaging of ionic surfaces. Appl. Surf. Sci. **140**, 327 (1999). https://doi.org/10.1016/S0169-4332(98)00549-2
4. U. Dürig, Relations between interaction force and frequency shift in large-amplitude dynamic force microscopy. Appl. Phys. Lett. **75**, 433 (1999). https://doi.org/10.1063/1.124399
5. L.D. Landau, E.M. Lifshitz, *Mechanics*, vol. 1, (Butterworth-Heinemann, Oxford 1976) ISBN 978-0-7506-2896-9
6. H. Hölscher, U.D. Schwarz, Theory of amplitude modulation atomic force microscopy with and without Q-control. Non-linear Mech. **42**, 608 (2007). https://doi.org/10.1016/j.ijnonlinmec.2007.01.018
7. L. Zitzler, S. Herminghaus, F. Mugele, Capillary forces in tapping mode atomic force microscopy. Phys. Rev. B **66**, 155436 (2002). https://doi.org/10.1103/PhysRevB.66.155436
8. J.R. Lozano, R. Garcia, Theory of phase spectroscopy in bimodal atomic force microscopy. Phys. Rev. B **79**, 014110 (2009). https://doi.org/10.1103/PhysRevB.79.014110
9. J.E. Sader, T. Uchihashi, M.J. Higgins, A. Farrell, Y. Nakayama, S.P. Jarvis, Quantitative force measurements using frequency modulation atomic force microscopy-theoretical foundations. Nanotechnology **16**, S94 (2005). https://doi.org/10.1088/0957-4484/16/3/018

Chapter 15
Mapping of Mechanical Properties Using Force-Distance Curves

The imaging modes considered in the previous chapters resulted mainly in topographic imaging. Contours of constant force in the static mode, or constant oscillation amplitude in the dynamic AM modes are measured. In Chap. 10 we have seen that force-distance curves give important information on the mechanical properties of the sample, like elasticity of the sample, adhesion properties and dissipation. The concept behind mapping of mechanical properties by force-distance curves is to acquire a force-distance curve at each image point and to extract images of elasticity, adhesion and other mechanical properties.

In the dynamic modes, the information about the tip-sample interaction is always averaged over the oscillation cycle, which complicates the extraction of information on the tip-sample interaction. Invoking force-distance curves during scanning gives more direct access to the mechanical properties. This method using force-distance curves for the mapping of mechanical properties of the sample has different names: peak force tapping [1, 2], force volume or pulsed force mode [3, 4]. Besides access to the mechanical properties, this mode also allows high-resolution imaging, it is a tapping mode under additional force control, while the ordinary tapping mode controls the amplitude whereas the tip-sample force remains unknown.

15.1 Principles of Force-Distance Curve Mapping

When measuring maps of force-distance curves, these curves are not acquired with a frequency close to the resonance frequency of the cantilever, but at a much lower frequency of several thousand Hz. Force-distance curves are acquired in the quasi-static mode i.e. measuring the force by the (quasi-static) bending of the cantilever. If several thousand force-distance curves are acquired per second, a force-distance curve can be acquired at each image point, while still maintaining a reasonable acquisition time of a few minutes for an image.

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3_15

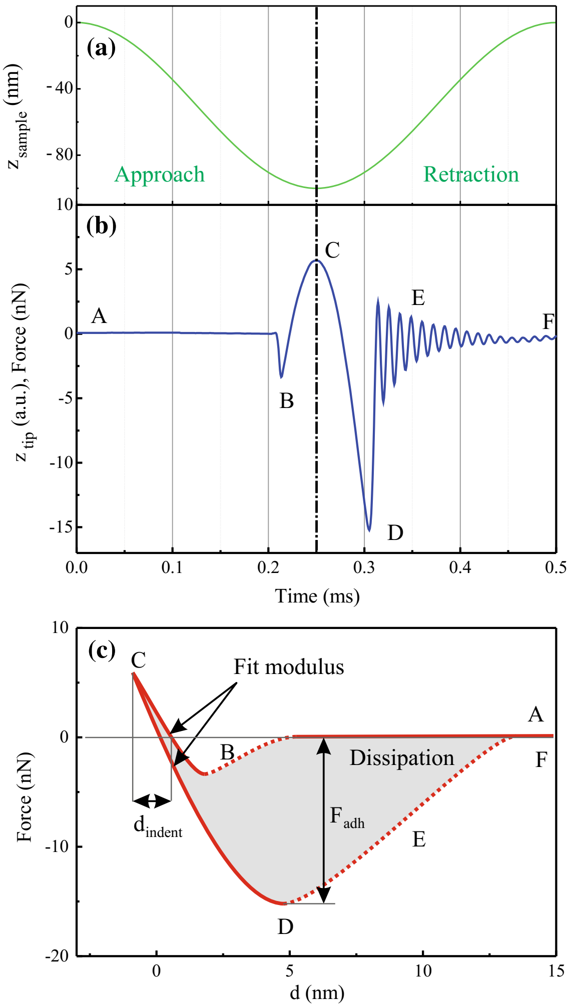


Fig. 15.1 **a** Sinusoidal change of the z-position of the sample during the acquisition of the force-distance curve. **b** Cantilever deflection z_{tip} (proportional to the tip-sample force) as a function of the time. **c** Tip-sample force as function of the tip-sample distance. From this curve, quantities like the adhesion force F_{adh}, the indentation depth d_{indent}, or the dissipation energy can be retrieved and maps (images) of these quantities can be acquired. The dissipation corresponds to the *shaded area* between the approach and the retraction curve. The Young's modulus of the sample can be determined by fitting a model for the mechanic contact to the approach force-distance curve

The force-distance curves considered in Chaps. 10 and 12 were taken only at one point on the sample within a acquisition time of typically a second. In force-distance curve mapping, the curves are typically acquired in less than a millisecond. In order to prevent cantilever excitations at higher harmonics the linear change of the z-position with sharp edges at the turnaround points is replaced by a sinusoidal excitation. The z-position of the sample is changed (modulated) at a frequency of several kHz, as shown in Fig. 15.1a. The larger z-values correspond to a large tip-sample distance with negligible tip-sample force, while at the lower z-values the tip comes into contact with the sample.

In Fig. 15.1b the corresponding cantilever deflection is shown, which is proportional to the tip-sample force. When the tip comes closer to the sample from region

A until close to point B, the attractive force increases slightly. At point B snap-to-contact occurs. The repulsive force increases towards point C. The maximum (peak) force is reached at point C. This peak force is of central importance and is used also as the signal for the z-feedback. During retraction of the tip the repulsive force turns into an attractive adhesive force. At point D the maximum attractive force is reached and snap-out-of-contact occurs. After snap out of contact the tip-sample force is negligible and a cantilever ring-down of the free cantilever is observed with an oscillation at its resonance frequency. The time constant of this exponential ring-down is given by the damping of the cantilever (region E). Thus, in region E it is not the tip-sample force which is shown, but the cantilever bending during ring-down. This "false" force signal due to the cantilever ring-down is undesired and has to be distinguished from other features of interest in the force-distance curve during the analysis of the curve. In region F, the tip has reached its quasi-free equilibrium position, and it is moved to the next lateral position (next image pixel) and the next force-distance curve will be acquired.

The force as a function of time can be converted into a curve of the force as a function of the z-position of the sample z_{sample}, which is considered to oscillate. Further, taking also the measured cantilever bending resulting from the force measurement into account, the dependence of the tip-sample force can be obtained as a function of the tip-sample distance $d = z_{\text{tip}} - z_{\text{sample}}$, which is shown schematically in Fig. 15.1c (cf. Fig. 11.6). From region A to B, a very small attractive force is measured during approach. At the snap-to-contact, the tip-sample distance decreases abruptly and the attractive force becomes abruptly more negative (dashed line in region B). Approaching more closely, the tip-sample force becomes repulsive and reaches the peak force (region C). The zero point for the tip-sample distance d is chosen at the point where the force is zero. At this point, the repulsive force at the tip apex is balanced by the attractive force from a larger volume of the tip. Negative values of d correspond to an indentation of the tip into the sample. Upon tip retraction from the surface, the force will be the same as for the approach for conservative interactions (such as an elastic force). If there is some dissipative tip-sample interaction (such as plastic deformation) the force during retraction will lie below the force curve for the approach. The larger attractive (more negative) force during retraction can be explained due to adhesion. At point D snap-out-of-contact occurs; here the tip-sample distance d increases abruptly and the tip-sample force drops to negligible values (dashed line in region E). In region F, the free cantilever state is reached before the next force curve is acquired.

The measured peak force is used for the z-feedback, i.e. the measured peak force is compared to a peak force setpoint and a feedback controller determines the appropriate z-signal needed in order to keep the measured peak force close to the setpoint. This feedback on the peak force has an advantage compared to the intermittent contact (tapping) mode. In tapping mode, the amplitude is kept constant, not the force. It is an advantage if the force is controlled, since a high peak force can induce undesired damage of the sample surface or the tip. Thus, controlling the force to a sufficiently small peak force is the best way to prevent unwanted sample and tip modifications. Since in tapping mode the amplitude and not the (peak) force is controlled, undesirable large forces may occur during scanning. Controlling the peak force is a gentle way of tapping, minimizing undesirably strong tip-sample interactions. Therefore, the peak force tapping mode is not only useful for mapping mechanical properties, but also for high-resolution imaging.

15.2 Mapping of the Mechanical Properties of the Sample

In the following, it will be shown how the peak force tapping mode can be used to determine the mechanical properties of the sample. For instance, the adhesion force F_{adh} and the indentation depth d_{indent} can be determined from each force-distance curve, as indicated in Fig. 15.1c. These quantities can be represented as images of (maximum) adhesion force or indentation depth at the peak force.

The dissipation energy can be obtained as the area between the approach and the retraction curves, as

$$E_{\mathrm{diss}} = \int_{z_{\mathrm{min}}}^{z_{\mathrm{max}}} \left(F_{\mathrm{approach}} - F_{\mathrm{retract}}\right) \mathrm{d}z, \tag{15.1}$$

with F_{approach} and F_{retract} being the forces during approach and retraction, respectively. The dissipation energy can be represented by the shaded area in Fig. 15.1c. The dissipation in the attractive regime (negative forces, which corresponds to dissipation due to adhesion) can even be distinguished from the dissipation in the repulsive regime, and those quantities can be mapped separately.

Another quantity of interest which can be mapped is the slope of the force-distance curve in the repulsive regime, which is related to the stiffness of the sample. More quantitatively, the force-distance curves can be fitted to an appropriate model of the tip-sample contact, for instance the Hertz model of the elastic contact, or models also including contributions from attractive forces, as the DMR, JKR, and MD models introduced in Sect. 10.2. In principle, Young's modulus can be obtained from a fit of the model to the measured force-distance curve. However, several parameters enter into the model which are often not known (precisely): the tip radius, the Young's modulus of the tip, and the Poisson ratios of the tip and the sample. If these parameters are known or estimated, the Young's modulus of the sample can be determined. Often it is not necessary to determine the absolute value of Young's modulus, but to detect differences if different materials are present at different areas of the sample.

The parameters characterizing the sample properties can be extracted "online" during scanning from the acquired force-distance curve using fast data processing. In this case, only the maps of the resulting parameters are stored as data and the individual force-distance curve is not stored. The challenge in this analysis is then to distinguish the desired points of the force-distance curve (such as peak force and maximum adhesive force) from undesirable features like the cantilever ring-down. In some cases the maximum due to cantilever ring-down may become the global maximum of the curve, while the peak force is only a local maximum. The curve analysis algorithm has to reliably identify the desired information. This is specifically important for the peak force, since this is used for the feedback and any false determination of the peak force will corrupt the feedback and can lead to a tip-sample crash. As an alternative to the "online" analysis each force-distance curve for each image point can also be stored and analyzed later ("off-line"). Of course this means there is a large amount of data to be stored.

This approach to detect force data as a function of the tip-sample distance can also be generalized to quantities other than the force. For instance, the phase can be acquired as a function of x, y, and z. This approach generates a data volume which has to be analyzed properly in order to extract useful information.

15.3 Summary

- In the peak force tapping mode thousands of force-distance curves are measured per second, one at each image point. The z-feedback for topographic imaging uses the maximal (peak) force as the signal. This force control allows sample and tip damage to be minimized.
- Parameters characterizing the mechanical properties of the sample are extracted from the force-distance curves. Corresponding maps of adhesion, indentation, dissipation, stiffness and other parameters are obtained.

References

1. O. Sahin, N. Erina, High-resolution and large dynamic range nanomechanical mapping in tapping-mode atomic force microscopy. Nanotechnology **19**, 445717 (2008). https://doi.org/10.1088/0957-4484/19/44/445717
2. P. Trtik, J. Kaufmann, U. Volz, On the use of peak-force tapping atomic force microscopy for quantification of the local elastic modulus in hardened cement paste. Cem. Concr. Res. **42**, 215 (2012). https://doi.org/10.1016/j.cemconres.2011.08.009
3. A. Rosa-Zeiser, E. Weilandt, S. Hild, O. Marti, The simultaneous measurement of elastic, electrostatic and adhesive properties by scanning force microscopy: pulsed-force mode operation. Meas. Sci. Technol. **8**, 1333 (1997). https://doi.org/10.1088/0957-0233/8/11/020
4. H. Krotil, T. Stifter, H. Waschipky, K. Weishaupt, S. Hild, O. Marti, Pulsed force mode: a new method for the investigation of surface properties. Surf. Interface Anal. **27**, 336 (1999). https://doi.org/10.1002/(SICI)1096-9918(199905/06)27:5/6<336::AID-SIA512>3.0.CO;2-0

Chapter 16
Frequency Modulation (FM) Mode in Dynamic Atomic Force Microscopy—Non-contact Atomic Force Microscopy

In Chap. 14 we introduced the intermittent contact mode (tapping mode), which is a very successful operation mode in dynamic atomic force microscopy. Since this mode has so many advantages, why should we use any other mode? In this chapter we introduce the FM detection scheme (often named non-contact atomic force microscopy) which in some cases has the following advantages over the tapping mode: (a) The FM detection scheme can be used with high Q cantilevers ($Q > 1{,}000$, occurring in vacuum). For high Q cantilevers the tapping mode results in unacceptably long measurement times. (b) The inelastic dissipation in the tip-sample interaction can be easily measured during scanning. (c) From the measured data the tip-sample force can be reconstructed as a function of the distance. (d) True non-contact atomic resolution imaging can be performed (in vacuum without the contamination layer) avoiding any repulsive tip-sample force and thus also avoiding wear.

In the FM detection scheme of AFM the cantilever does not oscillate at a fixed driving frequency (as in the tapping mode), but always oscillates at resonance [1–4]. If the resonance frequency shifts due to a tip-sample interaction, the cantilever oscillation frequency follows this shift. In the FM mode, the amplitudes are often so large that the tip-sample force cannot be approximated as linear. In spite of the nonlinear tip-sample force the resulting frequency shift can be calculated. The frequency shift in the FM mode is proportional to a weighted average of the tip-sample force over a cantilever oscillation cycle. For large amplitudes, the frequency shift depends almost exclusively on the tip-sample interaction at the lower turnaround point. We will describe in detail the experimental setup and the different FM detection modes and compare the FM and AM detection modes.

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3_16

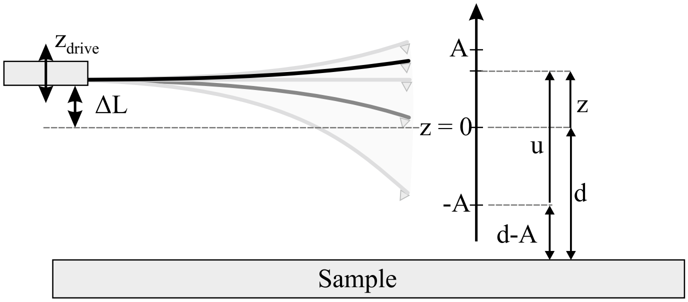


Fig. 16.1 Scheme of the cantilever vibration illustrating the corresponding coordinates

16.1 Principles of FM Detection in Dynamic Atomic Force Microscopy

In the following, we will consider again a driven damped harmonic oscillator under the influence of a conservative non-linear tip-sample force $F_{\text{ts}}(d+z)$, however, now for the case that the oscillation frequency is always at resonance. The driving force is given by an external sinusoidal oscillation $z_{\text{drive}} = A_{\text{drive}} \cos(\omega t)$ of the cantilever base. In FM detection, the driving at ω_{drive} is always applied at the actual resonance frequency[1] ω_0', which we call ω in the following, i.e. $\omega = \omega_{\text{drive}} = \omega_0'$. How it is experimentally achieved that the cantilever oscillates always at the (shifting) resonance frequency will be explained in the next section. In the following we just assume that the cantilever is always driven and oscillates always at its (shifted) resonance frequency. The equation of motion for the driven damped harmonic oscillator with an external tip-sample force $F_{\text{ts}}(d+z)$ added is written according to (13.4) as

$$m\ddot{z} + \frac{m\omega_0}{Q_{\text{cant}}}\dot{z} = -k(z - z_{\text{drive}} - \Delta L) + F_{\text{ts}}(d+z). \tag{16.1}$$

The relevant coordinates are indicated in Fig. 16.1. The zero point for z ($z = 0$) is given by the condition that the tip-sample force at $z = 0$ is compensated by the static cantilever bending ΔL, cf. Fig. 13.1 and (13.1). In this case the tip-sample distance is d.

In spite of the fact that a non-linear tip-sample force $F_{\text{ts}}(d+z)$ is included into the equation of motion, the motion of the tip z(t) is in a very good approximation a sinusoidal oscillation $z(t) = A\cos(\omega t + \phi)$, as known from experimental results and simulations [5, 6]. Since the oscillation in FM mode is always at resonance, $\phi = -90°$ and thus $z(t) = A\sin(\omega t)$. We will not solve the equation of motion (16.1), nevertheless, we will calculate the shift of the resonance frequency. The relation between tip-sample force and frequency shift $\Delta\omega$ will turn out to be more

[1] Under the influence of the tip-sample force the resonance frequency of the cantilever shifts from the resonance frequency of the free cantilever, ω_0, to ω_0'.

complicated than the simple proportional relation between $\Delta\omega$ and the force gradient obtained in the small amplitude limit (13.10). For the case of the non-linear tip-sample force, the final result will be that the frequency shift corresponds to a properly weighted average of the tip-sample force over an oscillation period.

An expression for the frequency shift can be derived if we insert the explicit expressions for the harmonic oscillation of the cantilever $z(t)$ and its derivatives as well as the expression for z_{drive} into (16.1). Subsequently we multiply (16.1) by $z(t) = A \sin \omega t$ and integrate over one period resulting in the following expression

$$\begin{aligned}
&-\int_0^T m\omega^2 A^2 \sin^2 \omega t \, \mathrm{d}t + \int_0^T \frac{m\omega_0}{Q_{\text{cant}}} A^2 \omega \cos \omega t \sin \omega t \, \mathrm{d}t + \int_0^T kA^2 \sin^2 \omega t \, \mathrm{d}t \\
&-\int_0^T kA_{\text{drive}} A \cos \omega t \sin \omega t \, \mathrm{d}t - \int_0^T k\Delta L A \sin \omega t \, \mathrm{d}t \\
&\quad = \int_0^T F_{\text{ts}}(d + z(t)) A \sin \omega t \, \mathrm{d}t. \qquad (16.2)
\end{aligned}$$

Since the integral of $\cos \omega t \sin \omega t$ over one period vanishes, the second and fourth terms on the left side in (16.2) vanish. The last term on the left side vanishes as well, since it is proportional to an integral of $\sin \omega t$ over one period. Thus, (16.2) can be written as

$$(k - m\omega^2)A^2 \int_0^T \sin^2 \omega t \, \mathrm{d}t = \int_0^T F_{\text{ts}}(d + z(t)) A \sin \omega t \, \mathrm{d}t. \qquad (16.3)$$

The integral $\int \sin^2 \omega t \, \mathrm{d}t$ within the limits from 0 to T can be calculated as $\frac{1}{2}T = \frac{\pi}{\omega}$, which results in

$$(k - m\omega^2)A^2 \frac{\pi}{\omega} = \int_0^T F_{\text{ts}}(d + z(t)) A \sin \omega t \, \mathrm{d}t. \qquad (16.4)$$

The left hand side of (16.4) can be further evaluated as follows

$$\begin{aligned}
\frac{A^2\pi}{\omega}\left(k - m\omega^2\right) &= \frac{A^2 m\pi}{\omega}\left(\frac{k}{m} - \omega^2\right) \\
&= \frac{A^2 m\pi}{\omega}\left(\omega_0^2 - \omega^2\right) = \frac{A^2 m\pi}{\omega}\left(\omega_0 + \omega\right)\left(\omega_0 - \omega\right). \qquad (16.5)
\end{aligned}$$

Since the tip-sample force is considered as a small perturbation, the frequency shift will be small as well, i.e. $\omega \approx \omega_0$ and $(\omega_0 + \omega) \approx 2\omega$. Thus, the left-hand side of (16.4) can be further written as

$$2\pi m A^2(\omega_0 - \omega) = -2\pi m A^2 \Delta\omega = -4\pi^2 m A^2 \Delta f. \tag{16.6}$$

Now also taking the right-hand side of (16.4) into account the following expression for the frequency shift arises

$$\Delta f = -\frac{1}{4\pi^2 m A^2} \int_0^T F_{\mathrm{ts}}(d + z(t)) A \sin \omega t \, \mathrm{d}t. \tag{16.7}$$

The time average of $F_{\mathrm{ts}}(t)$ times $z(t)$ over one period can be written as

$$\langle F_{\mathrm{ts}}(t) \cdot z(t) \rangle \equiv \frac{1}{T} \int_0^T F_{\mathrm{ts}}(d + z(t)) A \sin \omega t \, \mathrm{d}t. \tag{16.8}$$

Using the above equation, (16.7) can be rewritten as the following expression for Δf (using $T = 1/f_0$ and $m = k/\omega_0^2$)

$$\Delta f = -\frac{f_0}{A^2 k} \langle F_{\mathrm{ts}}(t) \cdot z(t) \rangle. \tag{16.9}$$

The frequency shift is proportional to $\langle F \cdot z \rangle$, which is the time average of force times distance over one oscillation period. The dependence as f_0/k on the resonance frequency and the spring constant is the same as in the small amplitude limit (13.11). In contrast to the case of small amplitudes, the frequency shift depends as $1/A^2$ on the oscillation amplitude. If the force is split into an even and an odd contribution (see Sect. 14.5), the contribution of an odd (dissipative) force $\langle F_{\mathrm{ts}}^{\mathrm{odd}}(t) \cdot z(t) \rangle$ vanishes and thus only the contribution due to an even (conservative) force contributes as $\langle F_{\mathrm{ts}}(t) \cdot z(t) \rangle = \langle F_{\mathrm{ts}}^{\mathrm{even}}(t) \cdot z(t) \rangle$.

As a consistency check we insert the force for a harmonic oscillator $F_{\mathrm{ts}} = -k'z$ as an approximation in the case of the small amplitude limit. This results in

$$\langle F_{\mathrm{ts}} \cdot z \rangle = -\langle k' \cdot z^2 \rangle = \frac{1}{T} \int_0^T -k' A^2 \sin^2 \omega t \, \mathrm{d}t = -\frac{1}{2} k' A^2, \tag{16.10}$$

which recovers the result of the frequency change found for the small amplitude limit $\Delta f = f_0 k'/(2k)$ (cf. (13.11)). In analogy to this result for the small amplitude limit an effective tip-sample spring constant can generally be defined as

$$k' \equiv -\frac{2\langle F_{\mathrm{ts}} \cdot z \rangle}{A^2}, \tag{16.11}$$

in order to recover from (16.9) an equation of the same form as in the small amplitude limit $\Delta f = f_0 k'/(2k)$.

16.1.1 Expression for the Frequency Shift

When analyzing the time average in (16.8) qualitatively, it can be seen that the parts of the oscillation path which make the largest contribution to the frequency change are the turnaround points. Here the velocity is lowest, so the tip stays longest at these positions (strongest contribution to the integral over time). The equilibrium position is passed quickly at the largest velocity, leading to a small contribution to the time average. This dominant contribution of the turnaround points can be obtained more quantitatively if we replace the time average in (16.8) by a spatial average. A spatial average over the positions of the tip in one oscillation cycle is also more appropriate because the tip-sample force is primarily a function of tip-sample distance. For the average $\langle F \cdot z \rangle$ we wrote in (16.8)

$$\langle F_{\mathrm{ts}}(d + z(t)) \cdot z(t) \rangle = \frac{1}{T} \int_0^T F_{\mathrm{ts}}(d + z(t)) \cdot z(t) \,\mathrm{d}t, \tag{16.12}$$

with $z(t) = A \sin \omega t$. In order to convert the time average to a spatial average over the trajectory, we substitute in (16.12) the variable t by z as

$$\frac{\mathrm{d}z}{\mathrm{d}t} = A\omega \cos \omega t = A\omega \sqrt{1 - \sin^2 \omega t} = \omega \sqrt{A^2 - z^2}. \tag{16.13}$$

Due to the square root this substitution is only valid for positive values of $\cos \omega t$ and we split the integral over the whole oscillation period in twice the integral over the halve period from the lower turnaround point to the upper one (as the tip-sample force is even with respect to the lower turnaround point). With the above substitution, the average $\langle F \cdot z \rangle$ can be written as

$$\begin{aligned} \langle F_{\mathrm{ts}}(d + z) \cdot z \rangle &= \frac{1}{T} \int_0^T F_{\mathrm{ts}}(d + z(t)) \cdot z(t) \,\mathrm{d}t \\ &= \frac{2}{\omega T} \int_{-A}^{+A} \frac{F_{\mathrm{ts}}(d + z) \cdot z}{\sqrt{A^2 - z^2}} \mathrm{d}z \\ &= \frac{1}{\pi} \int_{-A}^{+A} \frac{F_{\mathrm{ts}}(d + z) \cdot z}{\sqrt{A^2 - z^2}} \mathrm{d}z. \end{aligned} \tag{16.14}$$

Combining (16.9) and (16.14) the following expression for the frequency shift is obtained

$$\Delta f = -\frac{f_0}{\pi k A^2} \int\limits_{-A}^{+A} F_{\mathrm{ts}}(d+z)\frac{z}{\sqrt{A^2-z^2}}\mathrm{d}z = \frac{f_0}{\pi k A^2} \int\limits_{-A}^{+A} F_{\mathrm{ts}}(d+z)g(z)\mathrm{d}z. \quad (16.15)$$

This can be interpreted as the integral of the tip-sample force from $-A$ to A with a weighting function $g(z)$. Due to this weighting function, the largest contributions to the frequency shift come from the regions close to the turnaround points of the oscillation $z = \pm A$. Here the weighting function diverges (denominator becomes zero) as seen in Fig. 16.2a. From the weighting function alone a large contribution to the frequency shift is expected at both turnaround points. However, the second factor in the integrand of (16.15), the tip-sample force F_{ts}, must also be considered. For the situation of a large amplitude shown in Fig. 16.2a the contribution to the frequency shift at the upper turnaround point $z = A$ is eliminated by the vanishing tip-sample force F_{ts}. The product of weighting function and tip-sample force, i.e. the integrand of (16.15) is shown as a green line in Fig. 16.2a. In total, for large amplitudes the contributions to the frequency shift come only from regions close to the lower turnaround point, while the major part of the oscillation path does not result in a contribution to the frequency shift.

The case of a smaller oscillation amplitude is shown in Fig. 16.2b. For better comparability, the lower turnaround point of the oscillation was placed in the same position as in Fig. 16.2a. In this case, the integrand of (16.15) provides contributions to all parts of the oscillation cycle, since the force has appreciable values throughout the oscillation. The largest contributions to the frequency shift arise from both turnaround points, as shown by the green line in Fig. 16.2b.

This means that for smaller oscillation amplitudes a stronger frequency shift signal is expected. In addition to this contribution from the integral in (16.15) also the prefactor $1/A^2$ enhances the frequency shift for small amplitudes. If we compare this amplitude dependence of the frequency shift in the previously treated small amplitude limit (13.11), we note that in this case the frequency shift was found to be independent of the oscillation amplitude. The strength of the signal is one issue, another is the corresponding noise, which also increases with decreasing amplitude, as will be discussed in Chap. 17. Together, the important figure of merit, the signal-to-noise ratio, will be obtained (see Sect. 17.9).

Due to the antisymmetric behavior of the weighting function with respect to the point of origin of the oscillation, a constant force will not lead to a frequency shift. This corresponds to the result also obtained in the small amplitude limit that a constant force induces no frequency shift.

Often the total tip-sample force is considered as a superposition of different force contributions. Since the force enters linearly in the integral (16.15) the total frequency shift can be split into contributions arising from the individual forces.

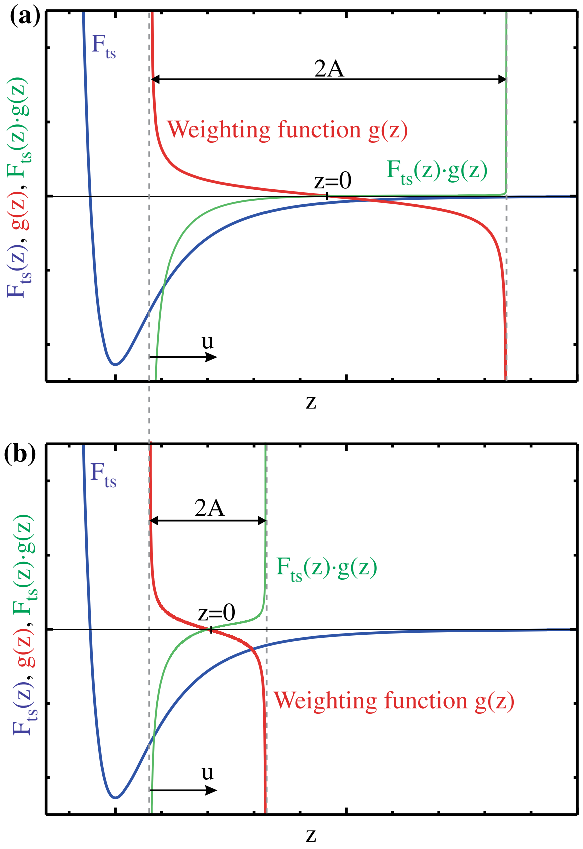


Fig. 16.2 The tip-sample force (blue), the weighting function $g(z)$ (red), and their product (green) are displayed as a function of distance z for two different oscillation amplitudes A. In the large amplitude limit **a** the frequency shift signal is mainly picked up close to the lower turnaround point of the oscillation, while in the smaller amplitude case **b** contributions to the frequency shift are picked up during the whole oscillation cycle with the main contributions coming from both turnaround points. For better comparison, the lower turnaround point is kept constant in (**a**) and (**b**)

The expression for the frequency shift (16.15) can be further evaluated with integration by parts. This results in

$$\Delta f = -\frac{f_0}{\pi k A^2}\left(-F_{\text{ts}}(d+z)\sqrt{A^2-z^2}\Big|_{-A}^{+A} + \int_{-A}^{+A} \frac{\partial F_{\text{ts}}(d+z)}{\partial z}\sqrt{A^2-z^2}\,\mathrm{d}z \right). \tag{16.16}$$

As the first term in (16.16) vanishes, the following expression for the frequency shift is obtained

$$\Delta f = -\frac{f_0}{2k}\int_{-A}^{+A} \frac{\partial F_{\text{ts}}(d+z)}{\partial z}\frac{\sqrt{A^2-z^2}}{1/2\pi A^2}\mathrm{d}z. \tag{16.17}$$

This corresponds to a weighted average of the tip-sample force-gradient, where the weighting function is a semicircle with a radius A divided by the area of the semicircle. This expression for the frequency shift is very similar to expression for the small amplitude limit (13.11), only instead of the tip-sample force gradient at the position $z = 0$ a weighted average of the force gradient over the oscillation path enters.

16.1.2 Normalized Frequency Shift in the Large Amplitude Limit

Up to now the coordinates have been chosen such that the reference for the position of the cantilever tip z was the equilibrium position of the cantilever (Fig. 16.1). This is the position in which the tip-sample force is compensated by the static bending force of the cantilever, also called the average tip position. Often, the lower turnaround point of the oscillation is a more useful reference point. Therefore, we now choose as a new distance variable $u = z + A$ in order to describe the tip position relative to the lower turnaround point (Fig. 16.1). If we substitute $z = u - A$ and express the tip-sample distance as $d + z = d - A + u$ the frequency shift (16.15) results in

$$\begin{aligned}\Delta f &= -\frac{f_0}{\pi k A^2}\int\limits_0^{2A}\frac{F_{\mathrm{ts}}(d-A+u)(u-A)}{\sqrt{A^2-(u-A)^2}}\mathrm{d}u\\ &= -\frac{f_0}{\pi k A^2}\int\limits_0^{2A}\frac{F_{\mathrm{ts}}(d-A+u)(u-A)}{\sqrt{(2A-u)u}}\mathrm{d}u.\end{aligned} \tag{16.18}$$

In the following, we consider the limit of a large oscillation amplitude, i.e. the oscillation amplitude A is much larger than the range of the tip-sample force. In this case the integrand in (16.15) or (16.18) has appreciable values only at tip positions very close to the lower turnaround point, as also indicated by the green line in Fig. 16.2a. The integrand $F_{\mathrm{ts}} \cdot g$ becomes negligible for larger values of u which, however, are still much smaller than A. Therefore, we take the limit $u \ll A$ and extend the integration limit to infinity, which results in

$$\Delta f = \frac{f_0}{\pi k A^2}\int\limits_0^{\infty}\frac{F_{\mathrm{ts}}(d-A+u)A}{\sqrt{2Au}}\mathrm{d}u = \frac{f_0}{\sqrt{2}\pi k A^{3/2}}\int\limits_0^{\infty}\frac{F_{\mathrm{ts}}(d-A+u)}{\sqrt{u}}\mathrm{d}u\ . \tag{16.19}$$

The dependences on resonance frequency and spring constant are the same as for the small amplitude limit (13.11). Furthermore, the frequency shift is proportional to $A^{-3/2}$. Unlike the original integral in (16.18) the integral in (16.19) does not depend on the oscillation amplitude, due to the large amplitude limit.

The expression for the frequency shift in (16.19) contains two contributions. The frequency shift depends on the tip-sample force and also on the cantilever parameters and experimental parameters. This allows to separate these parameters from the integral over the tip-sample force by defining a *normalized frequency shift* γ as

$$\gamma = \Delta f \frac{kA^{3/2}}{f_0} \ . \tag{16.20}$$

The normalized frequency shift has the following significance: Multiplying the experimentally measured frequency shift Δf by the factor $kA^{3/2}/f_0$, the expression (16.19) can be written as

$$\gamma = \frac{1}{\sqrt{2}\pi} \int\limits_0^\infty \frac{F_{\text{ts}}(d - A + u)}{\sqrt{u}} \text{d}u \ . \tag{16.21}$$

The normalized frequency shift depends only on an integral over the tip-sample force (which is also independent of the amplitude), while the dependence on the parameters k, f_0, and A is factored out.

The normalized frequency shift is particularly useful in order to compare experimental results obtained using different cantilevers (with different spring constants, and resonance frequencies) or results obtained using different oscillation amplitudes. The influence of all these parameters is factored out using the normalized frequency shift. In Fig. 16.3a measurements on a graphite sample are shown. The frequency shift is plotted as a function of tip-sample distance. Different frequency shift curves are obtained, for different oscillation amplitudes (always using the same cantilever). According to the previously obtained dependence, the measured frequency shift increases with decreasing oscillation amplitude. In Fig. 16.3b the normalized frequency shift is plotted, showing that all curves for different amplitudes collapse to one curve. This demonstrates the usefulness of the normalized frequency shift.

Now we evaluate the normalized frequency shift for a very simple model force which has a constant value of F_0 from the lower turnaround point up to a distance λ and is zero for larger distances. For this case, the normalized frequency shift can be evaluated using (16.21) as

$$\gamma = \frac{F_0}{\sqrt{2}\pi} \int\limits_0^\lambda u^{-1/2} \text{d}u = \frac{\sqrt{2}}{\pi} F_0 \sqrt{\lambda}. \tag{16.22}$$

To give some numbers: For $f_0 = 200\,\text{kHz}$, $F_0 = 2\,\text{nN}$, $A = 10\,\text{nm}$, $k = 10\,\text{N/m}$ and $\lambda = 0.1\,\text{nm}$ a normalized frequency shift of $9\,\text{fN}\sqrt{\text{m}}$ results, corresponding to a frequency shift of $\Delta f = 180\,\text{Hz}$. For an exponentially decaying force

$$F(z) = F_0 e^{-u/\lambda}, \tag{16.23}$$

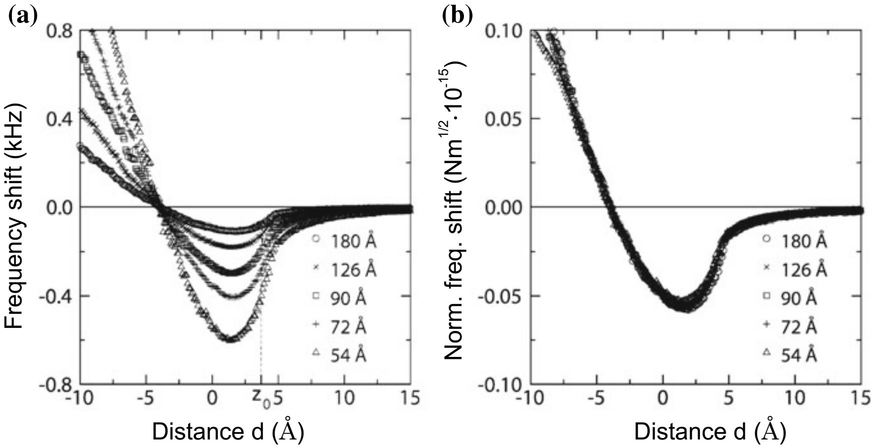


Fig. 16.3 **a** Experimentally measured frequency shift on a graphite sample as a function of the average tip-sample distance *d* for different values of the oscillation amplitude. The curves are shifted along the horizontal axis in order to make them comparable [3]. **b** If the normalized frequency shift is used as vertical axis, all curves for different amplitudes collapse to one curve, showing that the normalization has factored out the dependence on the amplitude (Reproduced with permission from [3])

the corresponding normalized frequency shift (16.21) can be calculated in the large amplitude limit as [7]

$$\gamma = \frac{1}{\sqrt{2\pi}} F_0 \sqrt{\lambda}, \tag{16.24}$$

which is (apart from a constant factor) the same result as obtained for a constant force F_0 with a range λ, shown in (16.22). Also for other forms of the tip-sample interaction, such as the Lennard–Jones interaction, the normalized frequency shift can be found in the literature [7].

16.1.3 Recovery of the Tip-Sample Force

In this chapter, we have derived equations of the (normalized) frequency shift for a given tip-sample force. Actually the reverse is desirable: It is desirable to recover the tip-sample force from the measured frequency shift. However, due to the integral present in (16.15) this equation cannot easily be inverted analytically to a solution for $F_{\text{ts}}(\Delta f)$. In the small amplitude limit the obtained equation

$$\Delta f(d) = -\frac{f_0}{2k} \frac{\partial F_{\text{ts}}(d+z)}{\partial z}\bigg|_{z=0}, \tag{16.25}$$

can be inverted to

$$F_{\mathrm{ts}}(d) = \frac{2k}{f_0} \int_{d}^{\infty} \Delta f(z')\mathrm{d}z'. \tag{16.26}$$

The integration up to infinity shows that the frequency shift should be measured up to a position relatively far from the surface. For larger oscillation amplitudes, (16.15) can be inverted using approximations which allow the determination of the force with an accuracy of ~5% [8, 9]. The use of a dynamic AFM mode in order to determine the tip-sample force, instead of the force-distance curves in static AFM, has also the advantage that no snap-to-contact occurs.

16.2 Experimental Realization of the FM Detection Scheme

We have mentioned that in the FM detection mode the cantilever oscillation is always at resonance, i.e. it always follows the resonance frequency which changes under the influence of the tip-sample force. Now we will describe how this is achieved by the experimental setup. In this section, we introduce detection schemes which are used in the FM detection mode. Here it is not the amplitude change that is measured in response to a shift of the resonance frequency, but rather the shift of the resonance frequency itself is measured.

16.2.1 Self-Excitation Mode

In the self-excitation mode no external oscillator is used, but the cantilever itself as a oscillator is the frequency-determining element in an electronic oscillator circuit. A positive feedback is used in order to self-excite the cantilever. A schematic of the implementation (Fig. 16.4) consists of an oscillator loop in which the measured oscillation signal is fed back (after a phase shift) as the driving signal of the cantilever. We will first discuss some essentials of this oscillator feedback loop and subsequently discuss its experimental realization. In addition to this oscillator feedback loop, the measured frequency shift of the resonance frequency Δf is used in an outer z-feedback feedback loop in order to control the tip-sample distance.

In a mechanical harmonic oscillator which is oscillating at resonance there is a $-90°$ phase shift of the displacement of the cantilever tip relative to the mechanical excitation, i.e. the cantilever oscillation is lagging behind the excitation. The detection of the cantilever deflection (by the photodiode and the preamplifier in the current example) is so fast that the deflection signal is sampled many times during one oscillation period. In the self-excitation scheme the measured cantilever oscillation signal is fed back as the excitation signal into the cantilever driving the piezo actuator (Fig. 16.4). In order to excite the cantilever with the correct resonance phase, a phase

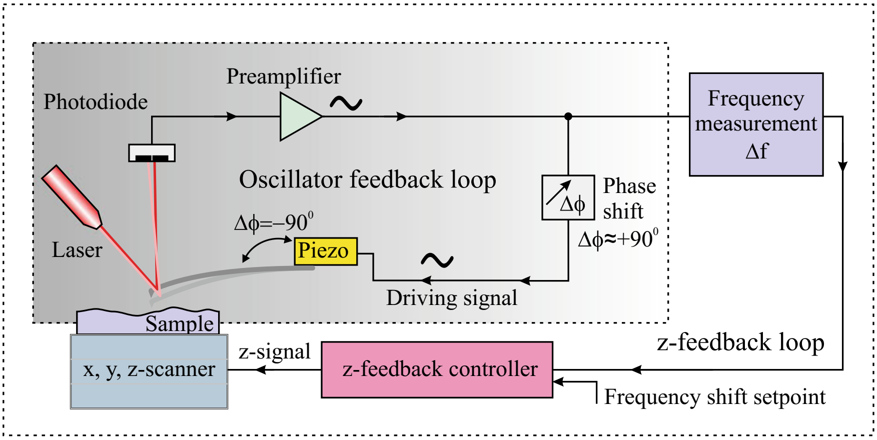


Fig. 16.4 Schematic of an FM detection setup operated in the self-excitation mode. In the circuit the measured cantilever oscillation signal is phase shifted (compensating the −90° phase shift between excitation and oscillation) and fed back to the piezo actuator driving the cantilever. In addition to this (inner) oscillator feedback loop, the measurement of the shift of the resonance frequency Δf is used in an outer z-feedback feedback loop in order to control the tip-sample distance

shift of +90° has to be applied to the oscillation signal before feeding it back as the driving signal.[2] This phase shift "compensates" the −90° phase shift between mechanical excitation and oscillation of the cantilever. For simplicity, we neglect now all other phase shifts present in the loop, for instance in the preamplifier.

If due to a tip-sample interaction the resonance frequency of the cantilever changes, the cantilever oscillation will adapt to this new resonance frequency (how fast this process occurs we will be discussed below). Since the driving in the self-excitation mode is performed using the (shifted) cantilever oscillation, the cantilever will be always driven at its shifted resonance frequency. If additionally the phase of the signal driving the oscillator is −90°, this means that the oscillator is automatically always fed at the resonance condition (frequency and phase). Thus, the oscillation frequency tracks (follows) the shift of the resonance frequency and the self-excitation mode maintains oscillation always at the resonance frequency.

Since there is no external oscillator included driving the cantilever, the question arises as to how the cantilever oscillation is excited in the first place. The cantilever is *thermally* excited in a broad frequency range. Thermal excitation can be considered as white noise, i.e. having frequency components at all frequencies (cf. Chap. 17). If a frequency component of the thermal noise does not "hit" the resonance, the oscillation amplitude at this frequency will be small. The frequency component of the white noise which "hits" the resonance will be amplified Q times due to the resonance enhancement (transfer function) of a harmonic oscillator at the resonance

[2] The technical realization of the phase shift depends on the actual implementation. It is easy to write $\Delta\phi = +90°$, but this has to be realized in practice. Let us assume that in a digital electronics the phase shift is implemented by a corresponding time delay of the oscillation signal.

frequency. Therefore, while uniformly excited over a wide frequency range by thermal noise, a large oscillation amplitude occurs only at the resonance frequency. Due to this resonance enhancement the self-excitation mode self-excites its oscillation at the resonance frequency from thermal noise. This self-excitation works best for cantilevers with high quality factors. In the case of systems with low quality factors (like measurements in liquids), starting the self-exciting oscillation is a problem. Also if the cantilever has multiple resonances, the self-excitation mode can be a bad choice. These problems are overcome in the PLL tracking mode of FM detection, which will be discussed in Sect. 16.2.3.

Another question is: How fast does the oscillation of the cantilever follow a change of the resonance frequency in the self-excitation mode? Let us assume an instantaneous change of the resonance frequency of the cantilever due to a change of the tip-sample interaction. For the case of AM detection, we have seen in Sect. 13.5 that after a change of the resonance frequency of the cantilever the new steady-state amplitude and phase are reached only after a large time constant $\tau_{\text{cant}} = 2Q/\omega_0$, corresponding to about Q oscillations.

The reason for the occurrence of the response time is that it takes time to transfer energy into, or remove energy from, the cantilever system during a transition to a new state with different amplitude/frequency. In order to see why the shifted resonance frequency is adapted very fast in the FM AFM mode, compared to the adaption of the new amplitude in the AM mode, let us compare the change of the energy of the cantilever oscillation upon a change of the resonance frequency in both cases.

The energy difference between the free oscillator and the state with tip-sample interaction present is compared for the two cases AM detection and FM detection.[3]

In the AM mode (e.g. tapping mode), a typical setpoint amplitude is 90% of the free amplitude. The energy difference between the free oscillator and the oscillator with tip-sample interaction present results as

$$\Delta E_{\text{AM}} = E_{\text{free}} - E_{\text{ts}} = \frac{1}{2} m\omega_0^2 A^2 - \frac{1}{2} m\omega_0^2 (0.9A)^2 = 0.19\ E_{\text{free}}. \tag{16.27}$$

In FM detection, the change of the energy occurs due to a change of the oscillation frequency, not the amplitude, which is kept constant in FM detection. A change of the resonance frequency from ω_0 to ω_0' leads to an energy change of

$$\begin{aligned}\Delta E_{\text{FM}} &= E_{\text{free}} - E_{\text{ts}} = \frac{1}{2} m\omega_0^2 A^2 - \frac{1}{2} m\omega_0'^2 A^2 \\ &= \frac{1}{2} m\omega_0^2 A^2 \left(1 - \frac{\omega_0'^2}{\omega_0^2}\right) \approx E_{\text{free}} \frac{2\Delta\omega}{\omega_0}.\end{aligned} \tag{16.28}$$

[3] This transition from the free state to the state with tip-sample interaction present (working point) gives an upper limit for energy changes occurring during scanning. Deviations from the setpoint values (amplitude/frequency shift) under feedback operation are smaller than the deviations in amplitude/frequency shift between the free cantilever and the situation with tip-sample interaction present.

Typical values for the frequency shift in the FM detection mode are $\Delta\omega/\omega_0 = 10^{-4}$. Due to the small frequency shifts involved, the energy difference in FM mode is very small. According to (16.28) the energy change between the free cantilever and the cantilever under tip-sample interaction is $2 \times 10^{-4} E_{\text{free}}$ in the FM mode, which is thousand times smaller than in the AM mode according to (16.27).

According to the definition of the Q-factor in (2.44), a damped harmonic oscillator can gain/lose roughly $1/Q$th of its energy in per cycle $E_{\text{diss}} = 2\pi E_{\text{osc}}/Q$. Thus, for a Q factor of 10,000 an energy of $6 \times 10^{-4} E_{\text{osc}}$ can be dissipated per cycle, which is three times more than the energy change occurring in the FM mode. Hence the FM mode is not limited by slow response times for high Q-factors occurring for operation under vacuum conditions, as is the case for AM detection.

The fundamental reason for the slow response in AM detection is that a large energy change is required in order to change the amplitude, while in the FM detection scheme the energy change due to a change of the oscillation frequency of the sensor is much smaller. This energy can be easily supplied or dissipated by the driving excitation within one oscillation cycle. This fast adaption to a new shifted resonance frequency leads to an intrinsically very high bandwidth of the FM detection scheme. However, to detect a frequency shift of e.g. $\Delta\omega = 10^{-4}\omega_0$ and below will require a certain measurement (averaging) time which reduces the intrinsically high bandwidth.

The self-excitation mode offers the fastest possible tracking of a shift of the resonance frequency, not limited by an external electronics as in the case of the tracking mode which will be considered in Sect. 16.2.3.

After clarifying the fundamental issues i.e. phase shift of $+90°$ in order to maintain the resonance phase, self-excitation of the oscillator from thermal noise, and the tracking of the shifted resonance frequency, we now discuss the experimental realization of the outer z-feedback loop.

As discussed above, in the self-excitation mode the frequency of the cantilever oscillation automatically follows the resonance frequency of the cantilever. This frequency shift is measured by the frequency measurement unit in Fig. 16.4. We will go into the details of the frequency measurement later. For the moment let us assume that the frequency measurement unit delivers a voltage signal proportional to the frequency shift. This frequency shift signal is used as the feedback signal in order to control the tip-sample distance (z-feedback) in a second (outer) feedback loop. A fixed frequency shift is chosen as the setpoint and corresponds to a certain tip-sample distance. During an xy-scan a height contour of constant frequency shift is considered as the topography of the sample.

16.2.1.1 Amplitude Control and Dissipation

In FM detection, conservative and dissipative tip-sample interactions can be measured separately. The conservative part is measured via the measurement of the frequency shift, as discussed above. A dissipative tip-sample interaction leads to a reduction of the amplitude at resonance, but does not change the resonance frequency, as discussed in Sect. 13.6 (Fig. 13.10) for the case of a harmonic oscillator. Therefore, in FM detection the conservative tip-sample interaction and the dissipative tip-sample interaction can be separated by measuring the frequency shift on the one hand, and the amplitude change on the other hand. In the actual implementation, the oscillation amplitude is controlled to a fixed value by adjusting the driving amplitude. If energy is dissipated by the tip-sample interaction the oscillation amplitude would decrease. However, an increased driving amplitude will restore the desired (setpoint) oscillation amplitude. This amplitude-controlling part of the self-excitation scheme is included in the setup shown in Fig. 16.5.

In order to maintain the oscillation amplitude at a certain setpoint value, the following scheme is applied. The amplitude of the cantilever oscillation signal is measured by an amplitude detection scheme (amplitude measurement block in Fig. 16.5). In a simple implementation an RMS-amplitude-to-DC converter can be used, in which the signal is rectified and low-pass filtered, resulting in a DC voltage proportional to the oscillation amplitude. The difference of this DC voltage to the amplitude setpoint value is taken as the error signal for an amplitude PI controller. The phase-shifted driving signal is multiplied by the appropriate amplitude factor obtained from the amplitude PI controller. In this way a constant cantilever oscillation amplitude is maintained by adjusting the amplitude of the driving signal.

The amplitude multiplication factor in the amplitude control depends on the tip-sample dissipation energy as follows. If energy is lost by an increasing tip-sample dissipation, the oscillation amplitude decreases. This is detected by the amplitude detection unit and compared to the desired amplitude setpoint. The output of the amplitude control unit (PI controller) is a multiplication factor by which the driving signal is multiplied in order to generate a constant cantilever oscillation amplitude. Therefore, this amplitude multiplication voltage can also serve as an output signal related to the dissipation. This dissipation signal can be recorded as a free signal dur-

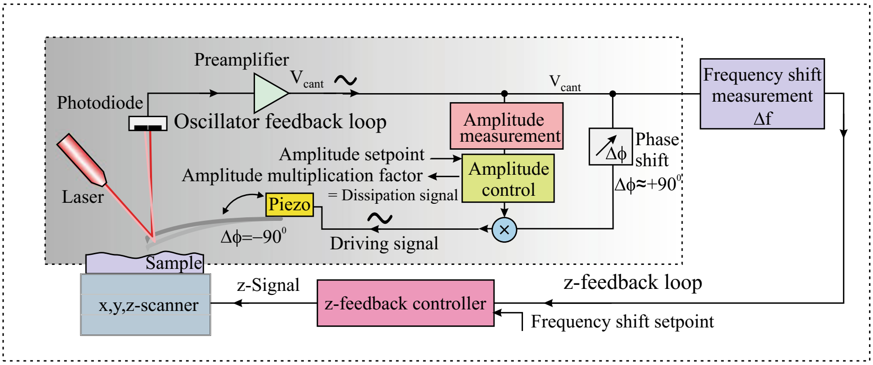


Fig. 16.5 Schematic of an FM detection setup operated with self-excitation including the amplitude control part. The cantilever oscillation amplitude is measured and maintained at a setpoint value by multiplying the driving signal by a proper multiplication factor. This factor relates to the energy dissipated by the tip-sample interaction

ing a scan. The relation between the oscillation amplitude and the energy dissipated by the tip-sample interaction is given by (14.19) with $\phi = +90°$.

The $+90°$ phase shift applied in the feedback circuit in order to drive the cantilever at resonance is an idealization. In practice additional phase shifts of other components (e.g. the preamplifier) in the circuit have to be compensated. The phase shift in the unit called the phase shift in Fig. 16.5 is adjusted to the resonance condition (deviating from +90°) in such a way that a minimum driving amplitude is required in order to establish a certain oscillation amplitude of the cantilever (resonance condition). If this adjustment of the phase shift is done for the free cantilever (at ω_0), It may be not the completely proper shift at the actual working point at the frequency $\omega_0 + \Delta\omega$.

To summarize, in the self-excitation mode the oscillation signal is fed back as the driving signal with a +90° phase shift maintaining both, resonance frequency and resonance phase. This sustains an oscillation which always follows the resonance frequency of the cantilever quasi instantaneously. The following actual measurement of this frequency will be discussed next. The amplitude multiplication factor applied to the measured oscillation signal provides information about the dissipation of the tip-sample interaction. Due to amplitude control, the cantilever oscillates at a constant amplitude. With high quality factor sensors, the oscillation will start by itself excited by thermal noise.

16.2.2 Frequency Detection with a Phase-Locked Loop (PLL)

There are several ways to measure a frequency (shift). In FM AFM the phase-locked loop detection (PLL) method is used often for this purpose, because with this method frequency shifts can be measured with high accuracy in a wide frequency range. As a starting point, we demonstrate that a change of the frequency of an oscillation can be alternatively expressed as a time-dependent phase. If the frequency of an oscillation is ω, the oscillation can be written as $\cos(\omega t + \phi_0)$. If the oscillation frequency changes at $t = 0$ form ω to $\omega + \delta\omega$, the oscillation can be expressed as $\cos\left[(\omega + \delta\omega)\, t + \phi_0\right]$. However, alternatively this expression can be rewritten as

$$\cos\left[(\omega + \delta\omega)\, t + \phi_0\right] = \cos\left[\omega t + (\delta\omega t + \phi_0)\right] = \cos\left(\omega t + \phi(t)\right), \qquad (16.29)$$

with $\phi(t) = \delta\omega t + \phi_0$. Thus, a frequency change can also be expressed as a time-dependent phase $\phi(t)$ which increases linearly with time, as shown in Fig. 16.6. The slightest frequency change corresponds to a linearly increasing phase signal. If the phase $\phi(t)$ is constant, the two frequencies are exactly the same.

In the following, the inner working of the frequency shift measurement unit in Fig. 16.5 will be explained for the case that a PLL is used for the frequency measurement. In a PLL the frequency of an internal oscillator is controlled to match (follow) the frequency of the sensor (e.g. cantilever) oscillation.

A PLL used in AFM is shown in Fig. 16.7 and consists of three main components: a phase detector, a Voltage-Controlled Oscillator (VCO), and a controller. First we

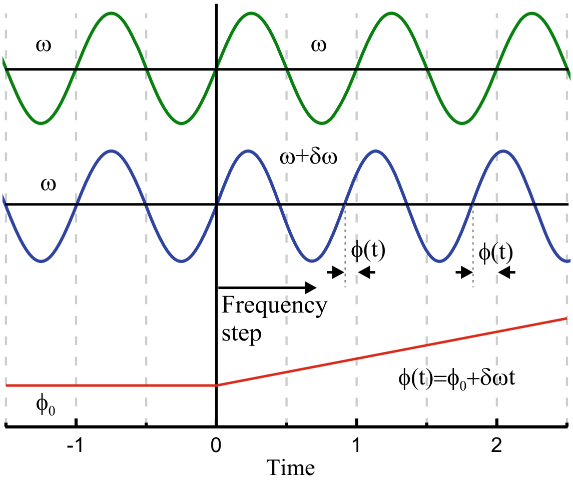


Fig. 16.6 The slightest frequency increase from ω to $\omega + \delta\omega$ leads to a linearly increasing phase $\phi(t)$. This phase (difference) can be detected using a phase detector. If the phase difference is maintained at zero, the two frequencies are the same

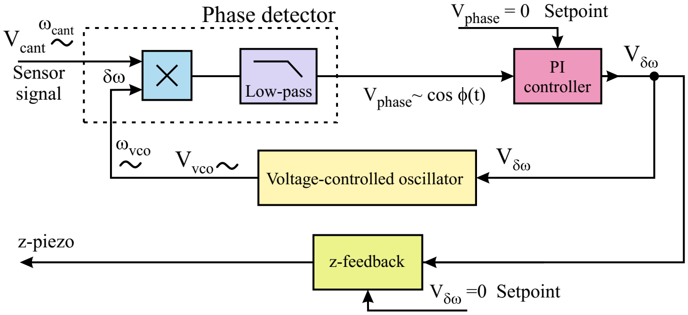


Fig. 16.7 The phase-locked loop consists of three main components: a phase detector, a Voltage-Controlled Oscillator (VCO) and a controller. These are combined to form a feedback loop in which the phase detector detects the phase difference between the AC sensor oscillation signal V_{cant} and the AC output signal (V_{vco}) of the VCO. The controller regulates the VCO frequency to a vanishing phase signal (V_{phase}). This means that the VCO frequency adapts the sensor (e.g. cantilever) frequency $\omega_{\text{vco}} = \omega_{\text{cant}}$ and the phase between the cantilever oscillation and the VCO signal is $\phi_0 = +90°$. Thus, the frequency of the VCO follows the cantilever oscillation frequency and a voltage proportional to the corresponding frequency shift $V_{\delta\omega}$ is obtained at the output of the controller and used for the z-feedback

introduce the phase detector and the VCO. Subsequently, their interaction in a phase-locked loop is described.

In the phase detector (cf. Chap. 6), the phase of the cantilever oscillation signal $V_{\text{cant}} \propto \cos(\omega_{\text{cant}} t)$ is compared to the phase of the signal from the voltage-controlled oscillator $V_{\text{vco}} \propto \cos(\omega_{\text{vco}} t + \phi_0)$ and the relative phase $\phi(t)$ is detected. In the phase detector, the two signals are multiplied and due to a mathematical identity the product can be written as

$$V_{\mathrm{vco}} \cdot V_{\mathrm{cant}} \propto \frac{1}{2} \left(\cos\left[(\omega_{\mathrm{vco}} + \omega_{\mathrm{cant}})t + \phi_0\right] + \cos\left[(\omega_{\mathrm{vco}} - \omega_{\mathrm{cant}})t + \phi_0\right] \right). \tag{16.30}$$

The low-pass filter in the phase detector removes the component with the sum of the frequencies. Thus, the signal at the output of the phase detector results as

$$V_{\mathrm{phase}} \propto \cos\left[(\omega_{\mathrm{vco}} - \omega_{\mathrm{cant}})t + \phi_0\right] = \cos(\delta\omega t + \phi_0) = \cos(\phi(t)), \tag{16.31}$$

with $\delta\omega = \omega_{\mathrm{vco}} - \omega_{\mathrm{cant}}$. The measured phase signal V_{phase} has the largest phase sensitivity for a phase close to +90°. Therefore, we consider $V_{\mathrm{phase}} = 0$ as the working point, corresponding to $\phi_0 = +90°$. Relative to this working point, the cosine function has a slope of minus one and the phase signal can be approximated (for small $\delta\omega t$) as $V_{\mathrm{phase}} \propto -\delta\omega t$. Including a proportionality factor K_{pd} which converts the phase into a voltage, the output voltage of the phase detector can be written as

$$V_{\mathrm{phase}} = K_{\mathrm{pd}} \cos\left(\delta\omega t + 90°\right) \approx -K_{\mathrm{pd}} \delta\omega t. \tag{16.32}$$

We do not consider the inner working of the voltage-controlled oscillator (VCO) here. For us the VCO is just a block in which the input voltage $V_{\delta\omega}$ controls the output frequency linearly relative to the working frequency as

$$\omega_{\mathrm{vco}} = \omega_{\mathrm{work}} - K_{\mathrm{vco}} V_{\delta\omega}, \tag{16.33}$$

with the proportionality factor K_{vco}, converting the input voltage $V_{\delta\omega}$ to a frequency shift relative to the working frequency (The minus sign in (16.33) is chosen, as a positive frequency shift $\delta\omega$ leads according to (16.32) to a negative phase voltage.). The working frequency is the frequency of the free cantilever $\omega_{\mathrm{work}} = \omega_{\mathrm{free}}$.

Now we discuss the frequency tracking capability of the PLL. For the moment, we do not consider the PI controller shown in Fig. 16.7 and assume that the phase signal V_{phase} is directly fed into the input of the VCO, i.e. $V_{\delta\omega} = V_{\mathrm{phase}}$. Let us assume that initially the frequency of the VCO matches the oscillation frequency of the cantilever, $\omega_{\mathrm{vco}} = \omega_{\mathrm{cant}} = \omega_{\mathrm{work}}$ and $\phi_0 = +90°$. At the working point $V_{\mathrm{phase}} = 0$ and thus $\delta\omega = 0$. In this case also the input voltage at the VCO vanishes, i.e. $V_{\delta\omega} = 0$.

Now we consider an increase of the actual oscillation frequency of the cantilever ω_{cant}, to ω'_{cant} by $\delta\omega$, e.g. due to a change in the tip-sample interaction. According to (16.32) this increase of the frequency by $\delta\omega$ leads to a phase signal measured by the phase detector $V_{\mathrm{phase}} \approx -K_{\mathrm{pd}}\delta\omega t$, which evolves linearly with time. With this input, the output frequency of the VCO increases according to (16.33) and since $V_{\delta\omega} = V_{\mathrm{phase}}$ as

$$\omega_{\mathrm{vco}} = \omega_{\mathrm{work}} - K_{\mathrm{pd}} K_{\mathrm{vco}} \cos\left(\delta\omega t + \phi_0\right) \approx \omega_{\mathrm{work}} + K_{\mathrm{pd}} K_{\mathrm{vco}} \delta\omega t. \tag{16.34}$$

According to (16.34), a linearly increasing phase $\delta\omega t$ leads to an increasing ω_{vco}. This reduces the frequency difference $\delta\omega$ between the cantilever frequency and the frequency of the VCO. For a varying $\delta\omega(t)$ the term $\delta\omega$ in (16.34) should be replaced by the integral $\int \omega(t)dt$. The closer ω_{vco} comes to ω'_{cant} (decreasing $\delta\omega$), the smaller is the contribution to the integral. Any remaining finite frequency mismatch $\delta\omega$ leads over time to an increasing phase $\delta\omega t$ bringing the VCO frequency closer to ω'_{cant}. In this way, the VCO frequency adapts to the increased frequency of the cantilever $\omega_{\mathrm{work}} + \delta\omega$. Due to this mechanism the VCO frequency is said to be locked to the cantilever frequency. In the steady-state $\omega_{\mathrm{vco}} = \omega_{\mathrm{cant}}$ and the frequency mismatch $\delta\omega$ vanishes.

In the terminology of the PLL: The VCO frequency is *locked* to the cantilever oscillation frequency by a *phase* comparison of both signals in a feedback *loop*. Hence, the name *phase-locked loop*. In this way, the PLL measures the frequency of the AFM sensor as the voltage $V_{\delta\omega}$. This voltage, which is proportional to the frequency shift $\delta\omega$, is used in the z-feedback loop to control the tip-sample distance. A certain tip-sample distance corresponds to a certain frequency shift voltage $V_{\delta\omega}$, which is kept constant by the z-feedback loop (Fig. 16.5).

The original cantilever signal is a high-frequency signal close to ω_0, which is modulated to slightly lower or higher frequencies (at a much lower frequency) by the tip-sample interaction, for instance during scanning of an atomic corrugation (yet without z-feedback). The PLL converts (demodulates) this modulated high frequency signal to a low frequency signal proportional to the frequency modulation of the high frequency signal. This is called FM demodulation and also occurs in an FM radio receiver, where a high-frequency carrier signal is modulated by a low-frequency audio signal and the demodulation of the audio signal is desired.

Up to now we have concentrated on the frequency tracking capability of the PLL, while we turn now to the offset phase difference ϕ_0. Remember, that two frequencies are the same, if the relative phase is constant, not necessarily $+90°$. If at a time t_0 the frequency of the VCO has adapted completely to the cantilever frequency (assumed to be increased by $\delta\omega$), a constant phase offset, different from $+90°$, can be present between both oscillations. However, the value $\phi_0 = +90°$ is desired, because it leads to maximum sensitivity of the phase detector.

The PI controller unit of the PLL maintains the desired $\phi_0 = +90°$ as follows. Any deviation from the working point of maximum phase sensitivity $\phi_0 = +90°$ leads, according to (16.31) to a finite V_{phase} (even if $\delta\omega = 0$). Thus, the PI controller enforces a vanishing V_{phase} signal, by the setpoint value $V_{\mathrm{phase}} = 0$. The PI controller controls the offset phase ϕ_0 to the desired value of $+90°$ between V_{cant} and V_{vco} by generating an appropriate controller output signal $V_{\delta\omega}$.

16.2.3 PLL Tracking Mode

We have considered the cantilever as an ideal harmonic oscillator. Due to the non-ideal properties of the mechanical cantilever oscillator, the cantilever oscillation can

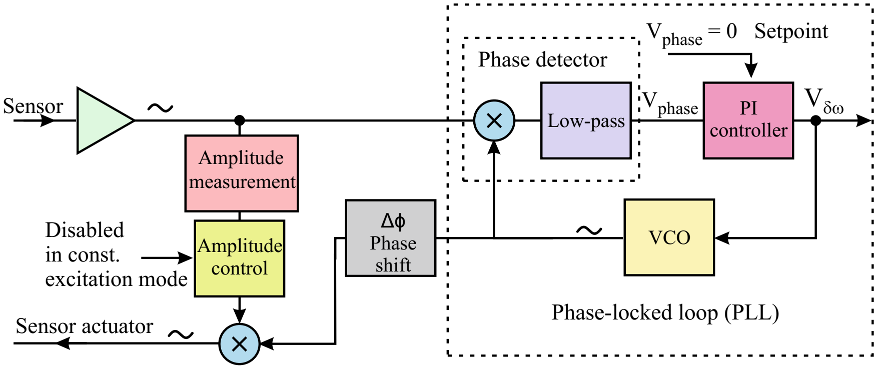


Fig. 16.8 Schematic of an FM AFM control in the PLL tracking mode. In this mode, the sensor is excited by a very clean sinusoidal driving signal taken from the voltage-controlled oscillator (VCO)

deviate from the ideal sinusoidal shape. Moreover, a cantilever and the whole AFM is a 3D object that has many modes which can sometimes be located at frequencies close to each other. An excitation of modes close to the desired resonance frequency can also lead to deviations from a clean sinusoidal oscillation. In order to feed the cantilever with a very clean sinusoidal signal (also free of noise from the detection system), at the correct frequency and phase, the PLL tracking mode is often used instead of the self-excitation mode.

In the PLL tracking mode, the signal at the output of the VCO, which has a very clean sine shape, is used to excite the cantilever (Fig. 16.8). The cantilever deflection signal (sensor signal) is fed to the input of the PLL (we neglect the amplitude control for the moment). Due to the phase setpoint of the feedback controller $V_{\text{phase}} = 0$, the VCO delivers a signal which is +90° phase shifted relative to the oscillation signal, resulting in a driving at resonance. In order to compensate for for additional phase shifts in the loop, another phase shift is applied to the VCO signal driving the sensor (phase shift unit in Fig. 16.8).

This mode of operation corresponds to the same (resonance) phase relation as in the self-excitation mode, however realized in a somewhat more indirect way via the PLL. Due to this, the time constant of the PLL electronics enters: the time to detect the changed resonance frequency plus the time to generate the corresponding driving signal. However, also in the tracking mode the cantilever excitation occurs at resonance (frequency and phase), like in self-excitation. Also in the tracking mode the response is not limited by the large Q-dependent time constant $\tau_{\text{cant}} = 2Q/\omega_0$ present in the AM detection mode.

The PI controller in the PLL loop (Fig. 16.8) is of specific importance if the VCO excites a harmonic oscillator (the cantilever), as is the case in the PLL tracking mode. Without the PI controller, any frequency deviation from the working point by $\delta\omega$ leads to a phase offset ϕ_0 different from +90°. Specifically for cantilevers with high Q-factors, even small frequency shifts lead (according to (Fig. 2.5)) to a large phase shift driving the cantilever at a phase different from +90°, i.e. out of the resonance condition. The required driving of the cantilever at resonance is maintained by the use of a PI controller in the PLL. Using the PI controller in the PLL loop, the phase signal ($V_{\mathrm{phase}} = \cos(\delta\omega t + \phi_0)$) is kept at zero by delivering a proper $V_{\delta\omega}$ signal. Thus, with a PI controller both the phase shift of $\phi_0 = +90°$ (driving the cantilever at resonance) as well as tracking the VCO frequency to the cantilever frequency ($\delta\omega = 0$) are maintained.

The oscillation amplitude control is usually implemented in the same way as in the self-excitation mode. In a variant of the PLL tracking mode the oscillation amplitude is not kept at a constant value, but the sensor excitation amplitude is set to a fixed value. This mode is called constant excitation mode.

16.3 The Non-monotonous Frequency Shift in AFM

FM detection can be operated both in the attractive and also in the repulsive regime of the tip-sample force. This advantage also involves a disadvantage. The measured property, the frequency shift, depends non-monotonously on the tip-sample distance, as can be seen in Fig. 16.3a and schematically in Fig. 16.9a. Due to this, the tip-sample distance can only be controlled by the feedback in a certain range of distances. As shown in the following, instabilities occur outside of this range.

In tapping mode AFM the measured signal (oscillation amplitude) increases monotonously (approximately linear) with increasing tip-sample distance (cf. Fig. 14.2). This leads to stable feedback, i.e. the feedback controller “knows what to do”. If the oscillation amplitude becomes smaller (e.g. due to moving over a step edge), the tip has to be withdrawn from the sample in order to recover the desired amplitude setpoint. A severe problem arises if the measured signal changes in a *non-monotonous* way with the tip-sample distance.

Let us assume that stable feedback is established at the tip-sample distance d_1 in the attractive regime at the frequency setpoint ω_1 (working point 1 in Fig. 16.9a). Here the frequency shift $\Delta\omega(d)$ has a positive slope. Due to some event, like a steep step edge, the tip-sample distance can potentially decrease suddenly to d_2, corresponding to a frequency ω_2 (assumed to be smaller than ω_1). The feedback would now try to restore the setpoint frequency (shift) ω_1. However, due to the opposite slope of the frequency shift at point 2, the feedback moves the tip closer and closer to the surface. The feedback “thinks” the tip has to be moved towards the sample in order to restore the more negative frequency shift ω_1. This will lead to a catastrophic event (positive feedback) in which the tip crashes into the sample up to the maximum range the z-piezo element can extend. The change from one branch of the frequency shift curve to that of the opposite slope can occur for various reasons: a steep slope in the surface topography, a protrusion on the surface, noise in the measurement signal and a material dependent lateral change of the interaction potential (i.e. a branch of opposite slope is reached at a different material on the sample, cf. Fig. 13.6).

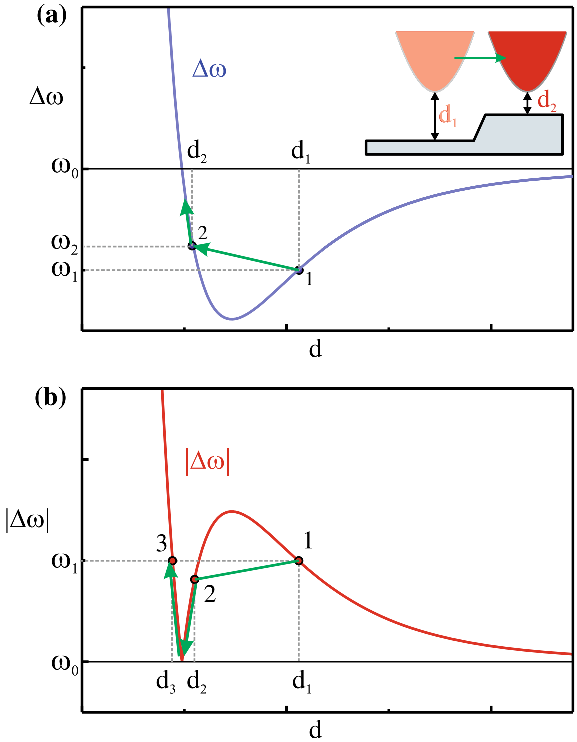


Fig. 16.9 Instabilities arise due to the non-monotonous dependence of the measured frequency shift as a function of the tip-sample distance. **a** A stable working point 1 (at a tip-sample distance d_1) in the attractive regime can be left, for instance due to a fast scan over a steep step edge (inset). This moves the system to a new working point at d_2 (before the feedback acts) with opposite slope of the frequency shift, leading to a wrong direction of the subsequent feedback action and to a crash of the tip into the surface. **b** Catastrophic events can be prevented by using the absolute value of the frequency shift as the signal for the feedback. Also in this case, the working point at d_1 is lost if the tip-sample distance changes suddenly to d_2, but instead of a catastrophic tip crash a stable working point in the repulsive branch at d_3 is reached

Stable feedback can be provided only for a range in which the measured signal monotonously increases (decreases) with the tip-sample distance. One way to improve the situation is not to use the frequency shift, but the *absolute value* of the frequency shift $\Delta\omega$ as the feedback signal, as shown in Fig. 16.9b [10, 11]. If here the working point at d_1 is left, also an instability occurs in the region of opposite (positive) slope, for instance at point 2. However, in this case no catastrophic event occurs since the tip approaches the surface only until stable feedback is resumed in the branch with a negative slope and an unintended, but stable working point 3 is reached at distance d_3 instead of d_1. Thus, using the absolute value of the frequency shift signal avoids catastrophic tip crashes and stabilizes the feedback, when moving to the branch of opposite slope in the repulsive regime. However, the intended working point in the attractive regime will be replaced by a working point in the repulsive regime.

Another way to cope with this non-monotonous frequency shift is to work in the constant height mode. In this case no instability will occur, since the feedback is off. However, the constant height mode can be operated only for very flat surfaces and under very stable conditions where drift does not change the height, i.e. at low temperatures.

16.4 Comparison of Different AFM Modes

In the previous chapters, we have discussed several modes of AFM operation, which we will now compare. In Table 16.1 operating modes are sorted along two coordinates: the operating mode can be static or dynamic and the interaction regime can be attractive or net repulsive. Often the static AFM is taken to be synonymous with contact AFM (net repulsive interaction), while dynamic AFM is taken to be synonymous with non-contact AFM (attractive interaction). However, also the off-diagonal elements in Table 16.1 are possible.

The static AFM is usually operated with tip and sample in contact (snap-to-contact), which corresponds to the upper left entry in the table. However, the static detection method can also be used in the regime of attractive interaction (non-contact). For instance, long-range electric or magnetic forces can be measured using static AFM in the non-contact mode (lower left off-diagonal element in the table). In this mode possible instabilities can lead to snap-to-contact.

In the dynamic modes, snap-to-contact is avoided and the contact/non-contact "coordinate" has to be assigned differently. The contact regime can be assigned to the range where a net repulsive force acts between the tip and sample, while in non-contact the force between tip and sample is attractive.

In the dynamic modes, we measure changes in the vibrational properties of the cantilever due to tip-sample interactions. The measured properties include the resonance frequency, the oscillation amplitude, and the phase between excitation and oscillation of the cantilever. The dynamic AFM can either operate in the non-contact mode (lower right entry in the table) or in the intermittent contact mode (tapping mode) where a repulsive tip-sample contact is established at the lower turnaround point of the oscillation (upper right off-diagonal entry in the table). In dynamic mode, snap-to-contact has to be avoided because no oscillation can be sustained in the snapped-in state. Therefore, cantilevers used in the dynamic mode have a higher force constant than cantilevers used in contact mode, or alternatively the amplitudes used are large.

Table 16.1 Operating modes of AFM ordered in two "coordinates": static/dynamic mode and attractive/net-repulsive interactions

	Static AFM	Dynamic AFM
Contact Net-repulsive interaction	Contact mode: $k \sim 1\,\text{N/m}$	Tapping mode: $k \sim 20$–$100\,\text{N/m}$
Non-contact Attractive interaction	Non-contact mode: $k \sim 1\,\text{N/m}$	AM/FM non-contact mode: $k \sim 20$–$10^6\,\text{N/m}$

16.5 Summary

- In the FM detection scheme the oscillation frequency follows the shift of the resonance frequency, i.e. the cantilever always oscillates at resonance.
- The frequency shift in the FM detection is given as

$$\Delta f = -\frac{f_0}{A^2 k}\langle F_{\text{ts}}(t) \cdot z(t)\rangle = -\frac{f_0}{\pi k A^2}\int\limits_{-A}^{+A} F_{\text{ts}}(d+z)\frac{z}{\sqrt{A^2 - z^2}}\text{d}z. \tag{16.35}$$

- In the large amplitude limit (amplitude much larger than the range of the tip-sample force) the normalized frequency shift γ factors the dependence on the experimental parameters out and is given by

$$\gamma = \Delta f \frac{kA^{3/2}}{f_0} \,. \tag{16.36}$$

  Thus, the normalized frequency shift depends only on an integral over the tip-sample force.
- In the self-excitation scheme the cantilever is self-excited from thermal noise at the momentary resonance frequency of the cantilever. The cantilever oscillation signal is measured and fed back (after an appropriate phase shift) as the cantilever driving signal.
- If in FM detection the amplitude is kept at a constant value (amplitude control), the corresponding multiplication factor contains information about the tip-sample dissipation.
- In the FM mode the frequency of the cantilever oscillation is usually measured by a phase-locked loop (PLL). The measured frequency shift signal is used to control the tip-sample distance via a z-feedback loop.
- In the PLL tracking mode the cantilever driving signal is taken from an oscillator of the PLL. This has the advantage of driving the cantilever with a very clean sinusoidal signal.

- The non-monotonous dependence of the frequency shift on the tip-sample distance can lead to instabilities. These can be prevented by taking the absolute value of the measured frequency shift as the signal for the z-feedback.
- The response time to adapt the steady-state oscillation signal after an instantaneous change of the tip-sample interaction is much shorter in the case of FM detection than for AM detection. Therefore, the FM detection scheme is used for with high Q-factors, i.e. in vacuum.
- The AFM modes can be ordered in two coordinates: static/dynamic and net repulsive (contact)/attractive (non-contact). The static AFM in the net repulsive regime is termed the contact mode and the dynamic mode in the attractive regime is called the non-contact mode. However, besides these regimes, the static mode can also be operated in the attractive interaction regime, and the dynamic mode can be operated in the net repulsive interaction regime (intermittent contact).

References

1. T.R. Albrecht, P. Grütter, D. Horne, D. Rugar, Frequency modulation detection using high Q cantilevers for enhanced force microscope sensitivity. J. Appl. Phys. **69**, 668 (1989). https://doi.org/10.1063/1.347347
2. S. Morita, R. Wiesendanger, E. Meyer (eds.), *Non-contact Atomic Force Microscopy* (Springer, Heidelberg, 2002). https://doi.org/10.1007/978-3-642-56019-4
3. S. Morita, F.J. Giessibl, R. Wiesendanger (eds.), *Non-contact Atomic Force Microscopy*, vol. 2 (Springer, Heidelberg, 2009). https://doi.org/10.1007/978-3-642-01495-6
4. S. Morita, F.J. Giessibl, E. Meyer, R. Wiesendanger (eds.), *Non-contact Atomic Force Microscopy*, vol. 3 (Springer, Heidelberg, 2015). https://doi.org/10.1007/978-3-319-15588-3
5. J.P. Cleveland, B. Anczykowski, A.E. Schmid, V.B. Elings, Energy dissipation in tapping-mode atomic force microscopy. Appl. Phys. Lett. **72**, 2613 (1998). https://doi.org/10.1063/1.121434
6. M.V. Salapaka, D.J. Chen, J.P. Cleveland, Linearity of amplitude and phase in tapping-mode atomic force microscopy. Phys. Rev. B **61**, 1106 (2000). https://doi.org/10.1103/PhysRevB.61.1106
7. F.J. Giessibl, H. Bielefeld, Physical interpretation of frequency-modulation atomic force microscopy. Phys. Rev. B **61**, 9968 (2000). https://doi.org/10.1103/PhysRevB.61.9968
8. J.E. Sader, T. Uchihashi, M.J. Higgins, A. Farrell, Y. Nakayama, S.P. Jarvis, Quantitative force measurements using frequency modulation atomic force microscopy-theoretical foundations. Nanotechnology **16**, S94 (2005). https://doi.org/10.1088/0957-4484/16/3/018
9. J.E. Sader, S.P. Jarvis, Accurate formulas for interaction force and energy in frequency modulation force spectroscopy. Appl. Phys. Lett. **84**, 1801 (2004). https://doi.org/10.1063/1.1667267
10. H. Ueyama, Y. Sugawara, S. Morita, Stable operation mode for dynamic noncontact atomic force microscopy. Appl. Phys. A **84**, 1801 (2004); **66**, S295 (1998). https://doi.org/10.1007/s003390051149
11. M. Heyde, M. Sterrer, H.-P. Rust, H.-J. Freund, Atomic resolution on MgO(001) by atomic force microscopy using a double quartz tuning fork sensor at low-temperature and ultrahigh vacuum. Appl. Phys. Lett. **87**, 083104 (2005). https://doi.org/10.1063/1.2012523

Chapter 17
Noise in Atomic Force Microscopy

In topographic images, the noise in the vertical position of the tip (i.e. the noise in the tip-sample distance) should be considerably smaller than the topography signal on the sample to be measured. If atomic steps are imaged, the noise should have an amplitude much smaller than 1 Å. In the following we do not consider noise due to floor vibrations or sound, but more fundamental limits of noise due to thermal excitation of the cantilever, or due to the detection limit of the preamplifier detecting the signal.

In Sect. 11.3 we studied the shot noise due to the discrete arrival of photons at the photodiode. The minimum detectable cantilever motion and the corresponding minimum detectable force were estimated. Additionally to this fundamental limit for the detector noise, noise from the detection electronics has to be considered. The detector noise depends on the specific detection method used. Another source of noise is the thermal noise of the cantilever. The cantilever is considered to be a harmonic oscillator which is thermally excited to a certain noise amplitude $\sqrt{\langle \Delta z_{\text{th}}^2 \rangle}$. In this chapter the effect of the thermal noise amplitude on the experimentally measured quantities in AFM such as the frequency shift is estimated.

17.1 Thermal Noise Density of a Harmonic Oscillator

The thermal displacement noise of a harmonic oscillator can be estimated from the equipartition theorem, which states that each degree of freedom carries an average energy of $1/2\, k_{\text{B}}T$ in thermal equilibrium. A degree of freedom is a parameter which enters into the expression of the total energy as a squared term. For the case of a one-dimensional harmonic oscillator the energy is written as $E_{\text{tot}} = 1/2\, kz^2 + 1/2\, mv^2$, and the number of degrees of freedom is two, as z and v enter as squared terms. Thus, the equipartition theorem states that the total energy of a thermally excited harmonic oscillator is $k_{\text{B}}T$.

B. Voigtländer, *Atomic Force Microscopy*, NanoScience and Technology, https://doi.org/10.1007/978-3-030-13654-3_17

Since the total mechanical energy in a harmonic oscillator is stored on average as one half in kinetic and one half in elastic energy, the average mean square displacement $\langle \Delta z_{\text{th}}^2 \rangle$ is related to the total energy by $1/2\, E_{\text{tot}} = 1/2\, k \langle \Delta z_{\text{th}}^2 \rangle$. From this the (time) average of the square of the vibrational amplitude due to thermal noise results as

$$\langle \Delta z_{\text{th}}^2 \rangle = \frac{k_{\text{B}} T}{k}. \tag{17.1}$$

At room temperature and for a spring constant of $k = 10\,\text{N/m}$, an amplitude of ~0.2 Å results. This is quite a large value and shows that soft cantilevers with high force sensitivity have quite a large thermally excited vibrational amplitude. On the other hand, as we discussed above, stiffer cantilevers have less force sensitivity in the static mode.

In the following, we will derive the thermal noise density of a harmonic oscillator (cantilever) in contact with a heat bath. The general concept for the power spectral density of a noise signal was introduced in Sect. 5.5. The noise signal is now the deflection of the cantilever Δz and the corresponding power noise spectral density is termed $N_{\text{z,th,osc}}^2(f)$. This thermal noise density consists of two contributions. First the excitation noise (thermal noise), which is assumed to be frequency-independent white noise $N_{\text{z,th,exc}}$. The value of this thermal excitation noise density still has to be determined in the following. A second contribution to $N_{\text{z,th,osc}}^2(f)$ comes from the harmonic oscillator. The constant thermal excitation noise density is sent through the harmonic oscillator with its resonance characteristics. Thus, the resulting thermal noise density of the harmonic oscillator $N_{\text{z,th,osc}}(f)$ can be written as (neglecting the subscript z)

$$N_{\text{th,osc}}(f) = N_{\text{th,exc}} G(f), \tag{17.2}$$

with $G(f)$ being the transfer function of the harmonic oscillator. In this chapter we use the natural frequency $f = \omega/(2\pi)$, since in actual measurements the natural frequency is used. As already discussed in Chap. 2 in (2.32), the transfer function of the harmonic oscillator is

$$\frac{A^2}{A_{\text{drive}}^2} \equiv G^2(f) = \frac{1}{\left(1 - \frac{f^2}{f_0^2}\right)^2 + \frac{1}{Q^2}\frac{f^2}{f_0^2}}. \tag{17.3}$$

The mean square thermal displacement can be calculated according to (5.13). Another expression for the mean square displacement was obtained from the equipartition theorem as (17.1). Thus, the following equation results

$$\langle \Delta z_{\text{th}}^2 \rangle = \int_0^\infty N_{\text{th,osc}}^2(f)\mathrm{d}f = N_{\text{th,exc}}^2 \int_0^\infty G^2(f)\mathrm{d}f = \frac{k_{\text{B}} T}{k}. \tag{17.4}$$

Fortunately, an anti-derivative for the integral over $G^2(f)$ exists (which can be found using a computer algebra system or a table of integrals). We omit this here, however. A very simple expression results ($\int_0^\infty G^2(f)\mathrm{d}f = \pi Q f_0/2$), when the integration

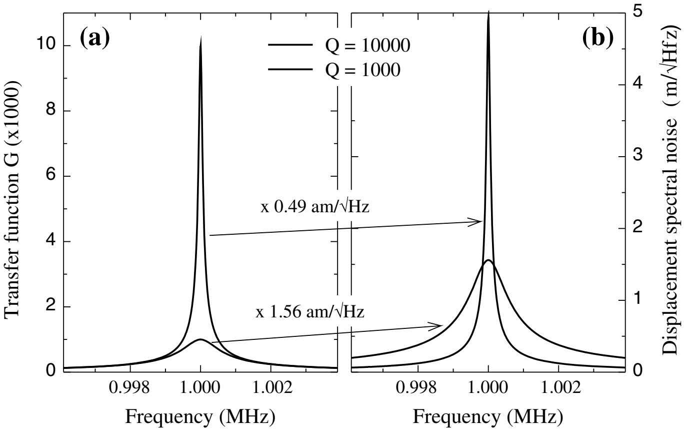


Fig. 17.1 **a** Transfer function of the harmonic oscillator $G(f)$. **b** Corresponding displacement spectral noise density $N_{\mathrm{th,osc}}$ at room temperature. The multiplication factor $N_{\mathrm{th,exc}}$ for going from **a** to **b** depends on the Q-factor

limits are inserted. With this, the spectral noise density of a harmonic oscillator results as

$$N_{\mathrm{th,osc}}(f) = N_{\mathrm{th,exc}} G(f) = \sqrt{\frac{2k_{\mathrm{B}}T}{\pi k Q f_0}} G(f). \tag{17.5}$$

Thus, the spectral noise density of the harmonic oscillator consists of the strongly peaked transfer function of the harmonic oscillator $G(f)$ and a frequency independent white thermal excitation noise density given by (17.5) as

$$N_{\mathrm{th,exc}} = \sqrt{\frac{2k_{\mathrm{B}}T}{\pi k Q f_0}}. \tag{17.6}$$

Since the white noise $N_{\mathrm{th,exc}}$ depends on the Q-factor, different multiplication factors have to be used when going from the transfer function to the displacement spectral noise density shown in Fig. 17.1b. Due to this, for high Q-factors the thermal noise of the oscillator is concentrated closer to the resonance frequency and suppressed everywhere else.

The mean square displacement is obtained by integration over the relevant frequency range. The mean square displacement noise within a bandwidth from f_1 to f_2 according to (5.14) as

$$\left\langle \Delta z_{\mathrm{th}}^2(f_1, f_2) \right\rangle = \int_{f_1}^{f_2} N_{\mathrm{th,osc}}^2(f)\mathrm{d}f = \frac{2k_{\mathrm{B}}T}{\pi k Q f_0} \int_{f_1}^{f_2} G^2(f)\mathrm{d}f. \tag{17.7}$$

This equation will be used in the following in order to evaluate the mean square displacement in various circumstances.

17.2 Thermal Noise in the Static AFM Mode

In the static case, the relevant frequencies are far below the resonance frequency and the transfer function can be approximated as $G^2 = 1$. Inserting this into (17.7), the mean square displacement in the static mode results with $B = f_2 - f_1$ as

$$\left\langle \Delta z^2_{\mathrm{th,stat}} \right\rangle = \frac{2k_\mathrm{B} T B}{\pi k Q f_0}. \tag{17.8}$$

The thermal noise amplitude of the sensor (cantilever tip) translates to the finally measured quantities, such as the minimum detectable force in static AFM. In the static AFM mode, the noise amplitude corresponds to a noise in the force measurement by Hooke's law via $\Delta F = k\sqrt{\left\langle \Delta z^2_{\mathrm{th,stat}} \right\rangle}$. Therefore, the minimum detectable force (due to thermal noise) in static AFM (i.e. at low frequencies off-resonance) is

$$F^{\mathrm{static}}_{\mathrm{min,th}} = \sqrt{\frac{2k k_\mathrm{B} T B}{\pi Q f_0}}. \tag{17.9}$$

17.3 Thermal Noise in the Dynamic AFM Mode with AM Detection

Here we consider the AM dynamic mode in which the cantilever (or more generally AFM sensor) is oscillated at, or very close to, the resonance frequency of the cantilever. Therefore, we consider $f = f_0$ and the transfer function results in $G^2 = Q^2$. Inserting this into (17.7), the mean square displacement in the dynamic mode results as

$$\left\langle \Delta z^2_{\mathrm{th,res}} \right\rangle = \frac{2k_\mathrm{B} T Q (2B)}{\pi k f_0}, \tag{17.10}$$

with $2B$ being the two sided bandwidth, i.e. from $f_0 - B$ to $f_0 + B$. The thermal displacement noise (17.10) is Q times higher in the dynamic case than in the static case (17.8). However, since also the signal (cantilever oscillation amplitude) is Q times larger in the dynamic mode due to the resonance enhancement, the signal-to-noise ratio of the cantilever deflection remains the same as in the static mode.

In the following, we derive the minimum detectable force gradient in the AM slope detection mode. The operating point in this mode is close to the maximum slope (roughly at half of the maximum amplitude) as discussed in Sect. 13.3. For simplicity, we assume that the measurement bandwidth is so narrow that the transfer function can be considered as constant with the value $1/2\,Q$ (instead of Q at the resonance). Thus, the thermal displacement noise $\sqrt{\left\langle \Delta z^2_{\mathrm{th}} \right\rangle}$ is one half of that derived from (17.10).

As we have shown before in (13.11), in dynamic AFM for small amplitudes, the force gradient is related to the measured frequency shift by $\partial F_{\mathrm{ts}}/\partial z = \Delta f 2k/f_0$ (we omit the factor -1 here). In the slope detection mode, the measured amplitude change is proportional to a frequency change with the inverse of the slope of the resonance curve at the working point as proportionality factor as

$$\frac{\partial F_{\mathrm{ts}}}{\partial z} = \frac{2k}{f_0}\Delta f = \frac{2k}{f_0}\frac{\Delta f}{\Delta A}\Delta A. \tag{17.11}$$

The inverse slope of the resonance curve at the working point can be written according to (2.40) as $\Delta f/\Delta A \approx f_0/(QA)$. If we identify the amplitude change ΔA with the thermal noise $\sqrt{\langle \Delta z^2_{\mathrm{th,res}} \rangle}$, the minimum detectable force gradient can be written as

$$\frac{\partial F}{\partial z} = \frac{2k}{f_0}\frac{f_0}{QA}\Delta A = \frac{2k}{QA}\sqrt{\frac{k_{\mathrm{B}}TQ(2B)}{\pi k f_0}} = \sqrt{\frac{4kk_{\mathrm{B}}T(2B)}{\pi Q f_0 A^2}}. \tag{17.12}$$

In order to decrease the noise large Q-factors are desirable. However, this limits the detection bandwidth due to a large time constant, as shown in Sect. 13.5. Also small k/f_0 ratios are desirable as long as no snap-to-contact occurs.

In tapping mode atomic force microscopy, a certain amplitude (attenuation) A corresponds to a certain tip-sample distance z' (i.e. distance between surface and the lower turnaround point of the oscillating tip). A noise in the deflection signal due to thermal excitation $\Delta A = \sqrt{\langle \Delta z^2_{\mathrm{th}} \rangle}$ translates to a noise in the topography signal z' via the slope of the amplitude distance relation $\mathrm{d}A/\mathrm{d}z'$ as

$$\Delta z' = \Delta A \frac{\mathrm{d}z'}{\mathrm{d}A} = \frac{\sqrt{\langle \Delta z^2_{\mathrm{th}} \rangle}}{\mathrm{d}A/\mathrm{d}z'}. \tag{17.13}$$

Here the mean square displacement has to be taken from (17.10). For stiff materials the slope $\mathrm{d}A/\mathrm{d}z'$ is about one, while it has a smaller value for soft materials and the noise in the topography signal becomes correspondingly larger.

17.4 Thermal Noise in Dynamic AFM with FM Detection

In FM modulation, a signal (or noise) component at frequency f_{mod} leads in the FM signal to two side bands at $f_0 \pm f_{\mathrm{mod}}$ above and below the carrier frequency[1] f_0, as shown in Appendix D. In the following we consider the deflection noise at $f_0 + f_{\mathrm{mod}}$. In order to evaluate the mean square displacement noise according to (17.7), we have to evaluate the transfer function at $f_0 + f_{\mathrm{mod}}$. When evaluating $G(f_0 + f_{\mathrm{mod}})$

[1] We do not indicate explicitly that the carrier frequency is the shifted resonance frequency f_0'.

we have to consider typical values of f_0 ranging from 30 kHz to 1 MHz, and f_mod lies typically in the range (far) below 1 kHz, i.e. $f_\mathrm{mod} \ll f_0$. In order to evaluate $G(f_0 + f_\mathrm{mod})$ in the limit very close to the resonance frequency, we start from (17.3) and use (2.37), resulting in

$$G^2(f_0 + f_\mathrm{mod}) = \frac{1}{\left(1 - \frac{(f_0+f_\mathrm{mod})^2}{f_0^2}\right)^2 + \frac{1}{Q^2}\frac{(f_0+f_\mathrm{mod})^2}{f_0^2}} \approx \frac{1}{4\frac{f_\mathrm{mod}^2}{f_0^2} + \frac{1}{Q^2}}. \tag{17.14}$$

If the condition $f_\mathrm{mod} > f_0/(2Q)$ is fulfilled, (which has to be checked) the term $1/Q^2$ in the denominator of (17.14) can be neglected. In this case, the square of the thermal displacement noise density can, according to (17.6), be written as

$$N_\mathrm{th,osc}^2(f_0 + f_\mathrm{mod}) \equiv N_\mathrm{z,th}^2(f_0 + f_\mathrm{mod}) = N_\mathrm{th,exc}^2 G^2(f_0 + f_\mathrm{mod}) = \frac{k_\mathrm{B} T f_0}{2\pi k Q f_\mathrm{mod}^2}. \tag{17.15}$$

We change the notation here in order to distinguish between the thermal displacement noise density to $N_\mathrm{z,th}(f_0 + f_\mathrm{mod})$, and the thermal frequency noise density after demodulation $N_{f,\mathrm{th}}(f_\mathrm{mod})$. In FM modulation, the displacement noise is transferred to a frequency noise according to (D.11) as shown in Appendix D and we can write

$$N_{f,\mathrm{th}}(f_\mathrm{mod}) = \frac{\sqrt{2} f_\mathrm{mod}}{A} N_\mathrm{z,th}(f_0 + f_\mathrm{mod}). \tag{17.16}$$

For the thermal displacement noise according to (17.15), the frequency noise density results for the case $f_\mathrm{mod} > f_0/(2Q)$ as

$$N_{f,\mathrm{th}} = \sqrt{\frac{k_\mathrm{B} T f_0}{\pi k Q A^2}} = \text{const.}, \tag{17.17}$$

which does not depend on f_mod.

The thermal frequency noise in FM detection can be calculated analogously to (5.14) by integration over f_mod up to the maximum $f_\mathrm{mod,max} = B$ as

$$\left\langle \Delta f_\mathrm{th}^2 \right\rangle = \int_0^B N_{f,\mathrm{th}}^2(f_\mathrm{mod}) d f_\mathrm{mod} = \frac{k_\mathrm{B} T f_0}{\pi k Q A^2} B. \tag{17.18}$$

The noise contributions from frequencies lower than f_0 are already included by the factor $\sqrt{2}$ in (17.16). Thus, in the FM case B is defined as $B = f_\mathrm{mod,max}$, i.e. as a single sided bandwidth.

The minimum detectable force gradient due to thermal noise can be written as

$$\frac{\partial F}{\partial z} = \frac{2k}{f_0}\sqrt{\left\langle \Delta f_\mathrm{th}^2 \right\rangle} = \sqrt{\frac{4 k k_\mathrm{B} T B}{\pi Q f_0 A^2}}. \tag{17.19}$$

17.5 Sensor Displacement Noise in the FM Detection Mode

Up to now we have considered the thermal noise of the cantilever which gives the fundamental limit of noise. Now we consider the sensor displacement noise, which in practical implementations of atomic force microscopy is often the dominant source of noise. Sensor displacement noise may be the shot noise of the photons arriving on the photodiode in the case of the laser beam deflection mode of detection. In an electrical detection scheme of the sensor displacement, the electrical noise of the preamplifier is the dominant source of detector noise. For any detection scheme, the actually measured noise of the detection voltage can be converted via a sensitivity factor into an equivalent displacement noise $N_{\mathrm{z,sens}}(f)$, which is expressed in units of m/$\sqrt{\mathrm{Hz}}$. For simplicity, we assume a white sensor displacement noise, i.e. $N_{\mathrm{z,sens}}(f_0 + f_{\mathrm{mod}}) = N_{\mathrm{z,sens}}$ is constant as a function of frequency within the considered detection bandwidth around f_0. Thus, the mean square displacement due to the sensor displacement noise results according to (5.14) as

$$\left\langle \Delta z_{\mathrm{sens}}^2 \right\rangle = \int_0^B N_{\mathrm{z,sens}}^2 \mathrm{d}f = N_{\mathrm{z,sens}}^2 B. \tag{17.20}$$

Further, the minimum detectable force in the static mode results as

$$F_{\mathrm{min,sens}}^{\mathrm{static}} = k\sqrt{\left\langle \Delta z_{\mathrm{sens}}^2 \right\rangle} = kN_{\mathrm{z,sens}}\sqrt{B}. \tag{17.21}$$

In the dynamic mode the minimum detectable force gradient due to the sensor displacement noise results according to (17.12) as

$$\frac{\partial F}{\partial z} = \frac{2k}{QA}\sqrt{\left\langle \Delta z_{\mathrm{sens}}^2 \right\rangle} = \frac{\sqrt{2}k}{QA} N_{\mathrm{z,sens}}\sqrt{2B}, \tag{17.22}$$

with $2B$ being the two-sided bandwidth. The frequency noise density of the demodulated Δf signal in FM detection results from the sensor displacement noise and can be written according to (D.11) as

$$N_{f,\mathrm{sens}}(f_{\mathrm{mod}}) = \frac{\sqrt{2} f_{\mathrm{mod}}}{A} N_{\mathrm{z,sens}}. \tag{17.23}$$

The mean square frequency noise resulting from the sensor displacement noise is

$$\left\langle \Delta f_{\mathrm{sens}}^2 \right\rangle = \int_0^B N_{f,\mathrm{sens}}^2(f_{\mathrm{mod}}) \mathrm{d}f_{\mathrm{mod}} = \frac{2N_{\mathrm{z,sens}}^2}{A^2} \int_0^B f_{\mathrm{mod}}^2 \mathrm{d}f_{\mathrm{mod}}. \tag{17.24}$$

Thus, the frequency noise due to the sensor displacement noise results as

$$\sqrt{\langle \Delta f_{\text{sens}}^2 \rangle} = \sqrt{\frac{2N_{\text{z,sens}}^2}{3A^2} B^3}. \tag{17.25}$$

In contrast to the thermal noise which did not depend on the frequency, the frequency noise due to the sensor increases with increasing bandwidth proportional to $B^{3/2}$.

The minimum detectable force gradient in FM detection due to the sensor noise results, using (17.12), as

$$\frac{\partial F}{\partial z} = \frac{2k}{f_0}\sqrt{\langle \Delta f_{\text{sens}}^2 \rangle} = \sqrt{\frac{8}{3}}\frac{kN_{\text{z,sens}}B^{3/2}}{f_0 A}. \tag{17.26}$$

17.6 Total Noise in the FM Detection Mode

As sources of noise in the FM detection mode we have considered the thermal noise (17.17) and the sensor displacement noise (17.23). Considering additionally also another source of noise, the oscillator noise $N_{\text{f,oscillator}}^2$ [1], in the general case (independent of the limit $f_{\text{mod}} > f_0/(2Q)$ used to derive (17.17)) the following expression for the combined noise density $N_{\text{f,total}}^2$ is obtained [1]

$$N_{\text{f,total}}^2 = \frac{k_{\text{B}}Tf_0}{\pi kQA^2} + \frac{N_{\text{z,sens}}^2 f_0^2}{2Q^2A^2} + \frac{2N_{\text{z,sens}}^2}{A^2} f_{\text{mod}}^2. \tag{17.27}$$

The corresponding frequency noise is obtained by integrating the frequency noise density (17.27) up to the bandwidth B, as

$$\Delta f_{\text{f,total}}^2 = \frac{k_{\text{B}}Tf_0}{\pi kQA^2} B + \frac{N_{\text{z,sens}}^2 f_0^2}{2Q^2A^2} B + \frac{2N_{\text{z,sens}}^2}{3A^2} B^3. \tag{17.28}$$

In most cases (except for very low Q-factors, as in liquids) the second terms in (17.27) and (17.28) are negligible compared to the other terms and can be neglected. If both of the other terms in (17.27) and (17.28) have non negligible contributions, two of the three involved parameters can be determined: the amplitude A which is related to the sensitivity factor S_{sensor}, the sensor displacement noise $N_{\text{z,sens}}$, or the spring constant k. Usually, the sensitivity factor S_{sensor} and the sensor displacement noise $N_{\text{z,sens}}$ are determined by fitting (17.27) to the experimentally measured frequency noise density.

An example of an actually measured frequency noise density as a function of the modulation frequency is shown in Fig. 17.2. The experimentally measured noise density is characterized by a small constant offset due to thermal noise (17.17) and a linear increase of the noise density with the modulation frequency according to

Fig. 17.2 Experimentally measured frequency noise density $N_{\mathrm{f,total}}(f_{\mathrm{mod}})$ of an FM atomic force microscopy setup

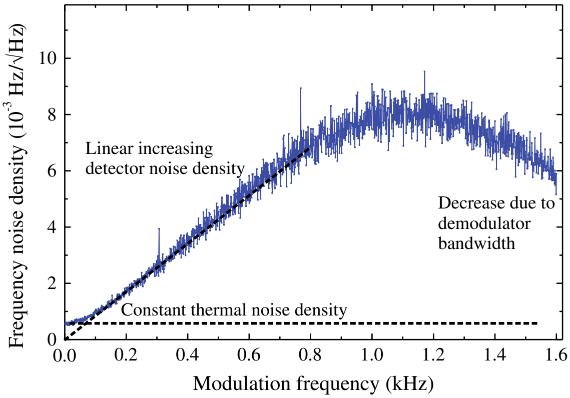


Table 17.1 Intrinsic parameters for different sensors used in atomic force microscopy

Sensor parameter	Si cantilever	qPlus tuning fork	Needle sensor
Quality factor	300	3,000	15,000
Resonance frequency (Hz)	100 k	32 k	1 M
Spring constant (N/m)	10	1,800	1.08 M
Oscillation amplitude (nm)	4	0.1	0.1
$N_{\mathrm{z,sens}}$ (fm/$\sqrt{\mathrm{Hz}}$)	100	50	2

(17.23). Due to the bandwidth of the frequency demodulator electronics, which has (in this example) a bandwidth limit of 1 kHz, the measured noise density levels off and decreases beyond this frequency.

In Table 17.1, characteristic intrinsic parameters for different sensors used in atomic force microscopy are listed for three different kinds of sensors. A typical silicon cantilever sensor is compared to a quartz tuning fork (qPlus sensor) and to a length extensional sensor (needle sensor). The detection noise densities are taken from [2]. In Table 17.2 numerical values for the noise estimated in this chapter are compared for the three different kinds of sensors.

17.7 Measurement of System Parameters in Dynamic AFM

Parameters of the AFM sensor and the measurement setup which are a priory unknown are: the spring constant of the cantilever k, the amplitude A which is determined by the sensitivity factor S_{sensor}, the sensor displacement noise $N_{\mathrm{z,sens}}$, as well as the quality factor Q and the resonance frequency f_0 of the AFM sensor used.

Methods for the determination of these parameters are considered in different sections of this book. Here we summarize these methods and consider under which circumstances the different methods can be used. The use of the methods to determine these parameters depends on (a) the type of AFM sensor (cantilever, or quartz sensor),

Table 17.2 Noise figures for different AFM sensors for a bandwidth of 1,000 Hz and $T = 300$ K

Mode	Noise figure	Si cantilever	qPlus tuning fork	Needle sensor	Equation number
Static	$\sqrt{\langle \Delta z^2 \rangle}$ (fm)	94	3.9	0.013	(17.8)
	$F_{\mathrm{min,th}}$ (pN)	0.94	6.9	14	(17.9)
	$F_{\mathrm{min,sens}}$ (pN)	32	2,800	68,000	(17.21)
AM	$\sqrt{\langle \Delta z^2_{\mathrm{th}} \rangle}$ (pm)	40	17	0.27	(17.10)
	$(\partial F/\partial z)_{\mathrm{th}}$ (N/m)	0.0005	0.1	0.3	(17.12)
	$\sqrt{\langle \Delta z^2_{\mathrm{sens}} \rangle}$ (pm)	3.2	1.6	0.063	(17.20)
	$(\partial F/\partial z)_{\mathrm{sens}}$ (N/m)	0.0001	0.019	0.091	(17.22)
FM	$(\partial F/\partial z)_{\mathrm{th}}$ (N/m)	0.0003	0.5	0.2	(17.19)
	$\sqrt{\langle \Delta f^2_{\mathrm{th}} \rangle}$ (Hz)	1.7	0.9	0.09	(17.18)
	$\sqrt{\langle \Delta f^2_{\mathrm{sens}} \rangle}$ (Hz)	0.6	13	0.52	(17.25)
	$(\partial F/\partial z)_{\mathrm{sens}}$ (N/m)	0.0001	1.5	1.1	(17.26)

(b) the quality factor of the sensor, or (c) if tip-sample contact is required or if the parameter can be determined already without tip-sample contact. The easiest is the determination of Q and f_0. Experimentally a resonance curve sweep is measured by, exciting the AFM sensor with a certain driving amplitude and sweeping the driving frequency in a range close to the resonance frequency of the cantilever, while measuring the resulting cantilever amplitude (and optionally also the phase). The experimentally measured resonance curve of a sensor is then fitted to the resonance curve of a harmonic oscillator in order to determine the resonance frequency and the quality factor. This relies on the assumption that the response of the AFM sensor can be approximated by a driven damped harmonic oscillator, which is usually fulfilled.

The spring constant k of cantilever type sensors can be determined using the geometrical data of the cantilever, or using the Sader method, as outlined in Sect. 11.6. For quartz sensors (tuning fork or needle sensors), the spring constant is usually calculated from the geometrical data, as the Sader method is not applicable for these sensors.

The amplitude sensitivity factor S_{sensor} converts the actually measured sensor output voltage $\Delta V_{\mathrm{sensor}}$ at the output of the sensor preamplifier to the deflection Δz, as $\Delta z = S_{\mathrm{sensor}} \Delta V_{\mathrm{sensor}}$. As shown in detail in Sect. 11.6 this amplitude sensitivity factor can be determined by pressing the cantilever tip to a hard sample while measuring the sensor voltage. One disadvantage of this method is that it can potentially lead to a modification (blunting) of the tip. This method can only be applied to cantilever type sensors, as the quartz sensors have a too high spring constant ($k > 2000$ N/m), which would lead to a damage of the sensor tip during this procedure. For quartz sensors the sensitivity can be calculated from the charge induced on the electrodes of these sensors, as outlined in [2]. The experimental determination of the amplitude sensitivity for quartz sensors is outlined in Sect. 18 and in [3].

The thermal method outlined in detail in Sect. 11.6 can be applied to cantilever type sensors as well as to quartz sensors considering their fundamental mode as a harmonic oscillator. In this method the spring constant and the mean thermal displacement are related to the thermal energy by $k\left\langle \Delta z_{\mathrm{th}}^2\right\rangle = k_{\mathrm{B}}T$. Due to this, either the spring constant k or the amplitude sensitivity factor can be determined if the respective other is known. The advantage of the thermal method is that it can be applied non-destructively, i.e. without tip-sample contact. If sensors with very high quality factors ($>10^5$) are used, and small vibrations other than thermal vibrations (e.g. due to floor vibrations or sound) are amplified Q times, these vibrations become stronger than the thermal vibration amplitude. In this case the thermal peak is no more a thermal peak and the thermal method cannot be applied [4].

The sensor displacement noise $N_{\mathrm{z,sens}}$ can also be determined using the thermal method. When the noise floor of the spectral deflection noise density (the quantity acquired by the thermal method) is measured somewhat off the resonance, i.e. not influenced by the resonance, it corresponds to the sensor displacement noise $N_{\mathrm{z,sens}}$.

If the FM detection mode is used, the experimentally measured frequency noise density of the Δf signal after FM demodulation, can be used in order to determine one or two of the following parameters: the sensitivity factor S_{Sensor}, the sensor displacement noise $N_{\mathrm{z,sens}}$, or the spring constant k, as outlined in Sect. 17.6 and [5].

17.8 Comparison to Noise in STM

In the following, we derive the fundamental thermal noise present in STM in order to compare it to the previously considered noise in atomic force microscopy. In (5.28) we have seen that the fundamental limit for the detection of a (tunneling) current using a transimpedance amplifier is the Johnson noise in the feedback resistor, which was written as

$$\Delta I = \sqrt{4k_{\mathrm{B}}TB/R}. \tag{17.29}$$

For a $100\,\mathrm{M\Omega}$ resistor and a bandwidth of 3 kHz, a (RMS) noise current of $\Delta I = 0.3\,\mathrm{pA}$ results. This fundamental noise limit for the measurement of the tunneling current transfers to a noise in the tip-sample distance (i.e. the vertical distance) via the dependence of the tunneling current on the tip-sample distance $I(z) \propto e^{-2\kappa z}$. The slope of the $I(z)$ curve at the working point $I_0(z_0)$ converts the noise in the current into a z-noise via

$$\Delta z = \frac{\Delta I}{|\mathrm{d}I/\mathrm{d}z|}. \tag{17.30}$$

Assuming a tunneling current of $I_0 = 0.1\,\mathrm{nA}$ at the working point and $\kappa = 0.1\,\mathrm{\AA}^{-1}$, the slope of the $I(z)$ curve results as $\mathrm{d}I/\mathrm{d}z = -2\kappa I_0$. This leads to a vertical noise of 0.15 pm, which is much smaller than the resolution required even in order to resolve an atomic corrugation. Moreover, according to (17.29) the vertical noise scales with the square root of the bandwidth $\Delta z \propto \Delta I \propto \sqrt{B}$. This weaker increase of the noise with the measurement bandwidth than the $B^{3/2}$ dependence found for the FM detection in atomic force microscopy allows to work with a larger bandwidth in STM compared to FM detection in AFM.

17.9 Signal-to-Noise Ratio in Atomic Force Microscopy FM Detection

Up to now we have considered the noise in AFM under different circumstances, however, the actual figure of merit is the signal-to-noise ratio. In the following we will discuss the signal-to-noise ratio for the case of the FM detection method in AFM. In this case, the signal-to-noise ratio is the frequency shift due to the tip-sample force gradient Δf divided by the corresponding noise. Specifically we will analyze this signal-to-noise ratio as a function of the oscillation amplitude and find the cantilever oscillation amplitude at which the signal-to-noise ratio is largest [6]. In order to perform this analysis we have to use a certain model for the tip-sample interaction. We assume a repulsive force, which is described by an exponential distance dependence with a range λ as

$$F(u) = F_0 e^{-u/\lambda}. \tag{17.31}$$

Now we evaluate the signal, i.e. the frequency shift in FM detection for the two limiting cases that the cantilever oscillation amplitude is either much larger than the interaction length λ, or much smaller. The following equations were derived under the condition that the minimum tip-sample distance at the lower turnaround point of the oscillation is kept constant when the amplitude is varied. In the limit that the oscillation amplitude is large compared to the interaction range, the frequency shift can (according to (16.19) and (16.24)) be expressed as

$$\frac{\Delta f}{f_0} = \frac{1}{\sqrt{2\pi}} \frac{F_0\sqrt{\lambda}}{kA^{3/2}}. \tag{17.32}$$

This means that for large amplitudes the frequency shift signal depends on the amplitude proportional to $A^{-3/2}$, as shown in Fig. 17.3.

In the opposite limit that the oscillation amplitude is much smaller than the interaction range, the frequency shift has been found proportional to the effective spring constant of the tip-sample interaction (13.11). This can be evaluated further using the force law in (17.31) as

$$\frac{\Delta f}{f_0} = \frac{k_{\mathrm{ts}}}{2k} = \frac{-F'}{2k} = \frac{F}{2k\lambda}. \tag{17.33}$$

This means there is no dependence of the frequency shift on the oscillation amplitude, which corresponds to the horizontal line for the frequency shift signal in Fig. 17.3

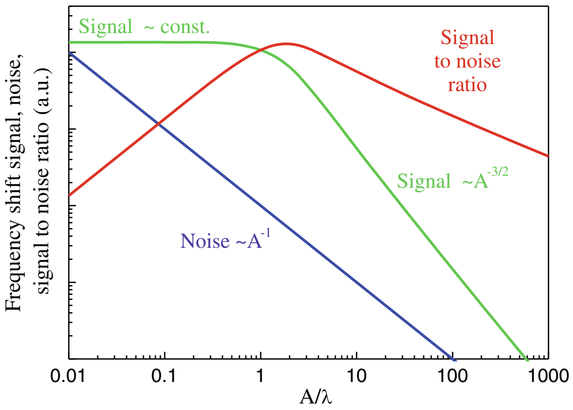


Fig. 17.3 The frequency shift signal, the corresponding noise and the signal-to-noise ratio in FM detection are shown as a function of the cantilever (sensor) oscillation amplitude A, which is normalized to the tip-sample interaction length λ. (Adapted from [6])

for small amplitudes. If the amplitude is close to the interaction length, there is a smooth transition between the limiting cases for small and large amplitudes as shown in Fig. 17.3.

As to the noise, we have seen in the previous section that both thermal noise and detector noise scale as $1/A$ with the amplitude given in (17.18) and (17.25). In Fig. 17.3 also the resulting signal-to-noise ratio is plotted. For amplitudes smaller than λ the signal is constant, while the noise decreases as $1/A$. Thus, the signal-to-noise ratio increases proportional to A for small amplitudes. For large amplitudes the amplitude dependences of signal and noise combine to $S/N \sim A^{-3/2}A \sim 1/\sqrt{A}$, which leads to a decrease of the signal-to-noise ratio for larger amplitudes. A maximum in the signal-to-noise ratio arises if the amplitude corresponds to the range of the interaction λ. These considerations show that the use of oscillation amplitudes in the order of the interaction length lead to the highest signal-to-noise ratio. Thus, if the aim is to use short-range interactions for high-resolution imaging, the oscillation amplitude should be small, possibly less than an ångström.

It is also interesting to compare the frequency shift signal of a short-range interaction to the signal of an interaction with a longer range for small and large values of the oscillation amplitude. In the following, we assume a short-range interaction with $\lambda^{\text{short}} = 0.1\,\text{nm}$ and an interaction with a range of $\lambda^{\text{long}} = 5\,\text{nm}$. If we consider the limiting case $A > \lambda^{\text{long}}$, using (17.32) we find that the signal of the long-range interaction is seven times larger than the signal of the short-range force (ratio of the square roots of the interaction length). However, in the limit of small amplitudes $A < \lambda^{\text{short}}$ the signal of the short-range interaction is, according to (17.33), 50 times larger than that of the long-range interaction, with the other parameters kept the same. This means for a large oscillation amplitude that the signal from a long-range force dominates, while for a small oscillation amplitude the signal comes predominantly from the short-range interactions.

17.10 Summary

- The fundamental limit for the deflection noise of a cantilever arises due to its thermal excitation. The thermal noise depends on the white noise excitation and on the transfer function of the cantilever, which peaks at the resonance frequency. At usual measurement conditions the thermal noise is not the limiting source of noise.
- Another independent contribution to the noise of the cantilever is the electrical noise of the sensor which measures the cantilever deflection.
- In FM detection, the sensor noise depends on the measurement bandwidth $\propto B^{3/2}$. This quite strong increase of the sensor noise with the bandwidth limits the measurement bandwidth in FM detection.
- The signal-to-noise ratio in FM detection is largest for amplitudes corresponding to the range of the interaction force.

References

1. K. Kobayashi, H. Yamada, K. Matsushige, Reduction of frequency noise and frequency shift by phase shifting elements in frequency modulation atomic force microscopy. Rev. Sci. Instrum. **82**, 033702 (2011). https://doi.org/10.1063/1.3557416
2. F.J. Giessibl, F. Pielmeier, T. Eguchi, T. An, Y. Hasegawa, Comparison of force sensors for atomic force microscopy based on quartz tuning forks and length-extensional resonators. Phys. Rev. B **84**, 125409 (2011). https://doi.org/10.1103/PhysRevB.84.125409
3. G.H. Simon, M. Heyde, H.-P. Rust, Recipes for cantilever parameter determination in dynamic force spectroscopy: spring constant and amplitude. Nanotechnology **18**, 255503 (2007). https://doi.org/10.1088/0957-4484/18/25/255503
4. J. Welker, F. de Faria Elsner, F.J. Giessibl, Application of the equipartition theorem to the thermal excitation of quartz tuning forks. Appl. Phys. Lett. **99**, 084102 (2011), https://doi.org/10.1063/1.3627184
5. J. Lübbe, M. Temmen, P. Rahe, A. Kühnle, M. Reichling, Determining cantilever stiffness from thermal noise. Beilstein J. Nanotechnol. **4**, 227 (2013). https://doi.org/10.3762/bjnano.4.23
6. S. Morita, F.J. Giessibl, R. Wiesendanger, (eds.), *Non-contact Atomic Force Microscopy*, vol. 2 (Springer, Heidelberg, 2009), https://doi.org/10.1007/978-3-642-01495-6

Chapter 18
Quartz Sensors in Atomic Force Microscopy

As an alternative to the most frequently used silicon cantilevers, quartz oscillators can be used as sensors in AFM. It is possible to obtain atomic resolution in FM atomic force microscopy using quartz sensors. These quartz sensors are characterized by a large spring constant ($>$1,000 N/m). Both quartz tuning forks, which are used in wristwatches, as well as quartz needle oscillators can be used as sensors in AFM. An advantage of using quartz sensors is that the detection of the oscillation signal can be performed completely electrically, without any optical elements, like a laser diode, a lens, a fiber, or a photodiode being needed. This simplifies the experimental setup.

18.1 Tuning Fork Quartz Sensor

One example of a quartz sensor is the quartz tuning fork, frequently used in wristwatches. In Fig. 18.1 a tuning fork quartz oscillator is shown without housing. The whole tuning fork has a length of 4 mm, and the prongs have a length of 2.4 mm. The resonance frequency of such a tuning fork is usually 32,768 Hz, which is related to its use in watches. The bending mode of such a tuning fork is like that known from a macroscopic tuning fork with the two prongs with a 180° phase difference (e.g. against each other). Since the tuning fork has no sharp tip at its end a (tungsten) tip is usually attached at the end of the prong. If a tip is fixed to one prong only, an asymmetry between the two prongs is induced which reduces the Q-factor substantially. In order to prevent this, the other prong can be fixed to a holder with high mass. This configuration is called qPlus configuration [1].

The excitation of the tuning fork is usually achieved mechanically by applying an AC voltage to a piezoelectric actuator exciting the sensor. The tuning fork is excited at its lowest resonance frequency, which leads to a bending of the sensor prong. Since single crystal quartz is a piezoelectric material, the detection of the

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3_18

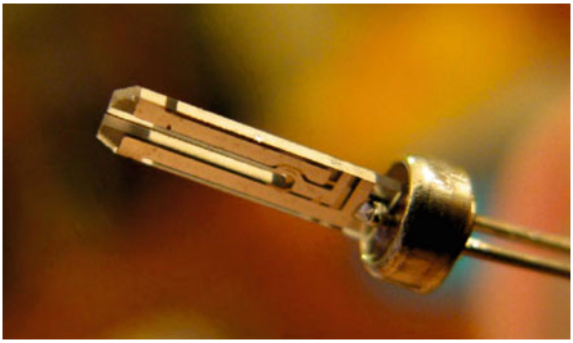


Fig. 18.1 A tuning fork quartz oscillator as used in wristwatches. The tuning fork oscillator can be used as a force sensor in AFM

bending oscillation of the prong of the tuning fork is performed electrically using the piezoelectric effect. A bending of the prong induces a voltage between the metal electrodes at the prong. Simultaneous STM operation can be achieved by attaching a wire to the tip, which guides the tunneling current to a preamplifier.

18.2 Quartz Needle Sensor

Another type of quartz crystal oscillator which can also be used as a force sensor in atomic force microscopy is shown in Fig. 18.2. This sensor is known as a "needle sensor" and is characterized by its small size (needle length 1.3 mm), an extensional oscillation of the quartz needle, a high resonance frequency ($\sim$1 MHz) and a high force constant ($\sim$1 MN/m). The needle has two Au electrodes as shown in Fig. 18.2, which allows for an electrical excitation without any additional driving piezo by applying the AC driving voltage to one of the two electrodes. This induces an oscillation of the needle along its axis via the (inverse) piezoelectric effect. An electrical detection also can be obtained due to the piezoelectric effect. The oscillating needle induces a voltage on the second electrode by the piezoelectric effect, which is amplified by a preamplifier and processed further using the FM detection scheme, as described previously. A sharp tip has to be attached to the top of the quartz needle. This can be a thin tungsten tip, as shown in Fig. 18.3a. Another way of attaching a tip to the needle sensor is to glue a Si cantilever to the top of the needle and to break the cantilever base off, as shown in Fig. 18.3b. If the attached tip plus glue mass is small, high Q-factors $>$10,000 can be achieved even in air.

A schematic of the control electronics of the needle sensor in which the needle sensor can be operated in the force detection mode (AFM) mode, or alternatively in the tunneling mode (STM) is shown in Fig. 18.4. In the tunneling mode (TCF = tunneling current feedback), in which the needle sensor can still oscillate, a DC tunneling bias voltage V_{bias} is added to the AC signal driving the needle oscillation. The tip is electrically connected to the needle electrode to which the DC bias is applied. The resulting tunneling current (averaged over one oscillation cycle) is

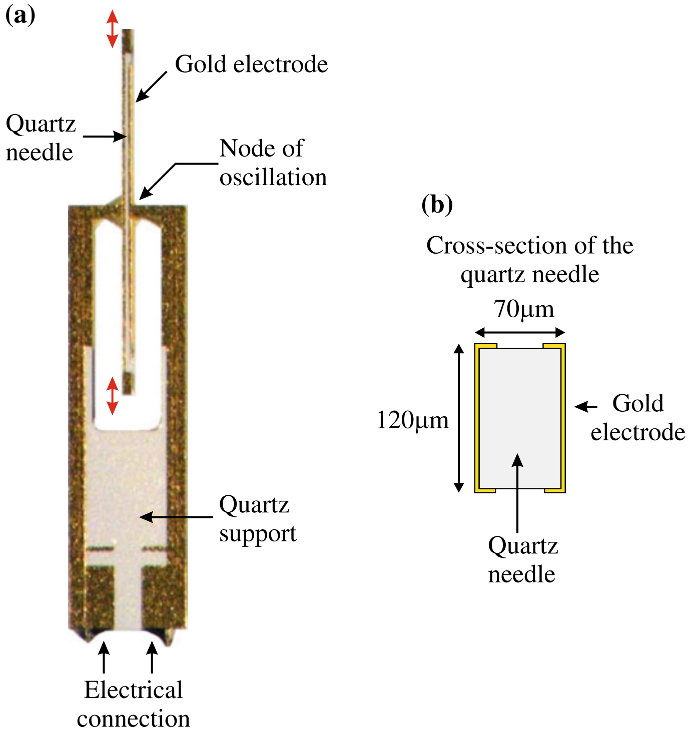


Fig. 18.2 Photo of a needle sensor (**a**) and schematic cross section through the needle (**b**). The needle sensor is an extensional type quartz oscillator which can be used as force sensor in AFM

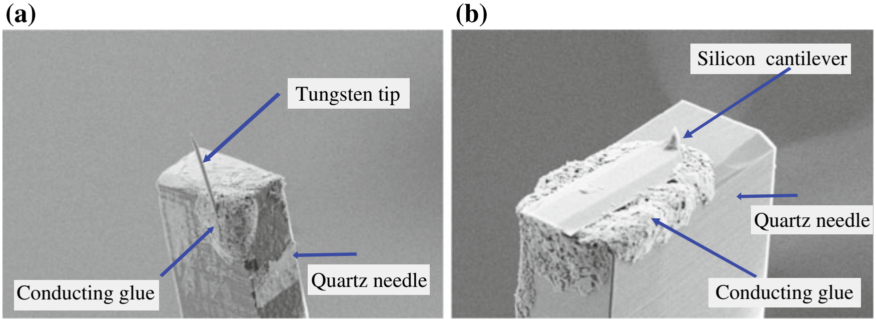


Fig. 18.3 Tips glued to the top of a needle sensor. **a** Electrochemically etched tungsten tip. **b** End part of a silicon cantilever

measured at the sample and used as a feedback signal for control of the tip-sample distance. This mode of operation is called tunneling current feedback mode (TCF). The frequency shift of the oscillating needle sensor can be recorded in parallel (as a free signal), however, it is not used for feedback.

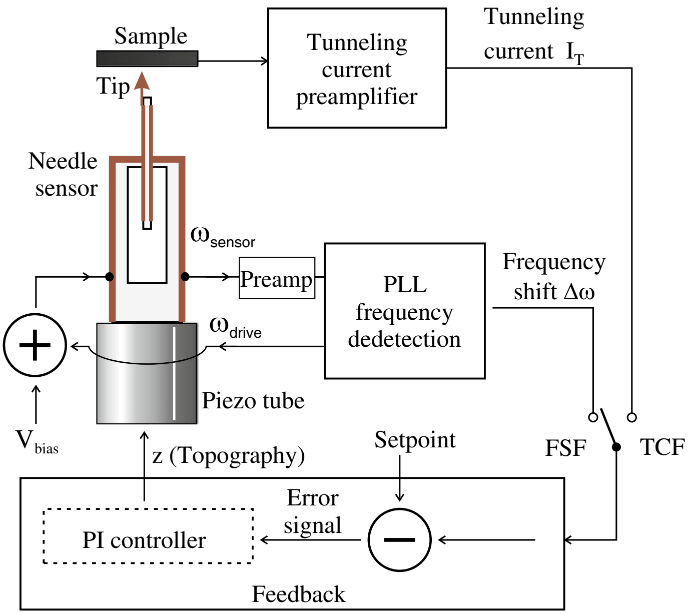


Fig. 18.4 Schematic circuit for driving the needle sensor as a force sensor in AFM. Alternatively the frequency shift signal (FSF) or the tunneling current (TCF) can be used as feedback signals [2]

Table 18.1 Comparison of the properties of the different AFM force sensors: silicon cantilever, quartz tuning fork and quartz needle sensor

	Cantilever	Tuning fork	Needle sensor
Spring constant (N/m)	1–50	1–20 k	600 k–1 M
Resonant frequency f_0 (kHz)	100–300	20–100	600–2,000
Quality factor Q	100–2 k	1–20 k	5–200 k
Frequency shift[a] Δf (Hz)	50	75	5
Min. amplitude[b] A_{min} (Å)	4	0.05	0.0002

[a] For a force gradient of 10 nN/nm the frequency shift is $\Delta f = -\frac{f_0}{2k} 10\,\text{nN/nm}$

[b] The minimum amplitude before snap to contact for a force of 10 nN is given by the condition $10\,\text{nN} < kA_{\text{min}}$

If the needle sensor is employed in the AFM mode with the FM detection scheme, the frequency shift signal is used for the z-feedback (FSF = frequency shift feedback). Additionally, the tunneling current can be recorded simultaneously as a free signal. In this way it is possible to combine atomic force microscopy and scanning tunneling microscopy.

In Table 18.1 typical properties of three types of AFM force sensors are compared: silicon cantilever, tuning fork and needle sensor [3]. The spring constant increases strongly from cantilever to tuning fork and the needle sensor. This is due to the larger dimensions of the tuning fork compared to the micro machined cantilevers. The high stiffness of the needle sensor is induced by its extensional vibration geometry (the axial extension of a bar is a hard spring). Also the quality factor (in air) increases in the order from cantilever via tuning fork to the needle sensor. For the cantilever sensors, the quality factor is low due to damping in air. The frequency shift for a force gradient of 10 nN/nm, as an example, is smallest for the needle sensor. Due to the higher force constant, the tuning fork and the needle sensor can be operated at close tip sample distances without the problem of snap-to-contact occurring.

18.3 Determination of the Sensitivity of Quartz Sensors

The mechanical oscillation amplitude A_{sensor} is related to the measured sensor voltage V_{sensor} (measured at the output of the preamplifier measuring the sensor signal) by the sensitivity factor as

$$A_{\mathrm{sensor}} = S_{\mathrm{sensor}} V_{\mathrm{sensor}}. \tag{18.1}$$

In the calibration procedure, the sensitivity factor (S_{sensor} in nm per volt) is determined which converts V_{sensor} to an oscillation amplitude A_{sensor} in nm. The voltage V_{sensor} and thus also S_{sensor} depend on the specific devices used to measure the amplitude voltage, e.g. the gain factors of the amplifiers enter into these quantities.

For silicon cantilevers the cantilever sensitivity was determined for instance via the force-distance curve, as described in Sect. 11.6. For the case of quartz sensors, this method cannot be applied due to the very high force constants of these sensors, which is in the same order as that of hard samples (the tip would be destroyed).

We assume that FM detection is used and the frequency shift is measured. In Fig. 18.5 we compare two cases of different oscillation amplitudes A_{sensor} and A'_{sensor}. When the tip is brought close to the surface and a certain frequency shift setpoint Δf is set, this will result in different values for the average tip-sample position of the cantilever d, for different oscillation amplitudes A_{sensor} and A'_{sensor}, as shown in

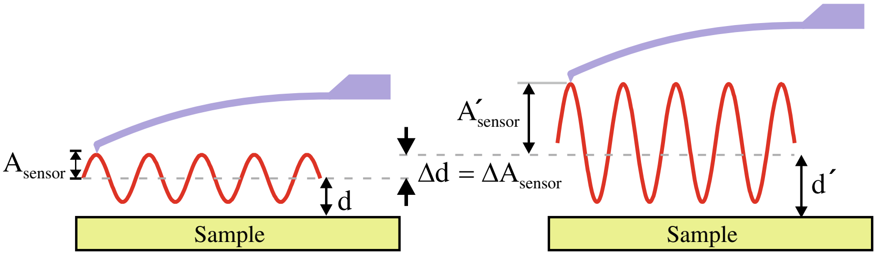


Fig. 18.5 Principle of the determination of the oscillation amplitude used for quartz sensors. The distance between the lower turnaround point of the tip oscillation and the sample is approximately the same for different oscillation amplitudes. Thus, the change of the sensor amplitude $\Delta A_{\mathrm{sensor}}$ is equal to the retraction of the equilibrium position of the tip Δd. A cantilever sensor is shown schematically instead of a quartz sensor

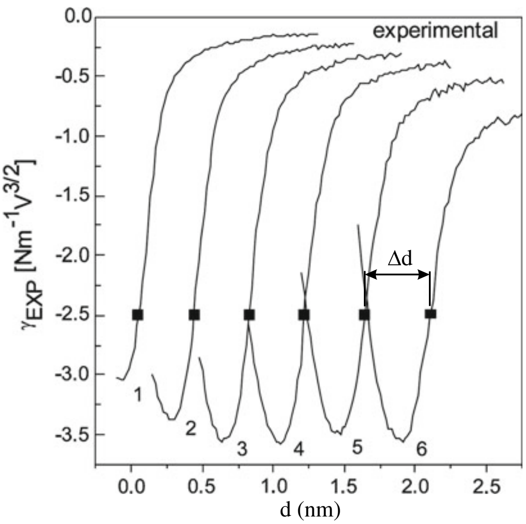


Fig. 18.6 Normalized frequency shift as a function of the average tip-sample position d, for different cantilever oscillation amplitudes. The sensor amplitude V_{sensor} increases from *curve 1* to *curve 6* from 0.2 to 0.7 V, respectively. Each increase of the amplitude by 0.1 V leads to a shift of the curves by $\Delta d = 0.5\,\text{nm}$, showing the proportionality $\Delta d \propto \Delta V_{\text{sensor}}$ (Reproduced with permission from [4])

Fig. 18.5. Since the main contribution to the frequency shift signal comes from the lower turnaround point of the oscillation (as shown in Chap. 16), the distance from the lower turnaround point to the sample is approximately the same in both cases, independent of the oscillation amplitude. Due to this the tip retraction Δd is equal to the amplitude change $\Delta d = \Delta A_{\text{sensor}} = A'_{\text{sensor}} - A_{\text{sensor}}$. By measuring Δd for the sensor voltage difference ΔV_{sensor} the sensitivity can be determined as

$$S_{\text{sensor}} = \frac{\Delta A_{\text{sensor}}}{\Delta V_{\text{sensor}}} = \frac{\Delta d}{\Delta V_{\text{sensor}}}. \qquad (18.2)$$

In this calibration procedure for the sensitivity the tip-sample interaction is kept constant (e.g. by keeping the frequency shift at a constant value), while A_{sensor} is varied. In a practical implementation of this method the normalized frequency shift (introduced in (16.20)) is measured as a function of the tip-sample distance d [4]. The measured frequency shift curves have the usual (Lennard-Jones-type) shape, as shown in Fig. 18.6. With increasing oscillation amplitudes, curves 1–6 are measured. Since the normalized frequency shift is plotted, all curves have approximately the same magnitude (as already shown in Fig. 16.3). However, they have a mutual shift: the larger the oscillation amplitude, the more the curves shift to larger tip-sample distances d, as also shown in principle in Fig. 18.5. The mutual shift (for a voltage increase ΔV_{sensor} of 0.1 V) amounts to about $\Delta d = 0.5\,\text{nm}$ as indicated in Fig. 18.6. A proportionality between these quantities is observed as $\Delta d \propto \Delta V_{\text{sensor}}$, with a proportionality factor of 0.5 nm/0.1 V. Thus, the sensitivity factor $S_{\text{sensor}} = 5\,\text{nm/V}$ can be obtained from the relation

$$\Delta A_{\text{sensor}} = \Delta d = S_{\text{sensor}} \Delta V_{\text{sensor}}. \qquad (18.3)$$

18.4 Summary

- Quartz sensors are used in AFM since they allow for completely electrical detection (and sometimes also excitation) via the piezoelectric effect, which simplifies the experimental setup.
- The two types of quartz sensors used most frequently are the tuning fork sensor and the needle sensor.
- A sharp tip has to be attached to the quartz oscillators for the use in AFM.
- The sensitivity of a quartz sensor can be determined experimentally by comparing the frequency shift versus distance curves for different oscillation amplitudes.

References

1. S. Morita, F.J. Giessibl, R. Wiesendanger (eds.), *Non-contact Atomic Force Microscopy*, vol. 2 (Springer, Heidelberg, 2009). https://doi.org/10.1007/978-3-642-01495-6
2. I. Morawski, B. Voigtländer, Simultaneously measured signals in scanning probe microscopy with a needle sensor: frequency shift and tunneling current. Rev. Sci. Instrum. **81**, 033703 (2010). https://doi.org/10.1063/1.3321437
3. F.J. Giessibl, F. Pielmeier, T. Eguchi, T. An, Y. Hasegawa, Comparison of force sensors for atomic force microscopy based on quartz tuning forks and length-extensional resonators. Phys. Rev. B **84**, 125409 (2011). https://doi.org/10.1103/PhysRevB.84.125409
4. G.H. Simon, M. Heyde, H.-P. Rust, Recipes for cantilever parameter determination in dynamic force spectroscopy: spring constant and amplitude. Nanotechnology **18**, 255503 (2007). https://doi.org/10.1088/0957-4484/18/25/255503

Appendix A
Horizontal Piezo Constant for a Tube Piezo Element

Here we will derive a more exact expression for the length extension ΔL of a bent piezo tube than the one used in (3.12). Using this expression for ΔL results in the equation for the horizontal piezo constant given in (3.13).

Before we come to the bending of a tube piezo element, we introduce the relevant concept for a very simple case. Let us assume the ceramic of the piezo tube is an elastic medium and we pull with a force (or force per area σ) at the end of the piezo tube as shown in Fig. A.1a. As a response to this externally applied stress, a strain ΔL develops which leads to a stress $\tau = E\Delta L/L$ in the opposite direction. In equilibrium σ and τ have the same value and opposite direction. Instead of pulling at the piezo tube, we can exert an elastic stress on the piezo tube also via the piezoelectric effect. The extension of the piezo element is (according to Hooke's law and (3.3)) accomplished by a stress $\sigma = Ed_{31}V/h$ (with h being the wall thickness), which is counterbalanced by the stress build-up in the elastic medium $\tau = E\Delta L/L$. Here due to the simple geometry the stresses have the same value throughout the cross section of the tube and counterbalance locally. This is different for the case of the bending of a segmented piezo tube. At this point, the stress σ resulting in an extension by the piezoelectric effect does not occur homogeneously, but only at the segments to which a voltage is applied. The elastic stress τ is also inhomogeneous, since the elastic strain which develops due to the bending of a piezo tube is also inhomogeneous throughout the tube cross section. In the following, we will discuss the geometry of bending, the stresses σ and τ, and the equilibrium condition in detail following the arguments given in [1].

We consider a piezo tube with voltages $+V_x$ and $-V_x$ applied to the x-electrodes, while the voltage at the other electrodes of the tube is zero. The geometry of bending of a tube piezo is shown in Fig. A.1b. As shown in (3.7), the bending angle can be written as $\alpha = 2\Delta L/D_m$, with D_m being the average diameter of the piezo tube (the wall thickness is considered as negligibly small), and ΔL being the length extension at the middle of the x-electrodes. In the following we will determine this length extension ΔL.

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3

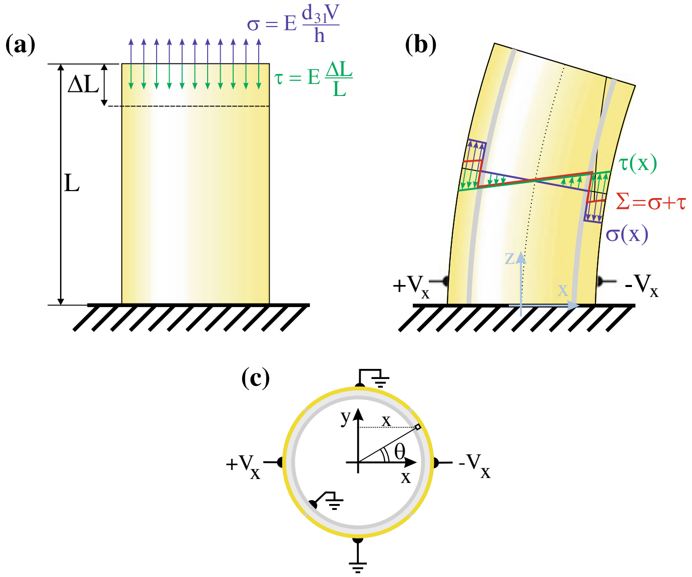


Fig. A.1 **a** When pulled with an external stress σ at an elastic object (piezo tube) the object extends by ΔL and an inner stress τ builds up as a response. In equilibrium the two stresses compensate each other. **b** In the case of a bending of the tube due to voltages on the x-electrodes, the externally applied stress σ is only different from zero at those electrodes (*blue arrows*), while the reaction stress in the elastic body τ is linear as a function of x. Thus, the stresses do not compensate locally as in **a**. However, in equilibrium the total torque has to vanish. **c** Cross section of the piezo tube with the applied voltages

A voltage V_x applied to the x-electrodes induces a stress σ which is homogeneous throughout the electrode, as sketched in Fig. A.1b. At the y-electrodes no external stress occurs, since no voltage is applied to those electrodes. This applied inhomogeneous stress distribution throughout any cross section through the piezo tube causes an elastic reaction (bending) of the tube, which results in a reaction stress τ in the piezo tube material. The strain is zero in the middle of the y-electrodes and is assumed to increase linearly along the bending direction x as shown in Fig. A.1b, while it is constant along the y-direction perpendicular to the bending. The corresponding stress τ also increases linearly with x and is shown in Fig. A.1b. We see that σ and τ do not have the same values at each point as for the vertical stretching along the z-axis (Fig. A.1a), but have different values across the piezo tube.

The sum $\sigma + \tau$ is also sketched in Fig. A.1b. What is now the equilibrium condition? Let us consider the cross-section of the tube in Fig. A.1b as a lever rotating about the center, on which the sum of the stresses $\Sigma(x) = \sigma(x) + \tau(x)$ is applied at each point of the tube cross section. The equilibrium condition is now, as for a lever, that the sum of all torques $\Sigma \cdot x$ applied to the lever has to vanish. The piezo extension induces a local torque $\sigma(x) \cdot x$ and the elastic response induces a local torque $\tau(x) \cdot x$. The bending of the tube is in equilibrium if the integral of the total torque

$\Sigma \cdot x$ over the whole cross section of the piezo tube vanishes. Now we perform this integration.

Due to the symmetry of the problem, we limit the integration to the first quadrant (Fig. A.1c). For the integration over the y-electrode ($45^\circ < \theta < 90^\circ$), σ is zero and

$$\Sigma(\theta) = \tau(\theta) = \tau_{\max} \cos\theta, \tag{A.1}$$

where the variable x has been replaced by $\cos\theta$ and $\tau_{\max}$ is the stress in the middle of the x-electrode. For the integration over the x- electrode ($0^\circ < \theta < 45^\circ$), the total stress can be written as

$$\Sigma(\theta) = \sigma(\theta) + \tau(\theta) = \tau_{\max} \cos\theta - \sigma_{\max}, \tag{A.2}$$

where $\sigma_{\max}$ is the stress applied to the x-electrodes due to the applied voltages. With this the equilibrium condition, i.e. the vanishing of the integral of the torque over the tube quadrant, reads as

$$\begin{aligned} &\int_0^{90^\circ} \Sigma(\theta) \cos\theta \mathrm{d}\theta \\ &= \int_0^{45^\circ} (\tau_{\max} \cos\theta - \sigma_{\max}) \cos\theta \mathrm{d}\theta + \int_{45^\circ}^{90^\circ} \tau_{\max} \cos\theta \cos\theta \mathrm{d}\theta = 0. \end{aligned} \tag{A.3}$$

The evaluation of these integrals leads to the equilibrium condition

$$\tau_{\max} = \frac{2\sqrt{2}}{\pi} \sigma_{\max}. \tag{A.4}$$

Replacing $\sigma_{\max} = Ed_{31}\Delta V/h$ and $\tau_{\max} = E\Delta L/L$, results in

$$\Delta L = \frac{2\sqrt{2}}{\pi} \frac{d_{31} L \Delta V}{h}. \tag{A.5}$$

This result for the extension ΔL is smaller by a factor of about 0.9 than that for the case where a "free" extension of the x-electrodes is considered (3.12), i.e. without any "hindrance" by the straining of the y-electrodes. In this way, (3.13) finally results for the horizontal piezo constant.

Reference

1. C.J. Chen, *Introduction to Scanning Tunneling Microscopy*, 2nd edn. (Oxford University Press, New York, 2008). https://doi.org/10.1093/acprof:oso/9780199211500.001.0001. ISBN: 9780199211500

Appendix B
Spectral Density, Spectrum and their Experimental Calibration

The spectral density of a signal can be measured with a spectrum analyzer. Nowadays, stand alone spectrum analyzer instruments (with all the calibration steps already included) are less frequently used in favor of analogue to digital conversion of the measured signal followed by a subsequent software discrete Fourier transform (DFT), the correct calibration is no more "included" by the spectrum analyzer hardware. Since the discrete Fourier transform just transforms n numbers to n new numbers, the user has to take care about the necessary calibration steps. While this is straightforward in principle, it involves a number of non-trivial details. Here, we also include the description of an experimental calibration procedure which gives an easy cross-check for the correct calibration of the spectrum or spectral density.

If a continuous signal $S(t)$ is sampled with a sampling frequency f_{sample}, this signal is represented as a discrete time series $S(k/f_{\text{sample}})$. The discrete Fourier transform (DFT) of a time series of length n is defined as

$$\hat{S}(m) = \sum_{k=0}^{n-1} S(k/f_{\text{sample}}) e^{-2\pi i k m/n}, \tag{B.1}$$

with $m = 0...n-1$. The power spectral density (termed PSD, or N^2_{PSD}) is proportional to the absolute square of the discrete Fourier transform (DFT) [1]. If we do not consider windowing yet (i.e consider a rectangular window) [2, 3], the PSD results as

$$N^2_{\text{PSD}}(m) = \frac{2}{f_{\text{sample}}\, n} \left|\hat{S}(m)\right|^2, \quad m = 0...n/2. \tag{B.2}$$

Here we consider the single sided PSD in which only $m = 0...n/2$, with n being even are considered (positive frequencies). Note that other definitions of the DFT than the one in (B.1) result in other factors in (B.2) [2]. The spectral density N_{PSD} is the square root of the power spectral density N^2_{PSD}.

For a continuous signal the power spectral density of a signal is related, via Parseval's identity, to the root mean square (RMS) S_{RMS} of the signal as

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3

$$S_{\mathrm{RMS}}^2 = \lim_{T\to\infty} \frac{1}{T}\int_0^T S^2(t)\,\mathrm{d}t \equiv \left\langle S^2(t)\right\rangle = \int_0^\infty N_{\mathrm{PSD}}^2(f)\,\mathrm{d}f. \tag{B.3}$$

If the signal is a discrete time series, the continuous quantities $S(t)$ and $N_{\mathrm{PSD}}(f)$ translate to discrete values as

$$S(t) \leftrightarrow S(k/f_{\mathrm{sample}}) \text{ and } N_{\mathrm{PSD}}(f) \leftrightarrow N_{\mathrm{PSD}}(m\, f_{\mathrm{res}}), \tag{B.4}$$

respectively with $k = 0 ... n - 1$, and $m = 0 ... n/2$. The width of the n frequency bins of the DFT is given by $f_{\mathrm{res}} = f_{\mathrm{sample}}/n$ [2]. For a discrete signal (B.3) translates to

$$\begin{aligned} S_{\mathrm{RMS}}^2 &= \frac{1}{T}\sum_{k=0}^{n-1} S^2(k/f_{\mathrm{sample}})/f_{\mathrm{sample}} \\ &= \sum_{m=0}^{n/2} N_{\mathrm{PSD}}^2(m\, f_{\mathrm{res}})\, f_{\mathrm{res}}. \end{aligned} \tag{B.5}$$

In the following we consider two simple examples for the power spectral density, a constant power spectral density (Fig. B.1a) and a power spectral density of a tonal sinusoidal signal (Fig. B.1b).

Spectral density. If the power spectral density of the signal is considered within a certain frequency bandwidth $B = f_2 - f_1$ between f_1 and f_2 (as indicated by the blue shaded area in Fig. B.1a), the power spectral density is zero outside the range of the bandwidth B. We assume further that $f_{\mathrm{res}} \ll B$, which is usually the case. If the power spectral density is constant for the j bins between f_1 and f_2, the N_{PSD}^2

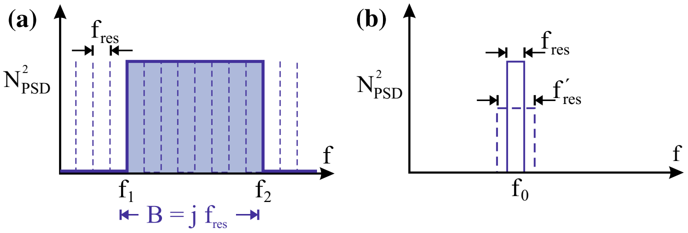


Fig. B.1 **a** Case of a constant power spectral density within B (blue shaded area). The DFT representation of the power spectral density has j (same) values with a frequency bin with of $f_{\mathrm{res}} = f_{\mathrm{sample}}/n$. In this case the power spectral density is independent of the width of the frequency bin of the DFT, f_{res} (c.f. (B.6)), while the power spectrum depends on the bin width (c.f. (B.10)). **b** Power spectral density of a tonal signal (sinusoidal), which has a non vanishing value only in one frequency bin. In this case DFT representation of the power spectral density depends on the frequency bin with f_{res} (c.f. (B.7)), while the power spectrum is independent of the bin width (c.f. (B.9)) in (B.5) can be written in front of the sum and the sum yields $j \cdot f_{\text{res}} = B$. Thus, for a constant PSD (B.5) results in

$$N_{\text{PSD}}^2 = \frac{S_{\text{RMS}}^2}{B}. \tag{B.6}$$

If for example the signal is a voltage, e.g. of RMS amplitude $S_{\text{RMS}} = 1\,\text{V}$, and $B = 100\,\text{Hz}$, a spectral density of $N_{\text{PSD}} = 0.1\,\text{V}/\sqrt{\text{Hz}}$ results. In this case the (power) spectral density is independent of the width of the frequency bins. This case is desirable as the value of the (power) spectral density has a significance independent of the width of the frequency bins, i.e. independent of details of the sampling.

For the case of a tonal (sinusoidal) signal the situation is different. For the sake of simplicity we consider cases without spectral leakage present [3]. Then the tonal signal is usually located within a single non-zero frequency bin of width f_{res} at a frequency f_t, as shown in Fig. B.1b. In this case only one term of the sum in (B.5) survives and the (power) spectral density of this bin depends (undesirably) on the width of the frequency bins f_{res}, as

$$N_{\text{PSD}}^2(f_{\text{res}}) = \frac{S_{\text{RMS}}^2}{f_{\text{res}}}. \tag{B.7}$$

This means that for instance a tonal signal with an RMS amplitude of 1 V results in different values for the (power) spectral density, depending on the width of the frequency bins f_{res}, as also shown in Fig. B.1b for two different values of the frequency bin width f_{res} and f'_{res}, respectively. This dependence of the value of the (power) spectral density on the frequency bin width f_{res}, which depends on the particular length of the time series used for the DFT and the particular sampling rate, is of course undesirable. Thus, the value of the (power) spectral density for a tonal signal has no unique significance without the knowledge of some details on the sampling process, such as the sampling rate f_{sample} and the length n of the DFT.

Spectrum. A different quantity, the power spectrum N_{spec}^2, or the spectrum N_{spec}, defined as

$$N_{\text{spec}}^2 \equiv N_{\text{PSD}}^2 \cdot f_{\text{res}}, \tag{B.8}$$

avoids this disadvantage. When inserting (B.7), valid for a tonal signal, into (B.8), the (power) spectrum of a tonal signal turns out to be independent of the width of the frequency bin, as

$$N_{\text{spec}}^2 = S_{\text{RMS}}^2. \tag{B.9}$$

As (B.9) shows, the value of the spectrum is equal to the RMS amplitude of the tonal (sinusoidal) signal, $N_{\text{spec}} = S_{\text{RMS}}$ (e.g. 1 V).

However, undesirably for a signal of constant (power) spectral density, the spectrum N_{spec} depends on the width of the frequency bin f_{res}, as evident when inserting (B.6) into (B.8), resulting in

$$N_{\mathrm{spec}}^2 = \frac{S_{\mathrm{RMS}}^2}{B} f_{\mathrm{res}}. \tag{B.10}$$

In conclusion, neither the (power) spectral density, nor the (power) spectrum deliver a value which is independent of the frequency bin for a tonal signal *as well as* for signal with constant PSD (representative of a broad band signal with a relatively flat PSD). A solution of this dilemma would be to choose the width of the frequency bin $f_{\mathrm{res}} = 1\,\mathrm{Hz}$, so that both, the spectral density and the spectrum have the same numeric value (but still different units, e.g. $\mathrm{V}/\sqrt{\mathrm{Hz}}$ and V, respectively). However, if low frequencies approaching 1 Hz and below are of interest, a frequency bin with of 1 Hz is too wide.

So far we have not considered the windowing in the DFT [3], which means we have so far implicitly considered a rectangular window function. When applying other window functions in the DFT, a quantity named "normalized equivalent noise bandwidth" (NENBW) can be defined and (B.8) is extended with $f_{\mathrm{res}}^{\mathrm{eff}}$ to

$$N_{\mathrm{spec}}^2 \equiv N_{\mathrm{PSD}}^2 \cdot f_{\mathrm{res}}^{\mathrm{eff}} = N_{\mathrm{PSD}}^2 \cdot f_{\mathrm{res}} \cdot \mathrm{NENBW}. \tag{B.11}$$

In order to present the complete information of a spectral analysis, both the (power) spectral density, as well as the (power) spectrum have to be presented, or one of them and $f_{\mathrm{res}}^{\mathrm{eff}}$.

In the signal processing from the time series of the signal to the spectral density or spectrum several proportionality factors are involved due to the use of, for example, either RMS amplitude or peak amplitude, either two-sided spectrum or single-sided spectrum, either natural frequency PSD or angular frequency PSD, or due to different window types, etc. So one has to consider all these factors carefully. Complementary also an experimental calibration of spectral density or spectrum is very desirable and will be considered in the following.

A tonal signal e.g. from a signal generator can be used to calibrate the spectrum or the spectral density. According to (B.9) the RMS signal amplitude of a tonal signal corresponds directly to the amplitude of the spectrum N_{spec}, independent of the width of a frequency bin. For the calibration of the power spectral density using a tonal signal, the effective width of a frequency bin enters. According to (B.7) (extended to $f_{\mathrm{res}}^{\mathrm{eff}}$), the RMS signal amplitude of a tonal signal S_{RMS}^2 has to be divided by $f_{\mathrm{res}}^{\mathrm{eff}}$ in order to obtain the power spectral density.

Alternatively to a calibration with a tonal signal a signal of constant power spectral density (white noise) and known amplitude can be used for the calibration. Such a signal is provided for instance by the Johnson-Nyquist noise (thermal noise) of a resistor R as as voltage source of known RMS voltage

$$U_{\mathrm{JN}}^{\mathrm{RMS}} = \sqrt{4\,k_{\mathrm{B}}\,T\;R\;B}, \tag{B.12}$$

and constant PSD (white noise spectrum). The bandwidth B in (B.12) corresponds to the effective width of a frequency bin of the DFT as $B = f_{\mathrm{res}}^{\mathrm{eff}}$. In order to obtain a reasonably large voltage, a resistor of large resistance should be used ($R \geq 500\,\mathrm{k}\Omega$) and considered in parallel with the input resistance of the measurement device.

In conclusion, if a case of a spectral analysis includes tonal peaks, as well as (locally) constant regions as function of frequency, both the spectrum, and the spectral density are required in order to deliver quantitative results for tonal peaks and broad band regions. The tonal peaks are represented quantitatively in the spectrum (e.g. in volts), while constant regions are represented quantitatively in the spectral density (e.g. as $\mathrm{V}/\sqrt{\mathrm{Hz}}$). Spectrum and spectral density can be converted into each other by the proportionality factor $\sqrt{f_{\mathrm{res}}^{\mathrm{eff}}} = \sqrt{(f_{\mathrm{sample}}/n) \cdot \mathrm{NENBW}}$.

References

1. C.W. de Silva, *Vibration: Fundamentals and Practice*, 2nd edn. (Taylor and Francis – CRC Press, London, 2006). ISBN 9780849319877
2. G. Heinzel, A. Rüdiger, R. Schilling, Spectrum and spectral density estimation by the Discrete Fourier Transform (DFT), including a comprehensive list of window functions and some new at-top windows (2002). http://pubman.mpdl.mpg.de/pubman/item/escidoc:152164:1/component/escidoc:152163/395068.pdf
3. R.G. Lyons, *Understanding Digital Signal Processing*, 3rd edn. (Pearson Education, 2011). ISBN: 8131764362

Appendix C
Corrections to the Thermal Method

In the following, we will present several corrections (going beyond the approximation of the cantilever as a simple harmonic oscillator) which have to be applied for a more exact determination of the force constant by the thermal method. We consider the most important case of rectangular cantilevers.

For an ideal harmonic oscillator represented in Fig. C.1a by a mass and a spring, the expression (11.21) holds. However, a real rectangular cantilever beam (Fig. C.1b) also has higher modes of oscillation. The first four modes of a free cantilever beam are shown in Fig. C.1c. For each higher mode one more node appears in the shape of the vibration modes. Each mode can be considered as a harmonic oscillator for which the equipartition theorem holds, i.e. each mode is thermally excited by $k_{\mathrm{B}}T$. Thus, in analogy to the ideal harmonic oscillator a (dynamic) spring constant k_i of the mode i can be defined by the relation

$$\frac{1}{2}k_i \left\langle \Delta z_{\mathrm{th},i}^2 \right\rangle = \frac{1}{2}k_{\mathrm{B}}T, \tag{C.1}$$

with $\left\langle \Delta z_{\mathrm{th},i}^2 \right\rangle$ being the mean square deflection arising from the ith mode. This mean square deflection can be calculated [1] as

$$\left\langle \Delta z_{\mathrm{th},i}^2 \right\rangle = \frac{k_{\mathrm{B}}T}{k}\frac{12}{\alpha_i^4} = \frac{k_{\mathrm{B}}T}{k_i}, \tag{C.2}$$

with the values of α_i ($\alpha_1 = 1.88$, $\alpha_2 = 4.69$, and $\alpha_3 = 7.85$) and correspondingly the (dynamic) spring constant k_i for each mode given in [1]. The spring constant for the first mode has been calculated as $k_1 = k/0.971$. While each mode is excited with the thermal energy $k_{\mathrm{B}}T$, the spring constants for the higher modes increase significantly. Thus the thermally excited deflection for higher modes becomes very small. Since the thermal excitation of the different modes are independent, the total mean square thermal amplitude is the sum over the mean square amplitudes of all

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3

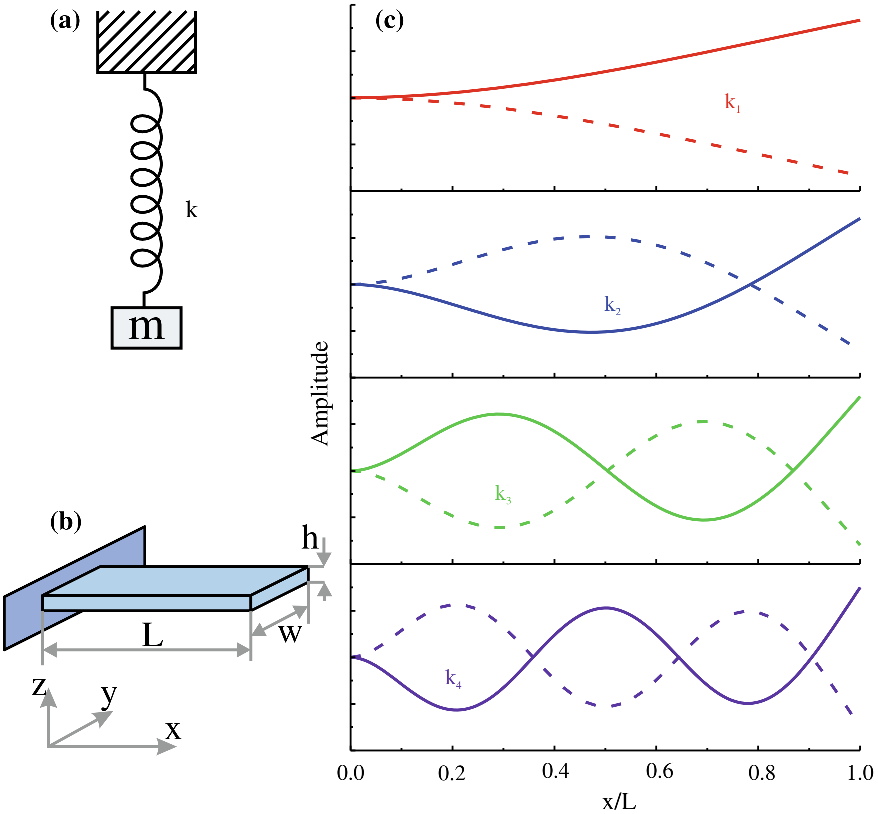


Fig. C.1 **a** Ideal one-dimensional harmonic oscillator represented by a mass m on a spring with spring constant k. **b** Sketch of a cantilever-type beam. **c** The first four modes of a rectangular cantilever. A (dynamic) spring constant k_i can be assigned to each mode

modes[1] $\left\langle \Delta z_{\mathrm{th}}^2 \right\rangle = \sum_0^\infty \left\langle \Delta z_{\mathrm{th},i}^2 \right\rangle$. It has been calculated that $\sum_0^\infty k_i \left\langle \Delta z_{\mathrm{th},i}^2 \right\rangle = k \left\langle \Delta z_{\mathrm{th}}^2 \right\rangle$ and (11.21) is also recovered for a rectangular cantilever beam with the “static” spring constant k for a rectangular beam from (11.17) [1]. From (11.21) and (C.1) it results that $\left\langle \Delta z_{\mathrm{th},1}^2 \right\rangle = 0.971 \left\langle \Delta z_{\mathrm{th}}^2 \right\rangle$, which means that the first mode already contains 97% of the total energy of the oscillating cantilever.

In the following, we discuss how the spring constant k can be obtained from the thermal deflection noise of the first cantilever mode. When measuring the cantilever deflection voltage $\Delta V_{\mathrm{sensor}}(t)$ and the corresponding deflection $\Delta z(t) = \Delta V_{\mathrm{sensor}}(t) S_{\mathrm{sensor}}$, generally deflection contributions from all modes enter into the RMS deflection signal. The Fourier transformation of the square of the time-

[1] It might be feared that this infinite sum might lead to an infinite total amplitude. However, the spring constants of the higher modes turn out to be very large. Thus, the corresponding thermal oscillation amplitudes become very low and it is generally well known that an monotonously increasing series can have a finite limit.

dependent noise signal is proportional to the noise power spectral density $N^2_{\mathrm{z,th}}(f)$, as introduced in Chap. 5. The noise spectral density is $N_{\mathrm{z,th}}(f) = \sqrt{N^2_{\mathrm{z,th}}(f)}$. In the following, we assume that the noise power spectral density has been measured (by Fourier transformation of the time signal) using a spectrum analyzer.[2] The thermal noise power spectral density as a function of frequency consists of several resonance type peaks, one for each mode at the resonance frequency of the mode. We will extract the spring constant from the strength of the deflection noise of the first mode. In Chap. 17 it was shown that the thermal noise spectral density of the first mode of a cantilever can be written (after the subtraction of a constant background, arising e.g. from electrical noise) as

$$N^2_{\mathrm{z,th,1}} = N^2_{\mathrm{z,th,exc}} G^2(f) = \frac{N^2_{\mathrm{z,th,exc}}}{\left(1 - \frac{f^2}{f^2_{0,1}}\right)^2 + \frac{1}{Q_1^2}\frac{f^2}{f^2_{0,1}}}, \tag{C.3}$$

with $N^2_{\mathrm{z,th,exc}}$ being the white noise arising from the thermal excitation, i.e. frequency-independent. From a fit of this function to the experimentally measured noise density, the parameters $N^2_{\mathrm{z,th,exc}}$, Q_1, and $f_{0,1}$ can be determined. The integral over $G^2(f)$ can be calculated and results as $\pi Q_1 f_{0,1}/2$ (compare Sect. 17.1). Thus, using (C.1) the following additional relation results

$$\left\langle \Delta z_1^2 \right\rangle = \int\limits_0^\infty N^2_{\mathrm{z,th,1}}(f)\mathrm{d}f = N^2_{\mathrm{z,th,exc}} \frac{\pi Q_1 f_{0,1}}{2} = \frac{k_{\mathrm{B}} T}{k_1}. \tag{C.4}$$

With this, the spring constant of the first mode results as

$$k_1 = \frac{2 k_{\mathrm{B}} T}{\pi N^2_{\mathrm{z,th,exc}} Q_1 f_{0,1}}. \tag{C.5}$$

Finally, the spring constant k can be obtained as $k = 0.971 k_1$. Importantly, this thermal method for the determination of the spring constant of the sensor can also be used for other types of sensors than the cantilever beams, for instance quartz sensors, discussed in Sect. 18.3. If the cantilever spring constant is known from other sources, (C.4) can be used to determine the thermal oscillation amplitude $\left\langle \Delta z_1^2 \right\rangle$ and thus S_{Sensor}.

There is another correction which has to be made. The sensitivity S_{Sensor}, which converts the sensor voltage signal to the sensor deflection, was obtained by bending the cantilever via a force applied to the end of the cantilever (Fig. 11.6). However, the thermal method for the spring constant determination is performed with a freely oscillating cantilever. It has been shown that the shapes of the cantilever deflection are slightly different in the two cases [1, 3, 4, 5]. Moreover, for the case of the laser

[2] Details of how to extract the noise power spectral density from the time signal without using a spectrum analyzer are given in [2] and in Appendix B.

beam deflection method, the relevant quantity is not the deflection itself, but the slope of the cantilever $\Delta z'(x)$. The slopes for a free cantilever and the end-loaded cantilever can be calculated. The sensitivity measured for an end-loaded cantilever $S_{\text{sensor,end}}$ has to be replaced by a corrected sensitivity $\chi\ S_{\text{sensor,end}}$ with the correction factor

$$\chi = \frac{S_{\text{sensor,free,calc}}}{S_{\text{sensor,end,calc}}} = \frac{\Delta z'_{\text{free,calc}}}{\Delta z'_{\text{end,calc}}}. \tag{C.6}$$

Thus, the desired sensitivity factor for the free cantilever needed for the thermal method is given by

$$S_{\text{sensor,free}} = S_{\text{sensor,end,measured}} \frac{S_{\text{sensor,free,calc}}}{S_{\text{sensor,end,calc}}} = \chi\ S_{\text{sensor,end,measured}}. \tag{C.7}$$

For the case of an infinitely small laser spot at the end of the cantilever, $\chi = 1.09$ has been calculated. For the cases in which the diameter of the laser spot on the cantilever is finite, and the laser is focused onto a location different from the end of the cantilever, the correction factor χ can be found in a graph shown in Fig. 5 of [5].[3]

References

1. H.-J. Butt, M. Jaschke, Calculation of thermal noise in atomic force microscopy. Nanotechnology **6**, 1 (1995). https://doi.org/10.1088/0957-4484/6/1/001
2. S.M. Cook, T.E. Schäffer, K.M. Chynoweth, M. Wigton, R.W. Simmonds, K.M. Lang, Practical implementation of dynamic methods for measuring atomic force microscope cantilever spring constants. Nanotechnology **17**, 2135 (2006). https://doi.org/10.1088/0957-4484/17/9/010
3. H.-J. Butt, B. Cappella, M. Kappl, Force measurements with the atomic force microscope: Technique, interpretation and applications. Surf. Sci. Rep. **59**, CO2 (2005). https://doi.org/10.1016/j.surfrep.2005.08.003
4. J.L. Hutter, H. Bechhoefer, Calibration of atomic-force microscope tips. Rev. Sci. Instrum. **64**, 1868 (1993). https://doi.org/10.1063/1.1143970
5. R. Proksch, T.E. Schäffer, J.P. Cleveland, R.C. Callahan, M.B. Viani, Finite optical spot size and position corrections in thermal spring constant calibration. Nanotechnology **15**, 1344 (2004). https://doi.org/10.1088/0957-4484/15/9/039

[3] The sensitivity factor which we term S is called *InvOLS* in [5].

Appendix D
Frequency Noise in FM Detection

Here we describe how an amplitude noise of an oscillation gives rise to a corresponding frequency noise. We start by describing some basic principles of the frequency modulation technique applied to our cantilever example as described in [1].

The oscillation of the cantilever at its shifted resonance frequency ω_0' is written (neglecting an offset phase) as

$$z(t) = A\sin(\omega_0' t). \tag{D.1}$$

In the following, we consider the modulation of this carrier oscillation at ω_0' with a modulation frequency ω_{mod}. Such a modulation of the cantilever oscillation can be considered to arise from a modulation of the cantilever resonance frequency due to a signal, e.g. by an (atomic) corrugation giving rise to a modulation with a (frequency) amplitude $\Delta\omega$ which we call here ω_Δ at a frequency ω_{mod} due to scanning. In the PLL FM demodulator the magnitude and frequency of the signal component are extracted.

In the following we will consider that a frequency modulation of the carrier signal arises due to a (sinusoidal) noise component with frequency ω_{mod}, resulting in a time-dependent modulated frequency

$$\omega(t) = \omega_0' + \omega_\Delta \cos(\omega_{\text{mod}} t), \tag{D.2}$$

with ω_Δ being the frequency deviation, i.e. the maximum shift away from ω_0'. Since ω is no longer constant, the phase (i.e. the argument of the sinusoidal oscillation cannot be written as $\phi = \omega t$, but has to be written as an integral over the instantaneous angular frequency $\phi = \int \omega(t)\mathrm{d}t$. With this the oscillation coordinate can be written as

$$z(t) = A\sin\left(\int \omega(t)\mathrm{d}t\right) = A\sin\left(\omega_0' t + \frac{\omega_\Delta}{\omega_{\text{mod}}}\sin(\omega_{\text{mod}} t)\right). \tag{D.3}$$

This expression can be written as an infinite sum over Bessel functions. However, in the limit that $\omega_\Delta \ll \omega_{\text{mod}}$, the oscillation of the cantilever can be approximated as

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3

$$z(t) = A\sin\omega_0' t + \frac{A\omega_\Delta}{2\omega_{\mathrm{mod}}}\left(\sin\left[\left(\omega_0' + \omega_{\mathrm{mod}}\right)t\right] - \sin\left[\left(\omega_0' - \omega_{\mathrm{mod}}\right)t\right]\right). \tag{D.4}$$

This corresponds to an oscillation at the resonance frequency ω_0' and two side bands at the frequencies $\omega_0' \pm \omega_{\mathrm{mod}}$. In the following, we consider a displacement noise component at frequency $\omega_0' + \omega_{\mathrm{mod}}$. The term $A\omega_\Delta/(\sqrt{2}2\omega_{\mathrm{mod}})$ corresponds to a (RMS) displacement noise amplitude which is renamed δA_+. Thus, the cantilever oscillation can be written as

$$z(t) = A\sin\omega_0' t + \sqrt{2}\delta A_+ \sin\left[\left(\omega_0' + \omega_{\mathrm{mod}}\right)t + \phi_0\right]. \tag{D.5}$$

Using the mathematical identity $\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$, the following expression results

$$\begin{aligned} z(t) = A\sin\omega_0' t &\left[1 + \frac{\sqrt{2}\delta A_+}{A}\cos\left(\omega_{\mathrm{mod}}t + \phi_0\right)\right] \\ &+ \sqrt{2}\delta A_+ \cos\omega_0' t \sin\left(\omega_{\mathrm{mod}}t + \phi_0\right). \end{aligned} \tag{D.6}$$

Since $\delta A_+ \ll A$, the second term in the square brackets can be neglected, which results in

$$\begin{aligned} z(t) = A\sin\omega_0' t \cos&\left(\frac{\sqrt{2}\delta A_+}{A}\sin\left(\omega_{\mathrm{mod}}t + \phi_0\right)\right) \\ &+ A\cos\omega_0' t \sin\left(\frac{\sqrt{2}\delta A_+}{A}\sin\left(\omega_{\mathrm{mod}}t + \phi_0\right)\right). \end{aligned} \tag{D.7}$$

In order to apply the above-mentioned identity for trigonometric functions in the next step, we included the factor $\cos\frac{\sqrt{2}\delta A_+}{A}\left(\omega_{\mathrm{mod}}t + \phi_0\right)$, which is very close to one, since $\delta A_+ \ll A$. Further, we also replaced the small term $\frac{\sqrt{2}\delta A_+}{A}\sin\left(\omega_{\mathrm{mod}}t + \phi_0\right)$ by its sinus. Due to this we can apply the above-mentioned identity in the reverse direction, which results in

$$z(t) = A\sin\left(\omega_0' t + \frac{\sqrt{2}\delta A_+}{A}\sin\left(\omega_{\mathrm{mod}}t + \phi_0\right)\right). \tag{D.8}$$

This means that an RMS displacement noise δA_+ at the frequency $\omega_0' + \omega_{\mathrm{mod}}$ translates into a phase noise of RMS amplitude $\delta A_+/A$ at the frequency ω_{mod}. The instantaneous frequency $\omega(t)$ is the time derivative of the phase and can be written as

$$\omega(t) = \omega_0' t + \frac{\sqrt{2}\delta A_+}{A}\omega_{\mathrm{mod}}\cos\left(\omega_{\mathrm{mod}}t + \phi_0\right). \tag{D.9}$$

Thus, the RMS displacement noise δA_+ at the frequency $\omega_0 + \omega_{\text{mod}}$ translates into a RMS frequency noise $\delta\omega_+$, as

$$\delta\omega_+ = \frac{\omega_{\text{mod}}}{A}\delta A_+ \qquad \text{or} \qquad \delta f_+ = \frac{f_{\text{mod}}}{A}\delta A_+, \tag{D.10}$$

correspondingly for the natural frequencies.

If we additionally consider a second independent noise component of the same magnitude from the lower side band at $\omega_0' - \omega_{\text{mod}}$, the frequency noise has to be multiplied by $\sqrt{2}$. While we here explicitly consider the amplitudes of displacement noise and frequency noise the reasoning can also be applied to the spectral noise densities, resulting in

$$N_f(f_{\text{mod}}) = \frac{\sqrt{2} f_{\text{mod}}}{A} N_z(f_0 + f_{\text{mod}}), \tag{D.11}$$

where $N_z(f_0 + f_{\text{mod}})$ is the spectral displacement noise density around f_0 and $N_f(f_{\text{mod}})$ is the demodulated spectral frequency noise density.

Reference

1. K. Kobayashi, H. Yamada, K. Matsushige, Frequency noise in frequency modulation atomic force microscopy. Rev. Sci. Instrum. **80**, 043708 (2009). https://doi.org/10.1063/1.3120913

Index

B. Voigtländer, *Atomic Force Microscopy*, NanoScience
and Technology, https://doi.org/10.1007/978-3-030-13654-3