NanoScience and Technology

Bert Voigtländer

Scanning Probe Microscopy

Atomic Force Microscopy and Scanning Tunneling Microscopy

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

Scanning Probe Microscopy

Atomic Force Microscopy and Scanning Tunneling Microscopy

Springer

Bert Voigtländer
Forschungszentrum Jülich
Peter Grünberg Institut (PGI-3)
Jülich
Germany

and

RWTH Aachen
Lehrstuhl für Experimentalphysik IV A
Aachen
Germany

ISSN 1434-4904 ISSN 2197-7127 (electronic)
NanoScience and Technology
ISBN 978-3-662-45239-4 ISBN 978-3-662-45240-0 (eBook)
DOI 10.1007/978-3-662-45240-0

Library of Congress Control Number: 2014958892

Springer Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer-Verlag GmbH Berlin Heidelberg is part of Springer Science+Business Media (www.springer.com)

For Ortraud

Preface

The aim of this textbook is to introduce *scanning probe microscopy* to graduate students and others wishing to learn about the subject from fundamental principles. The original literature is fascinating but hard going; 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 part of this book covers instrumental aspects and summarizes some basics like the harmonic oscillator. In the parts on atomic force microscopy and scanning tunneling microscopy, the book concentrates mainly on the principles of the methods. A few actually measured images and spectra are shown to demonstrate the principles. In reversed historical order, atomic force microscopy is introduced first since this technique is by far the most frequently used method today.

This book developed from a series of lectures that I gave for more than five years 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 scanning probe microscopy laboratories.

This textbook is not a historical survey of the field and thus will not concern itself with who did what first. I do not cite the original papers unless I feel that they add something that I could not include here. If as an author you perhaps do not feel cited properly, then you are in good company since Gerd Binnig and Heinrich Rohrer are not cited either. 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 you will not find 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. However, I included some details on frequency modulation atomic

force microscopy, since it is my belief that this technique will become more important in the future.

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 that 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 also like to thank Ruslan Temirov for supplying unpublished images from his work.

I am grateful to Michael Crommie, Don Eigler, Randy Feenstra, Franz Giessibl, Markus Heyde, Saw-Wai Hla, Wilson Ho, Gerhard Meyer, Oded Millo, Markus Morgenstern, Nacho Pascual, Udo Schwarz, Jens Wiebe, and Roland Wiesendanger for permitting me to reproduce images from their seminal works.

I would like to thank my former students Anna Strozecka, Stefan Korte, Martin Scheufens, Martin Lanius, Marcus Blab, and Richard Spiegelberg for intense discussions on various topics and for supplying material from their work. A special acknowledgement is due to Janet Carter-Sigglow for her language support. 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/SPMbook. On this page, supplementary material as well as errata will be posted. It also provides a discussion forum and the opportunity to contact me, in order to report errors, or ask questions.

Finally, it is my hope that this book will enable the reader to operate a scanning probe microscope successfully and understand the data obtained with the microscope.

Jülich, Aachen 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 close to the single digit nanometer range are now in production, consumer PCs are become real nanoelectronic devices. Also in many other areas the progress toward the nanoscale is under way.

An additional reason for the trend toward the atomic scale is that material properties are ultimately determined by the atomic structure. In order to understand material properties it is necessary to go down to the nano or atomic scale. However, since the atoms are very small 50 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 "grandfather" 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" 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:

> 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 already saw the potential of nanotechnology already in 1959 before anyone else did. Now more than 50 years later it is interesting to see how many of his predictions have been 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

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_1

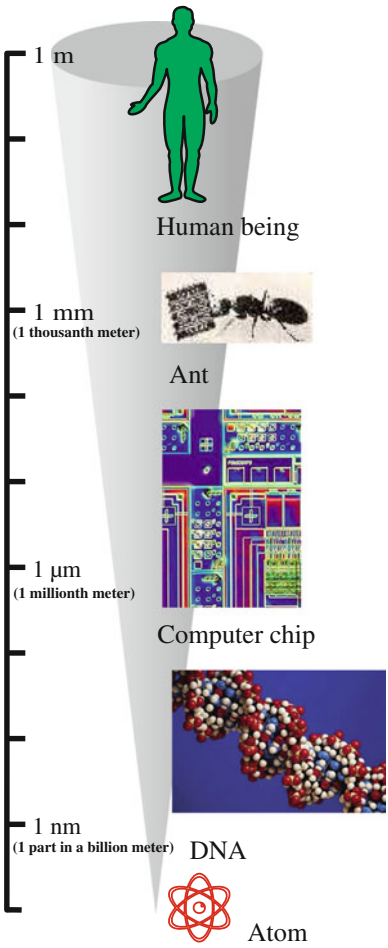

Fig. 1.1 Size scale from the human to the atom

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. The full 1959 speech is available on the internet.

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 first became possible to image matter on the nanoscale. However, scanning probe microscopy is today a unique tool on the nanoscale, because it cannot only image but also structure matter on the nanoscale or even on the atomic scale. In scanning probe microscopy, a small probe is used to detect the local properties at a surface or interface down to atomic resolution. By scanning a grid of points on the surface, the detected properties can be mapped and are usually represented as an image. Because of the scanning mechanism, all these techniques are summarized as scanning probe

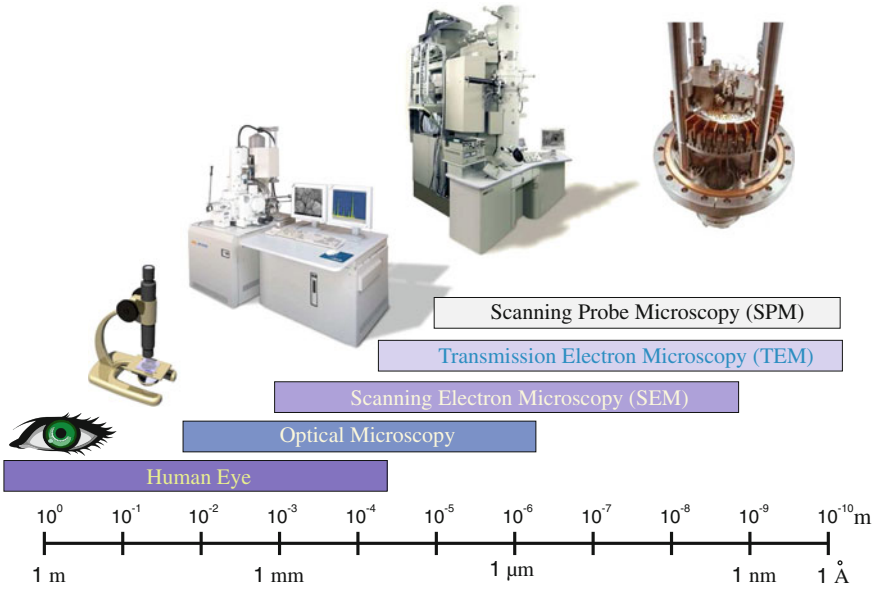

Fig. 1.2 Imaging ranges for different microscopy techniques in comparison

microscopes (SPM). If the interaction between the probe tip and the substrate is strong enough the substrate can be modified on the nanoscale.

One important figure of merit in microscopy is the resolution. Figure 1.2 compares the imaging ranges of different types of microscopy. The resolution of the human eye reaches down to one tenth of a millimeter. Optical 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. Transmission electron microscopy (TEM) 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 surface sensitivity is required SPM with its excellent surface sensitivity is the method of choice. If, however, features below the surface are to be imaged then TEM is the method of choice. If quick imaging within a few minutes down to the nanoscale is required then SEM should be used.

1.1 Introduction to Scanning Tunneling Microscopy

Today the scanning probe microscope is a very important tool in nanoscience. The principle of scanning probe microscopes is to move a sharp tip close to a surface in order to measure various properties with a spatial resolution on the nanometer or even atomic scale. The first kind of scanning probe microscope, the scanning tunneling microscope, (STM) was invented in 1981/1982 by Binnig and Rohrer who received the Nobel prize in physics 1986 for this invention. The most striking property of this kind of microscope is that it provides resolution down to the atomic scale in real space (Fig. 1.3b).

Here is an analogy which shows the precision of an STM 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!

A schematic of an STM, with fine metal tip used as a probe, is shown in Fig. 1.3a. 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 0.5–1 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

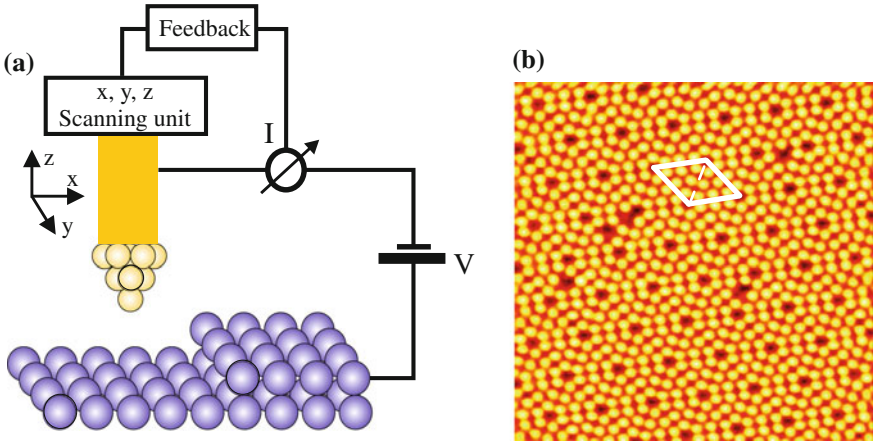

Fig. 1.3 **a** Schematic of a scanning tunneling microscope (STM). **b** STM image of the Si(111) surface. Individual atoms are observed as *yellow dots*. The rhombic unit cell is indicated by *white lines*. Besides the periodic arrangement of the atoms also defects such as single missing atoms can be observed

quantity can be used to measure (and 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-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 resolution.

While the tip is scanned along the surface in x and y directions, a feedback mechanism constantly adjusts the tip height by approaching or retracting the tip to a tip-sample distance at which the tunneling current remains constant. If there is an atomic step at the surface, as shown in Fig. 1.3a, and the tip approaches this step edge laterally during scanning, the tunneling current will rise due to the smaller distance between tip and sample. As a reaction to this the feedback circuit will retract the tip in order to maintain a constant tunneling current, i.e. a constant tip-sample distance. Recording the feedback signal (tip height) 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.

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 larger tip-sample distances) while for the case of the "less conducting atom" the tip has to approach closer to maintain the same tunneling current.

Figure 1.3b shows an atomically resolved image of a Si(111) surface. 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 imaged during operation with an SEM can be accessed at http://www.fz-juelich.de/pgi/pgi-3/microscope.

The tunneling junction (sample-gap-tip) can be treated in different approximations. Here in the introduction, we consider a simple one-dimensional approximation for one electron tunneling in order to grasp the very important exponential dependence of the tunneling current on the tip-sample distance. Later we will look more deeply into the theory of STM.

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 is an oscillating function. The one-dimensional Schrödinger equation is solved by the (not normalized) wave function

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

When drawing such a wave function, it should be always remembered that the quantum mechanical wave function is genuinely a complex function, which is difficult to

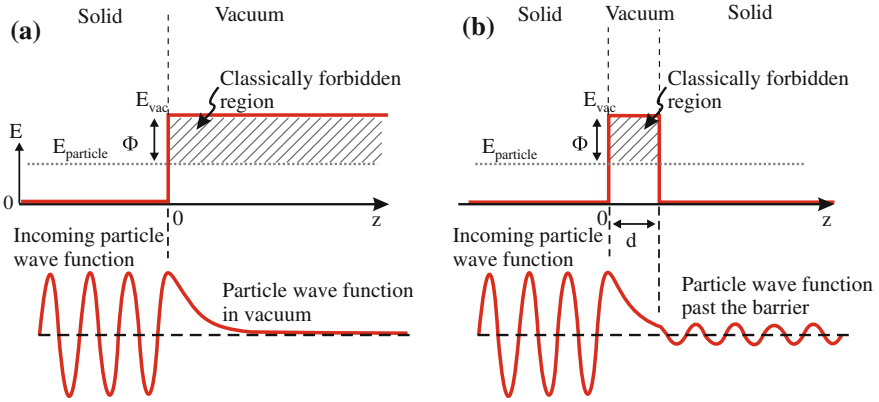

Fig. 1.4 **a** The *top graph* shows the potential diagram with a barrier of height Φ and the energy of an electron $E_{\text{particle}} = E_{\text{F}}$. The *lower graph* shows the real part of the electron wave function with an exponential decay of the wave function in the vacuum region. **b** The *top graph* shows the potential for a solid-vacuum-solid configuration. The *lower graph* shows the electron wave function oscillating in front of the barrier, exponentially decaying inside the barrier and again oscillating past the barrier

draw. Therefore, usually only the real or imaginary part is drawn, as in Fig. 1.4. 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 $E = E_{\text{particle}}$. 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. 1.4a.

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. An ansatz with an exponentially decaying wave function inside the barrier (in the vacuum) as $\psi(z) = \psi(0)e^{-\kappa z}$ leads to a solution of the Schrödinger equation inside this potential barrier (Fig. 1.4a). The probability of a particle being at a position z inside the barrier is approximately proportional to $|\psi(z)|^2$

$$|\psi(z)|^2 = |\psi(0)|^2 e^{-2\kappa z}, \quad \kappa = \sqrt{\frac{2m_{\text{e}}\Phi}{\hbar^2}}. \tag{1.2}$$

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. 1.4b. Since inside the solid the vacuum barrier is not present, the solution for the wave function is an oscillating wave, which is again a solution inside the second solid. This means that in quantum mechanics the electron has a finite probability in both metals. In the square barrier model a barrier, of height $\Phi = E_{\mathrm{vac}} - E_{\mathrm{F}}$ 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. (This is usually done in the quantum mechanics course. We will come to this in a later chapter.) 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 $|\psi(d)|^2$. A transmission coefficient T can be defined as

$$T = \frac{|\psi(d)|^2}{|\psi(0)|^2} \approx e^{-2\kappa d}. \tag{1.3}$$

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 at distance d from the surface. Later we will see that the tunneling current is proportional to the transmission coefficient.

Evaluating (1.2) using the free electron mass for m_{e} and a typical value for the work function of a metal ($\Phi \approx 4.5\,\mathrm{eV}$), 2κ is about $20\,\mathrm{nm}^{-1}$. Thus a variation of the barrier thickness of 0.1 nm results in a difference in the transmission factor of an order of magnitude ($\sim$7.4). Hence the tunneling current increases by about an order of magnitude if the tip approaches by one Å to the sample. This sensitivity in the tip-sample distance is the reason for the extremely high vertical resolution of the STM which can reach the picometer regime. Atoms on the tip which protrude only 2.5 Å ($\sim$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.

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 an atomic force microscope force microscopy the force between the tip and sample is measured. In Fig. 1.5, a qualitative sketch of the force between tip and sample is given. Three different regimes can be distinguished.

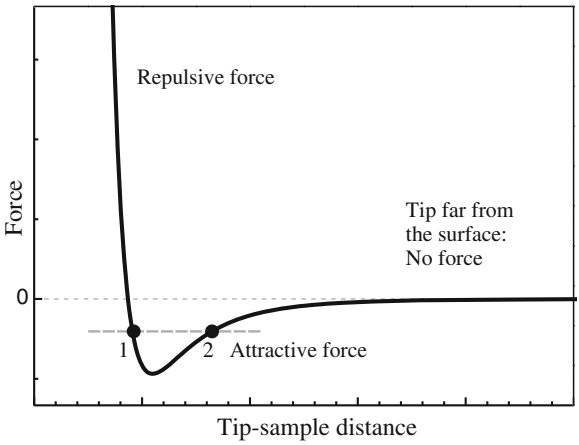

Fig. 1.5 Qualitative behavior of the force between tip and sample as function of tip-sample distance

(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. 1.5. Care has to 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 the cantilever on which a tip is mounted. 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 $F = -kz$, where F is the force and z is the distance the cantilever spring is bent relative to its equilibrium position without the sample present. Figure 1.6 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. 1.7.

In the static mode of operation, the surface contour is mapped while scanning by changing the z-position of the tip in such a way that the tip-sample force and, correspondingly, the tip-sample distance are kept constant. The tip position 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, as shown in Fig. 1.7. 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

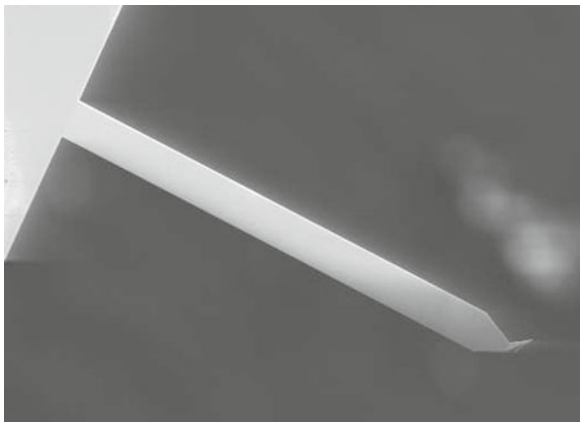

Fig. 1.6 SEM image of a silicon cantilever used in atomic force microscopy with a length of 450 μm

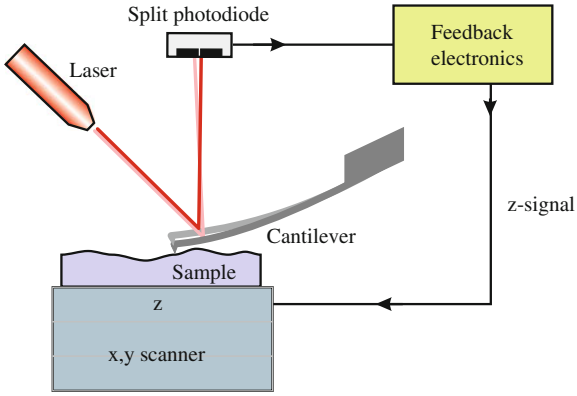

Fig. 1.7 Schematics of atomic force microscopy operation

the repulsive regime of the force-distance curve the operating mode is called contact mode. The last atoms of the tip are in direct contact with the surface atoms.

The atomic force microscope can also be operated in the so called dynamic mode with an oscillating cantilever. This dynamic mode is often 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 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 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 scheme of force detection 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.

1.3 A Short History of Scanning Probe Microscopy

It is a strange fact in the history of science that the scanning tunneling 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 STM 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 (precursor of the STM).
- 1981 Construction of the first STM by Binnig, Rohrer, Weibel and Gerber.
- 1982 First image of the atomic structure of the Si(111)-(7 × 7) surface by Binnig, Rohrer, Weibel and Gerber.
- 1985 Invention of the atomic force microscope (AFM) by Binnig, Quate and Gerber.
- 1986 Nobel prize in physics for the invention of the STM awarded to Binnig and Rohrer.
- 1987 Element-sensitive imaging of GaAs by Feenstra.
- 1990 Optical beam deflection method introduced by Meyer and Amer.
- 1990 First positioning of single atoms on a surface with a low temperature STM by Eigler.
- 1993 Tapping mode introduced by Zhong, Inniss, Kjoller, and Elings.
- 1995 First atomic resolution with an AFM by Giessibl.
- 1998 First vibrational spectroscopy with the STM by Stipe and Ho.

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- 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.
- In STM, during scanning the height of the tip is adjusted by a feedback loop (and recorded as the topography signal) such that the tunneling current and correspondingly the tip-sample distance is kept constant.
- Atomic force microscopy can be also applied to insulating samples. The deflection of a small cantilever senses the force between tip and sample.
- In the dynamic operation mode, the cantilever oscillates and the resonance frequency and subsequently the amplitude change due to the force between tip and sample.

Part I
Scanning Probe Microscopy Instrumentation

Chapter 2
Harmonic Oscillator

In scanning probe microscopy, vibrations play a central role in several areas. If, for instance, a scanning tunneling microscope is 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 STM. 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 atomic force microscopy. In the dynamical mode of atomic force microscopy, a cantilever vibrating close to (or at) its resonance frequency is used as a force detector. 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

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

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_2

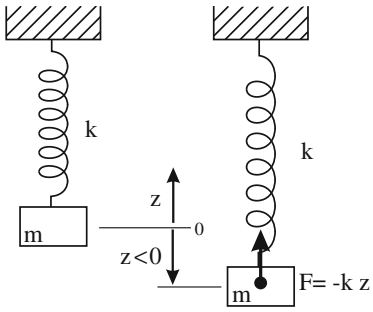

Fig. 2.1 The simplest example of a harmonic oscillator: a mass on a spring

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:

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

Formally (2.2) is solved if $\omega_0 = \sqrt{\frac{k}{m}}$. 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 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

$$\omega_0 = 2\pi/T. \tag{2.4}$$

The angular frequency ω_0 is the number of radians through which the oscillation proceeds per time, while the frequency f_0 is the number of oscillations per time ($\omega_0 = 2\pi f_0$). 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. 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).

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

$$z = A\cos(\omega_0 t) + B\sin(\omega_0 t). \tag{2.5}$$

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 (the argument of the cosine function is called phase) by a constant

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

phase shift ϕ. Thus all solutions are captured if the solution is shifted by a constant (but arbitrary) phase shift ϕ, and the general solution results as

$$z = a\cos(\omega_0 t + \phi). \tag{2.6}$$

The two solutions given in (2.5) and (2.6) are in fact equivalent. Using the mathematical identity

$$\cos(\alpha + \beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta, \tag{2.7}$$

the following relations between A, B in (2.5) and a, ϕ in (2.6) are obtained

$$B = a\cos\phi, \quad A = a\sin\phi. \tag{2.8}$$

Moreover, the solutions given in (2.5) and (2.6) 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, 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 (2.5) and its derivative

$$v(t) = -\omega_0 A\sin(\omega_0 t) + \omega_0 B\cos(\omega_0 t). \tag{2.9}$$

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

$$z_0 = A + B\cdot 0 = A \quad v_0 = -\omega_0 A\cdot 0 + \omega_0 B = \omega_0 B. \tag{2.10}$$

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

$$A = z_0 \quad \text{and} \quad B = v_0/\omega_0. \tag{2.11}$$

2.2 Driven Harmonic Oscillator

In dynamic atomic force microscopy, we will consider a cantilever which is exited, 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 (cf. Fig. 2.1) is oscillated (excited) sinusoidally with $z_{\text{drive}}(t) = A_{\text{drive}}\cos(\omega_{\text{drive}} t)$. The resulting force on the mass m is then $F = -k(z - z_{\text{drive}})$. The equation of motion results as

$$ma = m\ddot{z} = -k(z - z_{\text{drive}}). \tag{2.12}$$

The driving frequency ω_{drive} can be different from the natural frequency of the oscillator ω_0. The question arises at which frequency the driven harmonic oscillator will oscillate. At its natural 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

$$z(t) = A\cos(\omega_{\text{drive}}t). \tag{2.13}$$

Inserting this ansatz into the equation of motion (2.12) results in

$$-m\omega_{\text{drive}}^2 A\cos(\omega_{\text{drive}}t) = -m\omega_0^2 A\cos(\omega_{\text{drive}}t) + kA_{\text{drive}}\cos(\omega_{\text{drive}}t). \tag{2.14}$$

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

$$A = \frac{kA_{\text{drive}}}{m(\omega_0^2 - \omega_{\text{drive}}^2)}. \tag{2.15}$$

The special solution (2.13) means that m oscillates at the driving frequency with an amplitude which depends on the driving frequency and also on the natural 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 for an undamped driven harmonic oscillator are shown in (Fig. 2.2). If $\omega_{\text{drive}} \ll \omega_0$ the amplitude A approaches the

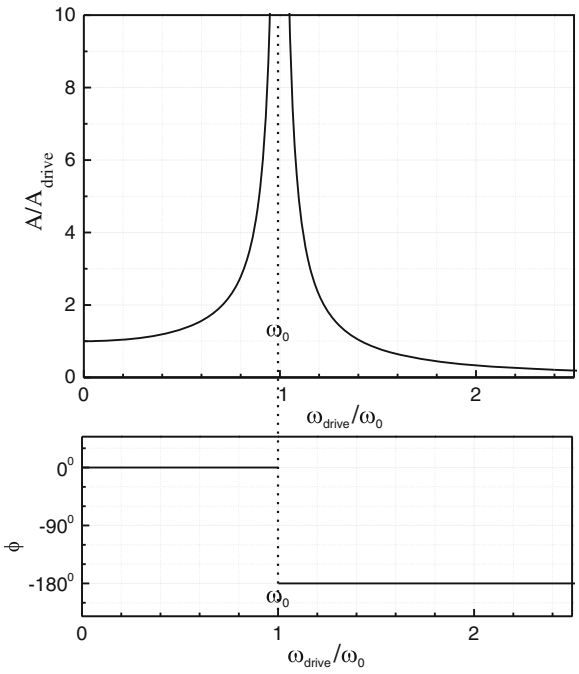

Fig. 2.2 Amplitude and phase of an undamped driven harmonic oscillator as a function of ω_{drive} showing a resonance at ω_0

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. 2.2 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".

2.3 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 (2.12). We consider this term as proportional to the speed at which the oscillating mass moves $F_{\text{frict}} = m\gamma\dot{z}$. Also here we assume 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

$$m\ddot{z} = -m\gamma\dot{z} - k(z - z_{\text{drive}}). \tag{2.16}$$

Replacing $\omega_0^2 = k/m$ results in

$$\ddot{z} + \gamma\dot{z} + \omega_0^2 z = \omega_0^2 z_{\text{drive}}. \tag{2.17}$$

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 amplitudes are regarded as complex numbers as

$$\tilde{z} = Ae^{i(\omega t+\phi)} = Ae^{i\phi}e^{i\omega t} = \hat{z}e^{i\omega t} \quad \text{and} \quad \tilde{z}_{\text{drive}} = A_{\text{drive}}e^{i\omega t}. \tag{2.18}$$

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 motion of the mass m. The nice thing about the complex notation is that differentiation of $\tilde{z}$ is now just multiplication with $i\omega$ ($\frac{d\tilde{z}}{dt} = \hat{z}i\omega e^{i\omega t} = i\omega\tilde{z}$). This means differentiation in (2.17) (with $z \to \tilde{z}$) can be easily executed and this differential equation converts to the simple algebraic equation

$$\left[(i\omega)^2\hat{z} + \gamma(i\omega)\hat{z} + \omega_0^2\hat{z}\right]e^{i\omega t} = \omega_0^2 A_{\text{drive}}e^{i\omega t}. \tag{2.19}$$

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

$$\hat{z} = \frac{\omega_0^2 A_{\text{drive}}}{\omega_0^2 - \omega^2 + i\gamma\omega}. \tag{2.20}$$

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

$$z = Re(\tilde{z}) = Re(\hat{z}e^{i\omega t}) = Re(Ae^{i(\omega t+\phi)}). \tag{2.21}$$

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

$$z = A\cos(\omega t + \phi), \tag{2.22}$$

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

$$A^2 = \frac{\omega_0^4 A_{\text{drive}}^2}{\left(\omega_0^2 - \omega^2 + i\gamma\omega\right)\left(\omega_0^2 - \omega^2 - i\gamma\omega\right)} = \frac{\omega_0^4 A_{\text{drive}}^2}{\left(\omega^2 - {\omega_0}^2\right)^2 + \gamma^2\omega^2}. \tag{2.23}$$

Now we introduce as a convenient abbreviation the quality factor $Q = \omega_0/\gamma$. The physical significance of the quality factor will be elucidated later. This replacement results in

$$A^2 = \frac{\omega_0^4 A_{\text{drive}}^2}{\left(\omega^2 - \omega_0^2\right)^2 + \frac{\omega_0^2\omega^2}{Q^2}}. \tag{2.24}$$

Furthermore, the oscillation amplitude A can be written as a function of the normalized frequency ω/ω_0 as

$$A^2 = \frac{A_{\text{drive}}^2}{\left(1 - \frac{\omega^2}{\omega_0^2}\right)^2 + \frac{1}{Q^2}\frac{\omega^2}{\omega_0^2}}. \tag{2.25}$$

The phase 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 $\tan\varphi = \frac{Im(x)}{Re(x)}$. In order to calculate the phase ϕ, we recall that $\hat{z} = Ae^{i\phi}$. However, according to (2.20) the real and imaginary parts of $1/\hat{z}$ are much easier to find. Therefore, we write

$$\frac{1}{\hat{z}} = \frac{1}{Ae^{i\phi}} = \frac{1}{A}e^{-i\phi} = \frac{1}{\omega_0^2 A_{\text{drive}}}\left(\omega_0^2 - \omega^2 + i\gamma\omega\right). \tag{2.26}$$

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

$$\tan\phi = \frac{-\gamma\omega}{\omega_0^2 - \omega^2} = \frac{-\omega_0\omega}{Q\left(\omega_0^2 - \omega^2\right)}. \tag{2.27}$$

Also the phase ϕ can be written as function of the normalized frequency ω/ω_0 as

$$\tan\phi = \frac{-\frac{\omega}{\omega_0}}{Q\left[1 - \left(\frac{\omega}{\omega_0}\right)^2\right]}. \tag{2.28}$$

With these results, the amplitude (2.25) and phase (2.28) in the solution (2.22) are calculated as a function of given variables. The resonance curve in Fig. 2.3 shows the amplitude and the phase of a driven damped harmonic oscillator. For small driving frequencies $\omega \ll \omega_0$, the motion of the oscillator mass just follows the outer excitation with a phase approaching zero; i.e. the oscillation is in phase with the excitation. On

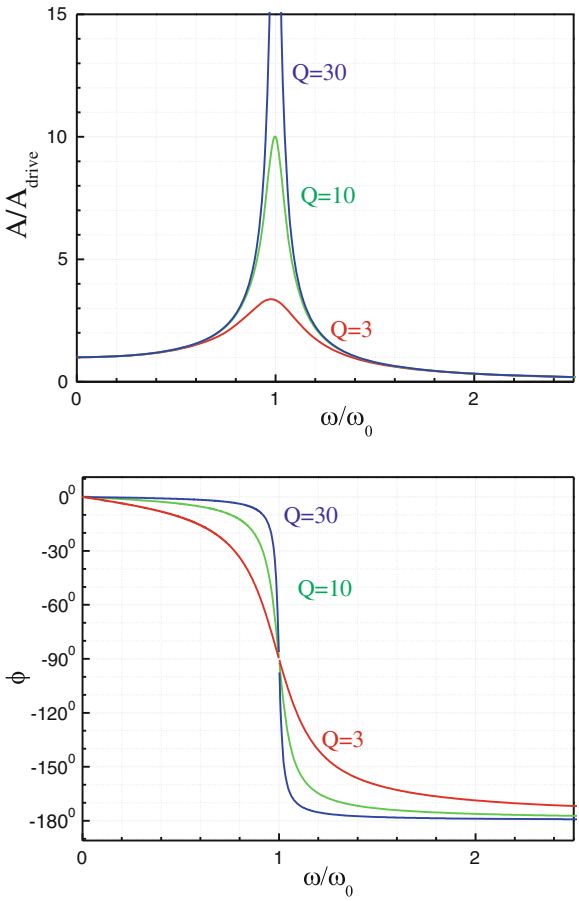

Fig. 2.3 Amplitude and phase of a damped driven harmonic oscillator as a function of $\omega \equiv \omega_{\text{drive}}$, for different values of damping $Q = \omega_0/\gamma$

the other hand for very large frequencies $\omega \gg \omega_0$, the amplitude A approaches zero. In this case the phase 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 (2.25) we find that the amplitude is proportional to $1/\omega^2$ for small damping, i.e. $\gamma \ll \omega_0$ or $Q \gg 1$. As seen in Fig. 2.3, 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 is smaller than zero, which means that the oscillator displacement z always lags behind the driving excitation (Fig. 2.3). The phase 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 (2.25) as

$$A(\omega_0) = Q A_{\text{drive}}, \tag{2.29}$$

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 (2.25) can be approximated as

$$A^2 = \frac{A_{\text{drive}}^2}{\left[\left(1+\frac{\omega}{\omega_0}\right)\left(1-\frac{\omega}{\omega_0}\right)\right]^2 + \frac{1}{Q^2}\frac{\omega^2}{\omega_0^2}} \approx \frac{A_{\text{drive}}^2}{4\left(1-\frac{\omega}{\omega_0}\right)^2 + \frac{1}{Q^2}}. \tag{2.30}$$

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[2] at ω_0. This condition for the amplitudes can be written as

$$A_{1/2}(\omega_{1/2}) = \frac{1}{\sqrt{2}} A(\omega_0) = \frac{1}{\sqrt{2}} Q A_{\text{drive}}. \tag{2.31}$$

If we insert $\omega = \omega_{1/2}$ in expression (2.30), the following relation results

$$\frac{1}{2} A_{1/2}^2(\omega_{1/2}) \approx \frac{A_{\text{drive}}^2}{4\left(1-\frac{\omega_{1/2}}{\omega_0}\right)^2 + \frac{1}{Q^2}} \approx \frac{1}{2} Q^2 A_{\text{drive}}^2. \tag{2.32}$$

[2] 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.

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 this, we obtain

$$\Delta\omega_{1/2} \approx \frac{\omega_0}{Q}. \tag{2.33}$$

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 (2.25) becomes minimal. Differentiating the denominator of (2.25) with respect to ω/ω_0, and equating this derivative to zero results in the following expression for the frequency $\omega_{\max}$ at which the resonance curve has its maximum

$$\omega_{\max}^2 = \omega_0^2 \left(1 - \frac{1}{2Q^2}\right). \tag{2.34}$$

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

$$\delta\omega = \omega_0 - \omega_{\max} = \omega_0 \left(1 - \sqrt{1 - \frac{1}{2Q^2}}\right). \tag{2.35}$$

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 $\omega_0 = 300\,\text{kHz}$ 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.

2.4 Transients of Oscillations

The solution for the damped driven harmonic oscillator (2.22) 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 (2.22) 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 (2.22) plus a solution of the corresponding homogeneous problem. The corresponding homogeneous problem is the damped harmonic oscillator without external driving. Here we do not derive the solution for the damped oscillator without driving but it should be remembered that this is (for small damping)

an exponentially decaying oscillation $z_{\text{hom}} = G\exp(-\omega_0/(2Q)t)\cos(\omega_{\text{hom}}t - \phi)$ with the oscillation frequency ω_{hom} being slightly lower than the natural frequency ω_0 of the free harmonic oscillator $\omega_{\text{hom}} = \omega_0\sqrt{1 - 1/(4Q^2)}$ and with G and ϕ as coefficients specified by the initial conditions.

If we call the specific solution z in (2.22) z_s, the general solution for the driven, damped harmonic oscillator is given as $z_{\text{general}} = z_{\text{hom}} + z_s$. 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 G 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. 2.4 the general solution for the initial condition: starting from rest, is shown to be composed of the specific solution of the inhomogeneous system (Fig. 2.4a) plus the solution for the homogeneous system (transient) z_{hom} (Fig. 2.4b). In Fig. 2.4c the sum of both is shown for the case that $\omega = \omega_{\text{hom}}$. The specific solution in Fig. 2.4a is approached within the decay time for the homogeneous solution Fig. 2.4b. The fact that the situation is not always simple is shown in Fig. 2.4d. Here the driving frequency deviates from ω_{hom}, which leads to a beating behavior before a steady-state solution is reached.

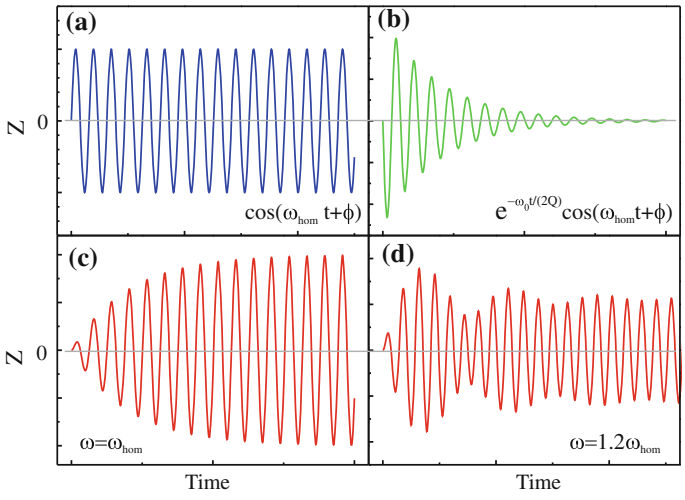

Fig. 2.4 The general solution for a damped driven harmonic oscillator is composed of the specific solution of the inhomogeneous driven system (steady-state solution), shown in (**a**) plus the solution of the homogeneous system without driving (transient), shown in (**b**). The initial conditions are 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

If the driven damped oscillator is oscillating in steady-state (Fig. 2.4a) 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. 2.4b. This is a sinusoidal oscillation with the envelope decreasing as $\exp(-\omega_0/(2Q)t)$. 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 time can be expressed in terms of the Q-factor as $\tau = 2Q/\omega_0 = TQ/\pi$. This means that the oscillation builds up (decays) within roughly Q oscillation cycles and Q can be expressed as

$$Q = \frac{1}{2}\tau\omega_0. \tag{2.36}$$

2.5 Dissipation and Quality Factor of a Damped Driven Harmonic Oscillator

When the mass 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 $F_{\text{frict}}v = \gamma m v^2$ and varies over one period, as v varies. The mean power consumed by the oscillator in steady-state can be written as

$$\langle P\rangle = \langle F_{\text{frict}}v\rangle = \gamma m\langle v^2\rangle. \tag{2.37}$$

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

$$\langle P\rangle = \gamma m\langle v^2\rangle = \frac{1}{2}\gamma m\omega^2 A^2. \tag{2.38}$$

With this the energy dissipated per cycle is

$$\text{Energy dissipated per cycle} = \langle P\rangle T = \langle P\rangle 2\pi/\omega = \pi\gamma m\omega A^2. \tag{2.39}$$

Another important quantity is the total energy stored in the oscillator. 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

$$\langle E\rangle \approx \frac{1}{2}kA^2 = \frac{1}{2}m\omega_0^2 A^2. \tag{2.40}$$

The efficiency of an oscillator is defined by how much energy is stored, compared with how much work is supplied (dissipated) by the external force per oscillation cycle. This is called the quality factor of the oscillator and is defined by 2π times the mean energy stored, divided by the energy dissipated per cycle

$$Q = 2\pi \times \frac{\text{Energy stored in the oscillator}}{\text{Energy dissipated per cycle}}. \tag{2.41}$$

Close to the resonance frequency ($\omega \approx \omega_0$), Q can be written using (2.39) and (2.40) as

$$Q \approx \frac{\omega_0}{\gamma}. \tag{2.42}$$

This is consistent with the abbreviation for Q introduced in the previous section.

2.6 Effective Mass of a Harmonic Oscillator

In this chapter, we always considered an idealized system consisting of a mass-less spring and a mass m at its end. However, in some 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. 2.5b). We introduce the concept of the effective mass for the example of a coil spring (with mass m_{spring}) and assume that the mass

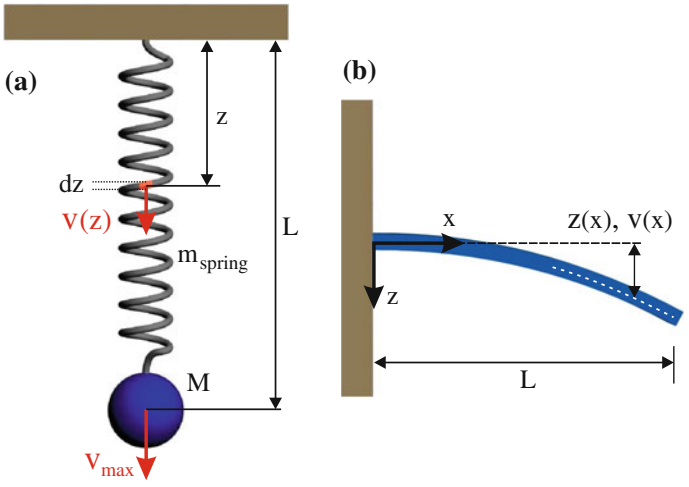

Fig. 2.5 **a** For a spring with mass m_{spring}, the velocity of a volume element depends on the position, i.e. $v = v(z)$. The effective mass turns out to be 1/3 of the spring mass. **b** For a cantilever beam the deflection and the velocity are non-linear as a function of x

is distributed homogeneously along its length. In the following, we calculate the maximum kinetic energy (which corresponds to the total energy) of the spring with a mass and we do not consider a mass M at the end of the spring.

When calculating the (maximum) kinetic energy of the spring, we regard $v(z)$ as the maximum velocity during one oscillation cycle. The (maximum) kinetic energy of a length element dz of the spring is given by

$$dE_{\text{kin}} = \frac{1}{2}\frac{m_{\text{spring}}}{L}v^2(z)dz. \tag{2.43}$$

According to Fig. 2.5a, the velocity distribution along the spring is linear with z and can be written as $v(z) = v_{\text{max}}z/L$, with v_{max} being the maximum velocity at the end of the spring, i.e. $v(L)$. Integrating the (maximum) kinetic energy along the spring results in

$$\begin{aligned} E_{\text{kin}} &= \frac{1}{2}\int_0^L \frac{m_{\text{spring}}}{L}v^2(z)dz = \frac{1}{2}\frac{m_{\text{spring}}}{L}\int_0^L v_{\text{max}}^2\frac{z^2}{L^2}dz \\ &= \frac{1}{2}\left(\frac{1}{3}m_{\text{spring}}\right)v_{\text{max}}^2 = \frac{1}{2}m_{\text{eff}}v_{\text{max}}^2. \end{aligned} \tag{2.44}$$

Thus a mass-containing spring is equivalent to a massless spring with an effective mass $m_{\text{eff}} = 1/3m_{\text{spring}}$ fixed to the end of the spring. If an additional mass M at the end of a spring is also considered, the effective mass becomes $m_{\text{eff}} = M+1/3m_{\text{spring}}$.

While we only considered the expression of the kinetic energy here, the same effective mass also enters into the equations of motion, and thus also into all following results. For instance, when calculating the natural frequency of a harmonic oscillator in which the spring contains mass, the effective mass has to be used instead of the mass M at the end of a massless 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. 2.5b. According to [1], the bending has the form $z(x) \propto -x^3 + 3x^2L$. 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 $v(L) = v_{\text{max}}$ as $c = v_{\text{max}}/(2L^3)$. Thus the maximum velocity at position x along the beam results as

$$v(x) = \frac{v_{\text{max}}}{2L^3}\left(-x^3 + 3x^2L\right). \tag{2.45}$$

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

$$E_{\text{kin}} = \frac{1}{2}\int_0^L \frac{m_{\text{cant}}}{L}\frac{v_{\text{max}}^2}{4L^6}\left(-x^3+3x^2L\right)^2 dx = \frac{1}{2}\left(\frac{33}{140}m_{\text{spring}}\right)v_{\text{max}}^2$$
$$= \frac{1}{2}m_{\text{eff}}v_{\text{max}}^2. \tag{2.46}$$

Thus the effective mass for a cantilever beam turns out to be ∼0.2357, instead of 1/3 for a coil spring with a linear extension.

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 haves to be used.

2.7 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

$$0 = a_1x + a_2\dot{x} + a_3\ddot{x}. \tag{2.47}$$

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

$$F(t) = a_1x + a_2\dot{x} + a_3\ddot{x}. \tag{2.48}$$

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

$$F(t) = a_1(x_1+x_2) + a_2(\dot{x}_1+\dot{x}_2) + a_3(\ddot{x}_1+\ddot{x}_2) = a_1x + a_2\dot{x} + a_3\ddot{x}. \tag{2.49}$$

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.

2.8 Summary

- The free harmonic oscillator has the natural 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 $F_{\text{frict}} = m\gamma\dot{z}$.
- The phase 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 the ratio between the energy stored in the oscillator to the energy dissipated per cycle. $Q \approx \frac{\omega_0}{\gamma} \approx \frac{\omega_0}{\Delta\omega} \approx A(\omega_0)/A_{\text{drive}}$, with $\Delta\omega$ being the width of the resonance curve and A_{drive} the excitation amplitude.
- The build up or the decay of the steady-state amplitude takes about Q oscillations, i.e. the corresponding time constant is $\tau = 2Q/\omega_0$.
- If a spring has a non-negligible mass, the effective mass has to be used in the equations of the harmonic oscillator.

Chapter 3
Technical Aspects of Scanning Probe Microscopy

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.

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 $\mathbf{r}$ inside the unit cell, the opposite dipole is also present at the position $-\mathbf{r}$ due to the inversion symmetry and the net dipole moment of the unit cell is zero. During a directional deformation of a piezoelectric material, microscopic dipoles are formed 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

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_3

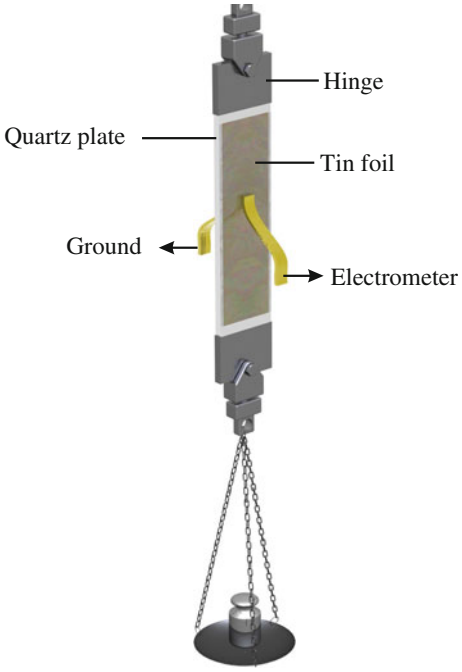

Fig. 3.1 Curie brothers' experiment demonstrating the piezoelectric effect

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 $Pb[Zr_xTi_{1-x}]O_3$). 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 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.

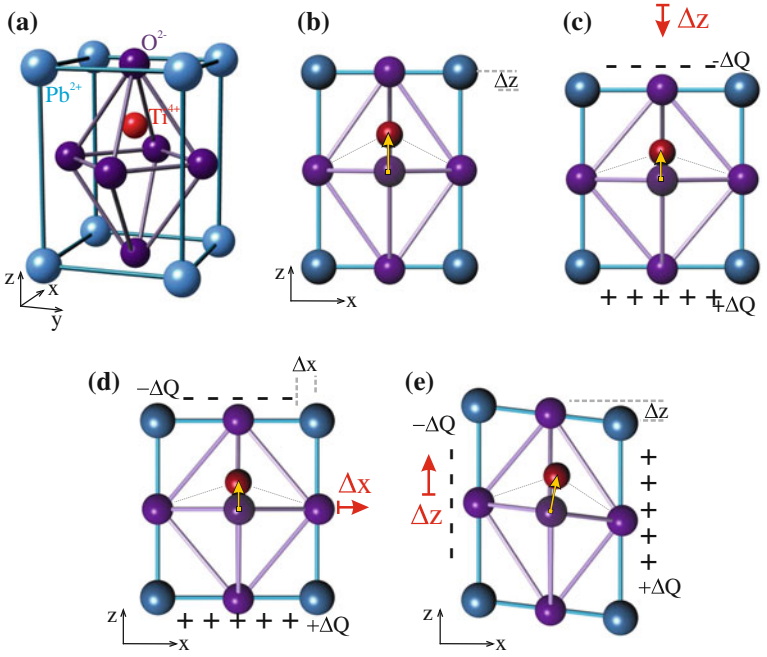

Fig. 3.2 **a** Schematic of the PZT unit cell. **b** *Side view* of the PZT unit cell with the dipole induced by the displaced Ti^{4+}. **c** Longitudinal piezoelectric effect: upon compression of the unit cell along the z-axis the dipole is reduced leading to a corresponding change of the surface charge. **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 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 an "inversion symmetry along the x-axis". 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. 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.

Here we discuss 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. The charge applied to the outer metallic electrodes leads to a change of the dipole moment in the piezoelectric 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

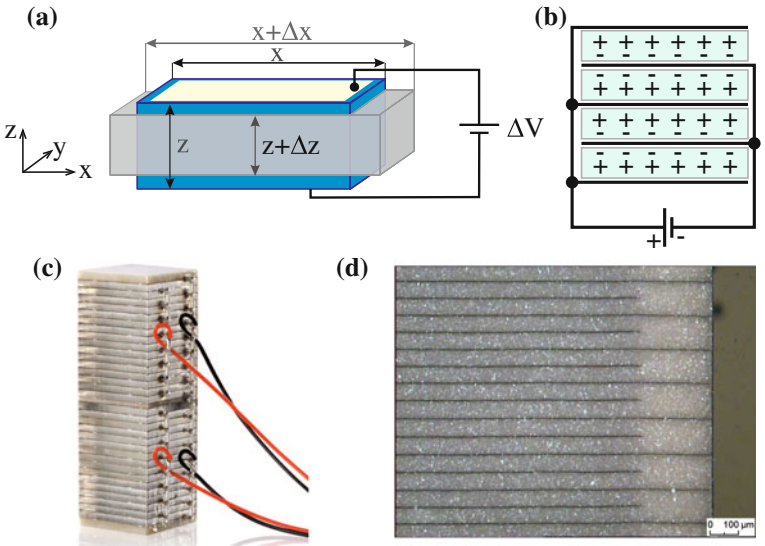

Fig. 3.3 **a** Sketch of a piezo plate (dimensions x, y, and z) poled in the z-direction. Considering the longitudinal piezo effect, an electric field in the z-direction induced by a voltage ΔV in z-direction induces a strain in z-direction, Δz. Considering the transverse piezoelectric effect a voltage 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])

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 contraction of the material (Poisson effect), a transverse strain in the x-direction. If a piezo plate as in Fig. 3.3a of thickness z 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 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

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

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 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. 3.3 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 length 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. 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 extension 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 STM operation, as we will see later. 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 neglect the change in the width of the plate

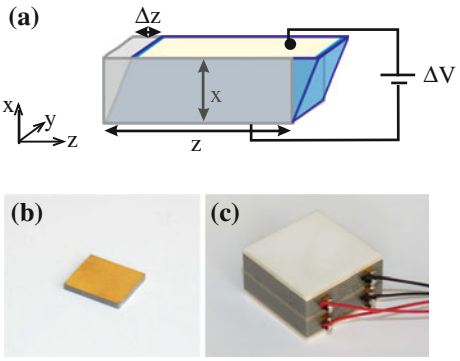

Fig. 3.4 **a** Sketch of a piezoelectric plate operated using the shear piezo effect. **b** Photo of a single shear piezo plate (6 mm × 7 mm). **c** Photo of a shear piezo stack (15 mm × 15 mm)

due to the transverse effect. On the other hand, when we focused on the transverse extension of a plate, we neglected 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 $\Delta z/x$. In this case, the piezo constant is independent of the dimensions and is called (due to some conventions)

$$\frac{\Delta z}{\Delta V} = d_{15}. \tag{3.4}$$

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, $Pb[Zr_xTi_{1-x}]O_3$) are used as materials for piezoelectric actuators because ceramics can be formed into various shapes and because of their large piezo constant. 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, 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

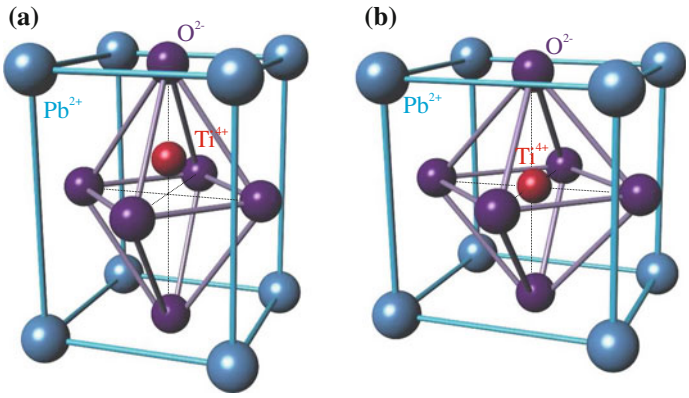

Fig. 3.5 Unit cell of the PZT crystal structure **a** below the Curie temperature **b** above the Curie temperature

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 man-

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_{51} (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

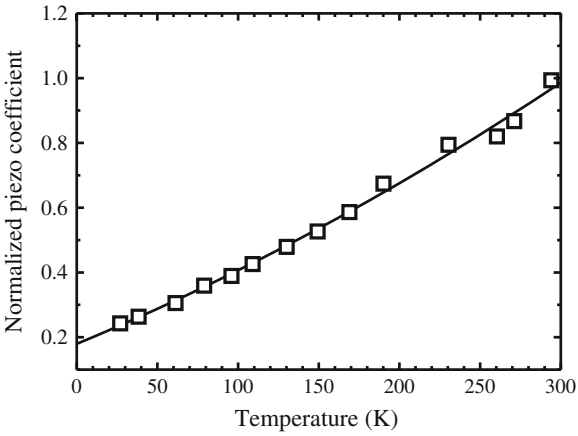

Fig. 3.6 Temperature dependence of the piezoelectric constants d_{31} for PZT-5A piezo ceramic material relative to the room temperature value (adapted from [3])

ufacturer 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.

3.4 Tube Piezo Element

One central task in scanning probe 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,

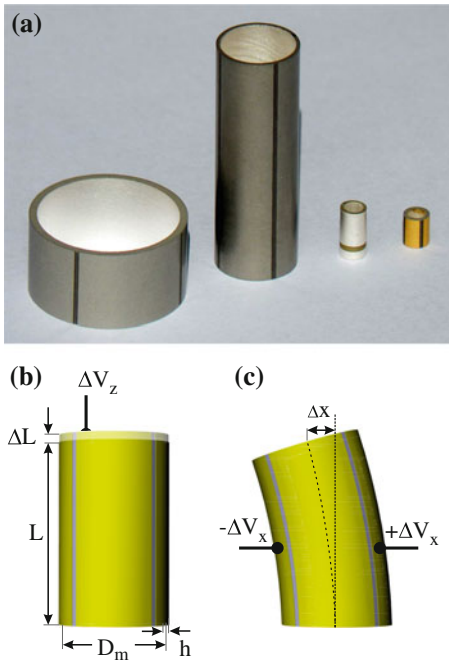

Fig. 3.7 **a** Photograph of several tube piezo elements. **b** Schematic *side view* of a tube scanner showing the vertical extension along z. **c** Schematic of the lateral movement in the x-direction

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 scanner is calculated using (3.3) (exchanging 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

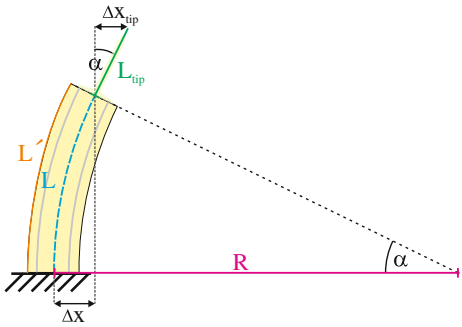

Fig. 3.8 Sketch of the geometry of a bent piezo tube with the relevant parameters

(due to the definition of the arc length) the bending angle as

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

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 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.

Typical dimensions of a piezo tube (PZT-5A) are as follows: length 25.4 mm, mean diameter 5.84 mm, wall thickness 0.51 mm, which 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) and (3.8) as

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

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 [4]. 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, good linearity in x and y directions is achieved and a coupling with 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.

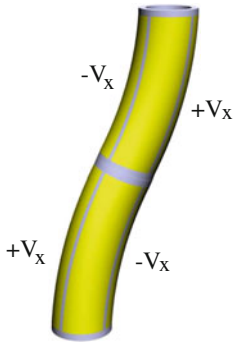

Fig. 3.9 Instead of an outside electrode divided into four segments the outer electrode has eight segments. The *upper part* of the piezo is bent in the opposite direction to prevent a displacement in the z-direction

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 [1]. 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 lowest longitudinal (axial) vibrational stretching mode of a rod or tube with one end clamped and one end free is

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

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

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

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

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

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

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

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) $I = \int z^2 dA$, where z is the direction of bending. The expression $\sqrt{I/A} = \kappa$ is called the radius of gyration and has the following expressions: for a

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

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

$$f_{\text{bend}} = \frac{1}{2\pi}\sqrt{\frac{3EI}{L^3(M + 0,2357m)}}. \tag{3.20}$$

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 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. 3.12) should be used for several hours at room temperature, or rather at elevated temperature but still below the Curie temperature.

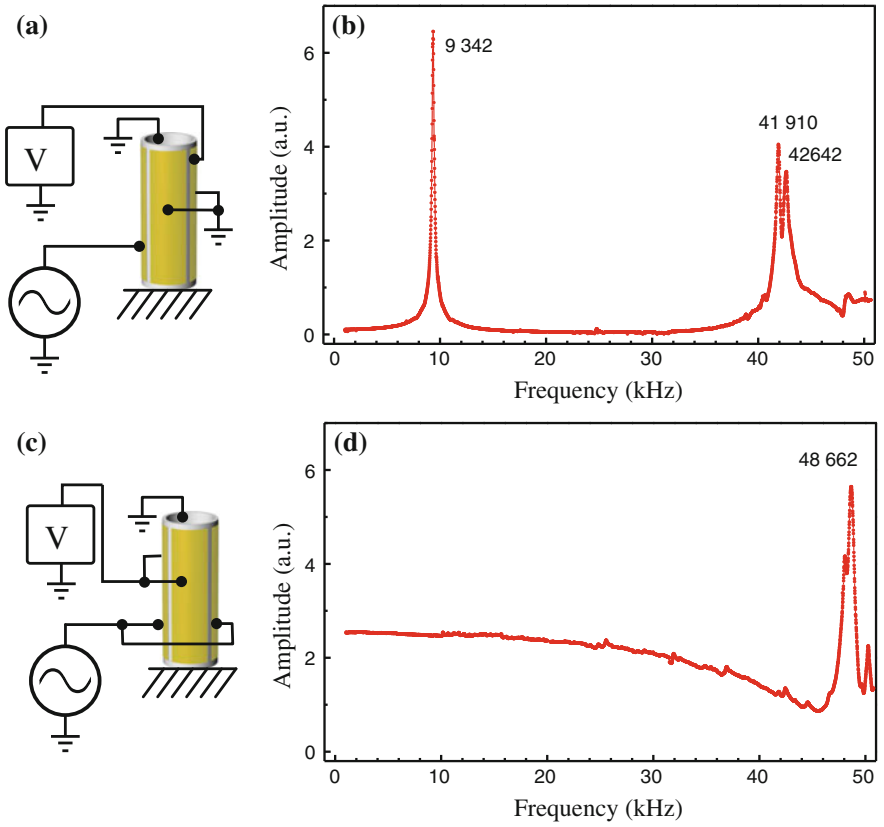

Fig. 3.10 **a** Schematic of the measurement setup with an electric excitation of the mechanic oscillation of a tube piezo element (bending mode). The amplitude of the mechanically excited oscillation is detected by the piezoelectric effect. **b** Amplitude of the mechanic oscillation. Resonances are 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

3.5 Flexure-Guided Piezo Nanopositioning Stages

A further continuously moving nanopositioning system uses flexure guides. It relies on the elastic deformation of a spring-like structure which confines the motion in only one direction and is driven by a piezo element. The working principle can be seen in Fig. 3.11a. 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 allow a spring-like motion along one direction for the material inside, while being stiff along the other directions. A second set of trenches forms flexures to guide the motion along the orthogonal direction. Stacks of piezo elements (blue in Fig. 3.11a) are used to move the flexures. Sometimes a mechanical lever is included

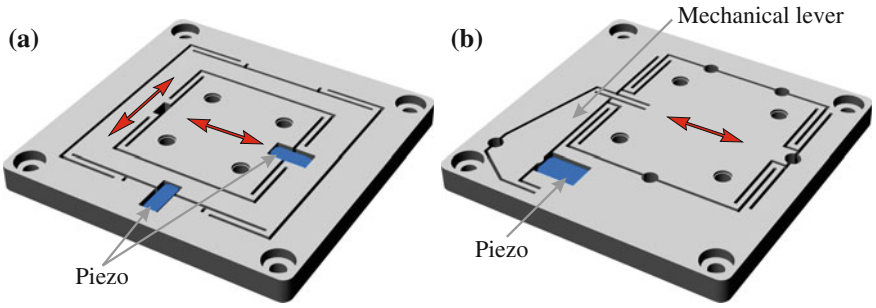

Fig. 3.11 **a** Flexure-guided piezo nanopositioning *xy*-stage. **b** Flexure-guided piezo stage with an integrated mechanical lever amplifying the motion

in the flexures (Fig. 3.11b) in order to amplify the motion ranges up to hundreds of micrometers. Capacitive position sensing detectors can be integrated to allow a precise measurement of the motion. One disadvantage of the flexure-guided piezo nanopositioning stages is that they are relatively large.

3.6 Non-linearities and Hysteresis Effects of Piezoelectric Actuators

The positioning performance of piezoelectric actuators is limited by the effects of hysteresis and non-linearities, which will be discussed in the following.

3.6.1 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. 3.5a. 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 parallel to the outer applied field 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

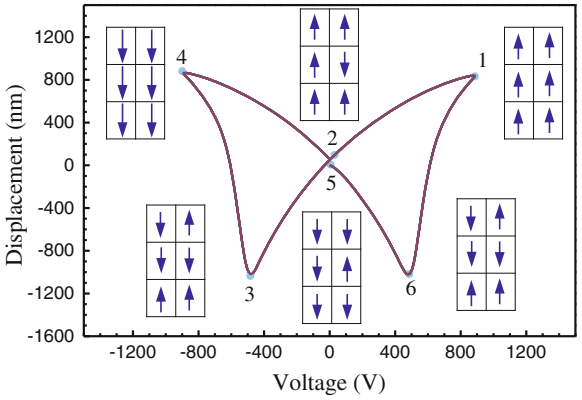

Fig. 3.12 The *butterfly curve* of the piezoelectric material PIC 151 [2] for the applied field and the displacement, both in 3-direction. The strain is shown in dependence of the applied electric field for large electric fields. The corresponding polarization of ferroelectric domains is also indicated in a simplified scheme. The *butterfly curve* shown here was kindly measured by aixACCT [5]

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. 3.12). This figure also shows a schematic sketch of the polarization in the domains. The domains are considered to be square and 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

[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.

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. 3.13, smaller voltage signals which are used for scanning in SPM are shown together with the corresponding displacement. Also here a hysteresis is visible indicated by the elliptic 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 SPM 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

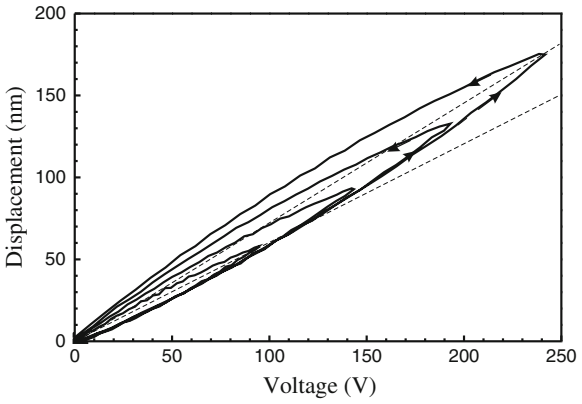

Fig. 3.13 The displacement induced by an applied voltage also shows hysteretic behavior in a range up to 200 V for the applied voltage and the displacement, both in 3-direction. The average piezo constant indicated by the *dashed lines* increases for increasing voltage amplitudes. Due to this the piezo constants and the corresponding displacements can vary by 10–25 %. The curves shown here was kindly measured by aixACCT [5] on a PIC 151 ceramic [2]

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. 3.13 for the two voltage sweeps with smallest and largest amplitudes. The average piezo constant for the smallest and the largest voltage sweeps in Fig. 3.13 differ by about 18 % in this case. This means that due to the effect of hysterisis 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.

3.6.2 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. 3.14. For an ideal piezo actuator without creep the displacement would occur only at the time of the voltage jump and not change afterward.

In SPM, 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

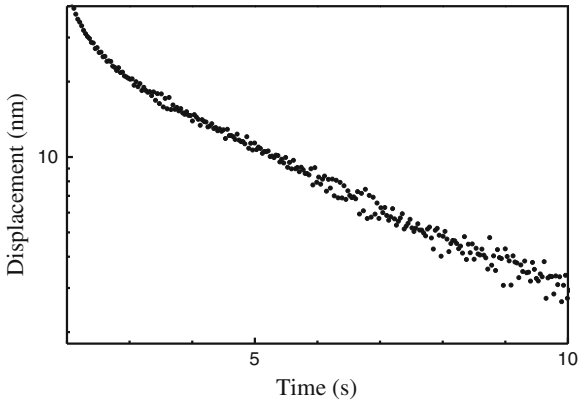

Fig. 3.14 Creep is the piezo displacement after an instantaneous voltage jump. The curve shown here was kindly measured by aixACCT [5] on a PIC 151 ceramic [2]

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.

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.

3.6.3 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.

In conclusion, due to all the above mentioned limitations for piezoelectric scanners, scanning probe techniques are generally not suitable tools for a *quantitative* measurement of distances in the micrometer range (without careful separate calibration). If atomic resolution is achieved the lateral calibration can be performed by taking atomically resolved images of a known surface structure. The vertical calibration is usually performed at (single) monoatomic step edges. If no atomic resolution is obtained, commercially available calibration grids can be used for horizontal and vertical calibration.

An absolute calibration of scanners is also possible using interferometric or capacitive position sensors. In this case, a closed loop operation can be realized. In a feedback loop, the voltage at the piezoelectric actuator is adjusted such that the desired and measured displacement of the actuator is reached. 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 standard in atomic force microscopes.

3.7 STM Tip Preparation

Tip preparation is an important point, which defines the resolution of the scanning tunneling microscope and the quality of the images. The tip should have a minimal radius of curvature at the end and a narrow diameter to penetrate into trenches and pits on the surface. The tip material should be stable in high electric fields.

Tips for STM under ambient conditions are typically made of platinum or a Pt-Ir wire in order to prevent oxidation of the tip material in air. A more or less sharp tip can be produced by cutting and/or grinding. These crude tip preparation techniques are only used for scanning very flat surfaces like graphite. For STM in vacuum, electrochemically etched tungsten tips are most frequently used. The most common procedure of electrochemical etching is the DC drop-off method [6]. A tungsten wire (diameter 0.25 mm) is put into a solution of NaOH (e.g. 5 g NaOH in 50 ml water) and kept at a positive potential towards a stainless steel counter electrode (Fig. 3.15a). The etching process takes place predominately near to the surface of the solution. Due to convection, fresh OH^- is supplied from the air-electrolyte interface. The downward flow of the heavy W anions protects the lower part of the wire in the electrolyte from the supply of fresh OH^-. These specific conditions lead automatically to the formation of a narrow neck shown in Fig. 3.15a. When the neck is etched thin enough the wire fractures due to its weight. Additionally, in order to prevent any further etching, the etching voltage is shut down by the control electronics. The remaining top part will be used as the tip (Fig. 3.15b) and has to be cleaned with deionized water. Most often the tip is covered with an oxide layer and contaminations from the etchant. Thus other in vacuum treatments of the tip, like annealing or field evaporation, are often applied.

There are several different types in situ (in vacuum) tip treatment. Due to the fact that the real sharpness of the tip on the atomic scale cannot be accessed these treatments often have the character of highly empirical procedures. In the following, some examples of further cleaning and characterization in vacuum are given.

Heating. The freshly etched tip is fixed in a special tip-holder and installed into a load-lock chamber for transfer to vacuum. Resistive heating of the tip apex can be performed in order to remove the oxide layer and other contaminations remaining

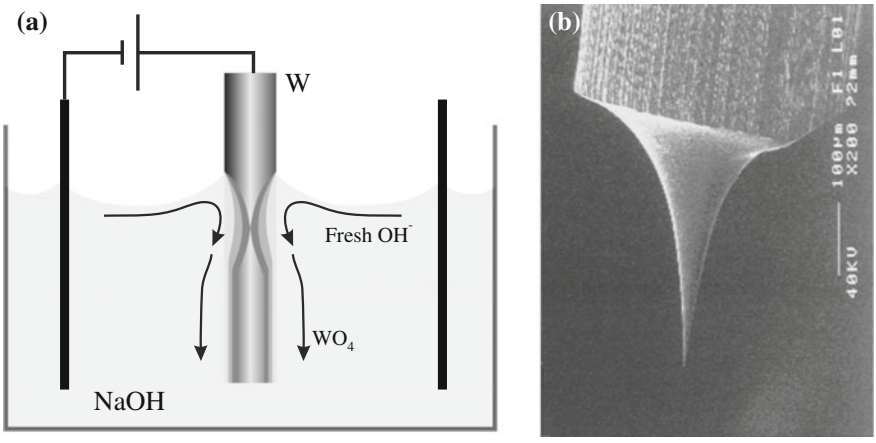

Fig. 3.15 **a** Schematic of electrochemical tip etching. **b** SEM image of an etched tip, original wire diameter 0.25 mm

after the chemical etching [7]. A direct current is applied between the tip and a tungsten wire (diameter 0.5 mm) that touches the tip wire at a point near to the tip apex. The tip should be heated to a temperature above 800 °C for several seconds. The sharpness of the tip is controlled by the value of the applied voltage required in order to achieve a certain field emission current from the apex of the tip. It was found that to obtain an emission current of 1 nA, the applied voltage should not exceed 600 V. If a higher voltage than 600 V is required the tip has a poor sharpness and has to be changed. The pressure during this operation should be less than about 10^{-8} mbar. After this second step of tip preparation, the tip is introduced into the tunneling microscope by the transfer system. Another way of heating the tip is heating by electron bombardment.

Sputtering. Ion bombardment of the tip under vacuum conditions (for instance Ar ions at several hundred volts) can be used to clean and sharpen the tip.

High field treatment. It is also possible to sharpen the tip during tunneling. The bias voltage is raised for a short time (for several scan lines) to several volt (negative at the sample). By this treatment some W atoms may diffuse to the tip apex due to the non-uniform electric field and form a nanotip.

Tip indentation into metal. It is also possible to reshape a blunt tip by indenting (pressing) it several nm into a soft metal sample. In this way a new microtip can be formed. This is also the reason why, when working on metal samples, the tip is rarely replaced.

3.8 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 μm have to be reduced by a factor of 100,000 for the 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?

3.8.1 Isolation of the Microscope from Outer Vibrations

If the microscope is considered as a rigid mass, outside vibrations are transmitted from the ground and the air. An effective vibration isolation can be obtained by

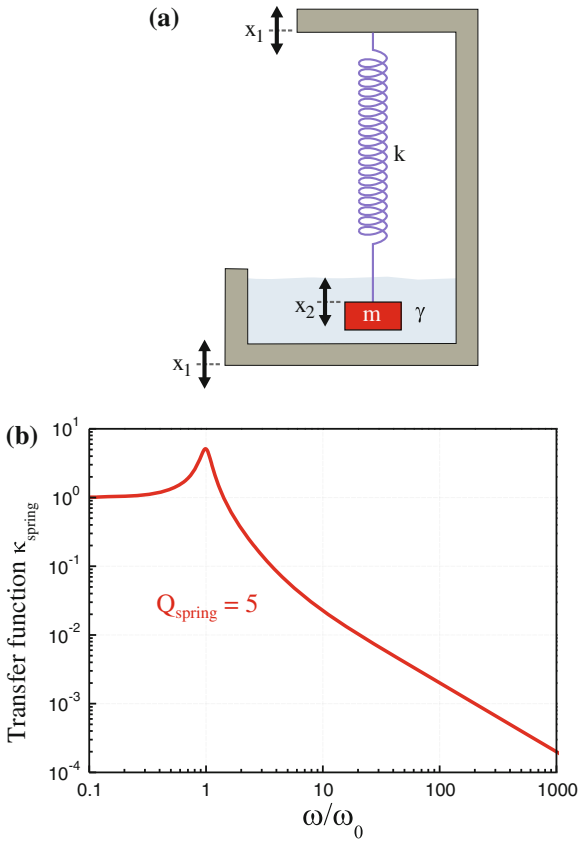

Fig. 3.16 **a** Vibration isolation of a microscope (represented by a mass m) against external vibrations x_1 using a spring suspension. **b** Transfer function of the vibration isolation system for $Q_{\text{spring}} = \omega_0/\gamma = 5$

a spring suspension (Fig. 3.16a). The microscope assembly (mass m) is fixed to a spring with spring constant k. This harmonic oscillator has a natural frequency of $\omega_0 = \sqrt{k/m}$. The oscillating system is damped with a damping factor γ (or the corresponding quality factor $Q_{\text{spring}} = \omega_0/\gamma$). An external (sinusoidal) vibration $x_1(t)$ with amplitude x_1^0 and frequency ω (vibration from of the building floor) is coupled into the system (Fig. 3.16a). As a reaction to this outside forced excitation, the mass m performs an oscillation $x_2(t)$ with amplitude x_2^0 at the driving frequency ω. We refer the motions x_1 and x_2 relative to a fixed (not oscillating) reference system. The elastic force on the mass depends on the *difference* of the positions $(x_2 - x_1)$. Thus the restoring force of the spring acting on the mass m is

$$F_{\text{spring}} = -k(x_2 - x_1), \tag{3.22}$$

In the current case, it is assumed that the frictional damping force depends on the *difference* of the velocities[3] $(\dot{x}_2 - \dot{x}_1)$. Therefore, the damping force F_{frict} is

$$F_{\text{frict}} = \gamma m(\dot{x}_2 - \dot{x}_1). \tag{3.23}$$

The equation of motion for the mass m reads now

$$\ddot{x}_2 + \gamma(\dot{x}_2 - \dot{x}_1) + \omega_0^2(x_2 - x_1) = 0, \tag{3.24}$$

or reordered slightly

$$\ddot{x}_2 + \gamma\dot{x}_2 + \omega_0^2 x_2 = \gamma\dot{x}_1 + \omega_0^2 x_1. \tag{3.25}$$

For a sinusoidal vibration of the frame x_1 can be written in the complex notation (skipping the tilde)

$$x_1(t) = x_1^0 e^{i\omega t}, \tag{3.26}$$

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

$$x_2(t) = x_2^0 e^{i\omega t}. \tag{3.27}$$

with x_1^0 and x_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 $i\omega$)

$$-\omega^2 x_2 + i\gamma\omega x_2 + \omega_0^2 x_2 = i\gamma\omega x_1 + \omega_0^2 x_1. \tag{3.28}$$

or

$$(-\omega^2 + i\gamma\omega + \omega_0^2)x_2^0 e^{i\omega t} = (i\gamma\omega + \omega_0^2)x_1^0 e^{i\omega t}. \tag{3.29}$$

Finally, we obtain

$$\frac{x_2^0}{x_1^0} = \frac{\omega_0^2 + i\gamma\omega}{\omega_0^2 - \omega^2 + i\gamma\omega}. \tag{3.30}$$

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

$$\kappa_{\text{spring}}(\omega) = \frac{|x_2^0|}{|x_1^0|} = \sqrt{\frac{\omega_0^4 + \gamma^2\omega^2}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}}. \tag{3.31}$$

[3] If the damping medium is at rest relative to a fixed external coordinate system, (i.e. not oscillating together with x_1, as assumed here), the term $\dot{x}_1$ has to be neglected in the following. This case applies to a cantilever in atomic force microscopy damped in air.

The response of the system to a driven oscillation $\kappa_{\text{spring}}(\omega)$ can be divided into three regimes (Fig. 3.16b). For $\omega \ll \omega_0$ the outside excitation is transmitted with a transfer function of one, i.e. without any damping. For a frequency close to the natural frequency of the system (in resonance), the outside excitation is even amplified, i.e. the vibrations are increased instead of damped. At $\omega = \omega_0$ the transfer function at becomes

$$\kappa_{\text{spring}}(\omega_0) = \sqrt{\frac{\omega_0^4 + \gamma^2\omega_0^2}{\gamma^2\omega_0^2}} = \sqrt{1 + \frac{\omega_0^2}{\gamma^2}} = \sqrt{1 + Q_{\text{spring}}^2}. \tag{3.32}$$

For small damping ($\gamma \ll \omega_0$ or equivalently $Q_{\text{spring}} \gg 1$), the transfer function can be approximated by

$$\kappa_{\text{spring}}(\omega_0) \approx \frac{\omega_0}{\gamma} = Q_{\text{spring}}. \tag{3.33}$$

If the Q-factor is very large, the external vibration would be amplified tremendously at ω_0. To avoid such resonance excitation, appropriate damping must be applied.

In the third regime $\omega \gg \omega_0$ and γ approaching zero (or correspondingly Q_{spring} very large), the transfer function (3.31) reduces to

$$\kappa_{\text{spring}}(\omega) \approx \left(\frac{\omega_0}{\omega}\right)^2. \tag{3.34}$$

This shows that for excitation frequencies ω much larger than the natural frequency ω_0 and for small damping, the external vibrations are suppressed $\sim 1/\omega^2$. We have seen that damping (small Q-factor or large γ) avoids resonance excitation. However, on the other hand damping deteriorates vibration isolation at higher frequencies. The transfer function becomes asymptotically $\sim 1/\omega$ for $Q_{\text{spring}} = 1$. In Fig. 3.17 the transfer function is shown for different values of Q_{spring}. In typical spring suspen-

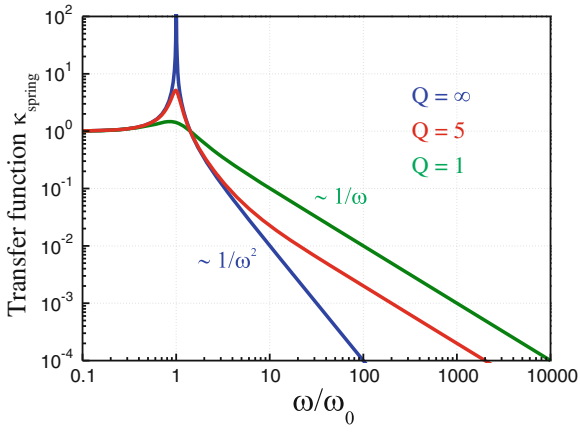

Fig. 3.17 Transfer function of a spring suspension system for different values of the quality factor Q_{spring}

sion systems, a compromise between good damping at high frequencies and large resonance enhancement is chosen for $Q_{\text{spring}} \approx 2 - 5$.

The best vibration isolation (for instance from building vibrations) is achieved with the lowest natural frequency of the spring system. Therefore, the natural 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 $k\Delta l = mg$. If we insert the result for m into the equation for the natural frequency of the system $f_0 = \frac{1}{2\pi}\sqrt{k/m}$ the natural frequency for the system can be written as

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{k}{\Delta l k/g}} = \frac{1}{2\pi}\sqrt{\frac{g}{\Delta l}}. \tag{3.35}$$

To achieve a natural frequency of 1 Hz the spring should be stretched by 25 cm. To achieve a natural 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 natural frequency.

Note that the mass and the spring constant do not enter explicitly into the expression for the natural 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.

3.8.2 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 microscope can also be characterized as a vibrating system with a (quite high) resonance frequency ω_{STM} and a damping constant γ_{STM}, or corresponding quality factor $Q_{\text{STM}} = \omega_{\text{STM}}/\gamma_{\text{STM}}$ (Fig. 3.18). The softest part in the mechanical loop is the piezo material with a typical quality factor of 100. Let x_2 describe the oscillation of the microscope body (or sample in Fig. 3.18a), and x_3 the vibration of the scanner/tip assembly (Fig. 3.18a). Here one point is important (which makes life much easier): it is not the vibration amplitude of the tip x_3 (relative to the floor x_1) that has to be reduced to a minimum but only the *difference* of the motion between tip and sample $x_3 - x_2$. Only the relative motion of the tip relative to the sample counts! The differential equation for the vibrating tip x_3 relative to an external fixed reference is

$$\ddot{x}_3 + \gamma_{\text{STM}}(\dot{x}_3 - \dot{x}_2) + \omega_{\text{STM}}^2(x_3 - x_2) = 0. \tag{3.36}$$

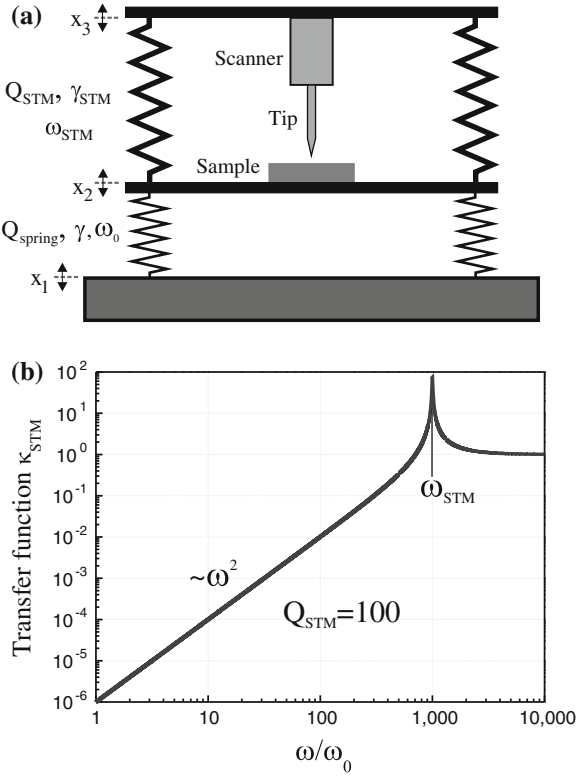

Fig. 3.18 **a** The microscope itself is considered as an oscillating system characterized by ω_{STM} and γ_{STM}. Tip and sample oscillate against each other. **b** Transfer function κ_{STM} according to (3.40) for the microscope with resonance frequency ω_{STM}

The spring force is proportional to $x_3 - x_2$ and the frictional force is proportional to $\dot{x}_3 - \dot{x}_2$. Using the complex method to solve the differential equation results in

$$-\omega^2 x_3 + i\gamma_{\mathrm{STM}}\omega(x_3 - x_2) + \omega_{\mathrm{STM}}^2(x_3 - x_2) = 0, \tag{3.37}$$

or

$$-\omega^2 x_2 - \omega^2(x_3 - x_2) + i\gamma_{\mathrm{STM}}\omega(x_3 - x_2) + \omega_{\mathrm{STM}}^2(x_3 - x_2) = 0. \tag{3.38}$$

The (complex) ratio of the difference of the amplitudes $x_3^0 - x_2^0$ to the amplitude of the base of the microscope x_2^0 is obtained as

$$\frac{x_3^0 - x_2^0}{x_2^0} = \frac{\omega^2}{\omega_{\mathrm{STM}}^2 - \omega^2 + i\gamma_{\mathrm{STM}}\omega}. \tag{3.39}$$

The transfer function results in

$$\kappa_{\mathrm{STM}}(\omega) = \left|\frac{x_3^0 - x_2^0}{x_2^0}\right| = \sqrt{\frac{\omega^4}{(\omega_{\mathrm{STM}}^2 - \omega^2)^2 + \gamma_{\mathrm{STM}}^2\omega^2}}. \tag{3.40}$$

The resulting transfer function is plotted in Fig. 3.18b and can be approximated by

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

for $\omega \ll \omega_{\mathrm{STM}}$, and small damping, with ω_{STM} being the natural frequency of the STM (mechanical loop between tip and sample). When the excitation frequency ω is much lower than the natural frequency of the microscope ω_{STM}, good damping of the external vibrations is achieved. In Fig. 3.18b we use $Q_{\mathrm{STM}} = 100$, 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.

3.8.3 *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 natural frequency for the vibration isolation system and a high natural frequency for the mechanical loop of the microscope. According to (3.31), a vibration of the frame with amplitude $\left|x_1^0\right|$ is transmitted to the STM base with amplitude $\left|x_2^0\right|$ as

$$x_2^0 = \kappa_{\mathrm{spring}} x_1^0. \tag{3.42}$$

(From now on, we consider the amplitudes as real and omit the absolute signs.) Furthermore the vibration amplitude of the STM base x_2^0 induces (according to (3.40)) a relative amplitude between tip and sample of

$$x_3^0 - x_2^0 = \kappa_{\mathrm{STM}} x_2^0. \tag{3.43}$$

In total, an outer vibration of amplitude x_1^0 induces a relative tip sample vibration of amplitude $x_3^0 - x_2^0$ as

$$x_3^0 - x_2^0 = \kappa_{\mathrm{STM}} x_2^0 = \kappa_{\mathrm{STM}} \kappa_{\mathrm{spring}} x_1^0. \tag{3.44}$$

or the total transfer function can be written as

$$\kappa_{\mathrm{total}} = \frac{x_3^0 - x_2^0}{x_1^0} = \kappa_{\mathrm{STM}} \kappa_{\mathrm{spring}}. \tag{3.45}$$

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

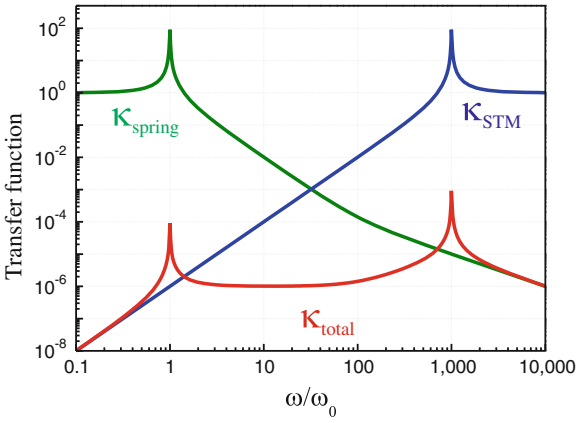

Fig. 3.19 Transfer function of the combined system κ_{total} given by the product of the individual transfer functions of the spring suspension system κ_{spring} and the STM itself κ_{STM} for the case of small damping, i.e. $Q_{\text{STM}} = Q_{\text{spring}} = 100$

According to (3.34) and (3.41), the total transfer function can be approximated in the frequency range $\omega_0 < \omega < \omega_{\text{STM}}$ as

$$\kappa_{\text{total}} \approx \left(\frac{\omega_0}{\omega}\right)^2 \left(\frac{\omega}{\omega_{\text{STM}}}\right)^2 = \left(\frac{\omega_0}{\omega_{\text{STM}}}\right)^2 . \tag{3.46}$$

This behavior of an approximately constant transfer function in between the resonance frequencies ω_0 and ω_{STM} can be seen in Fig. 3.19 in which the transfer function is shown in the limit of negligible damping ($Q_{\text{STM}} = Q_{\text{spring}} = 100$).

If, for example, the natural frequency of the spring suspension system is 1 Hz and the natural frequency of the STM is 1 kHz, the overall transfer function for intermediate frequencies has a constant value of 10^{-6}, as shown in Fig. 3.19. If we would be able to raise the resonance frequency of the STM 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 $Q_{\text{STM}} = 100$. When damping is included the total transfer function is not constant. The total transfer function according to (3.31) and (3.40) is plotted in Fig. 3.20 together with the individual transfer functions of the spring suspension and the STM. It is assumed that the STM mechanical loop can be approximated by a single natural frequency 1,000 times higher than the natural frequency of the spring suspension. With this assumption, the transfer function stays below the initial desired value of 10^{-5} up to $\omega/\omega_0 < 40$. 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 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 natural frequencies of the spring ω_0, while it damps the vibrations at larger frequencies. On the other hand, the STM assembly acts as a high-pass for

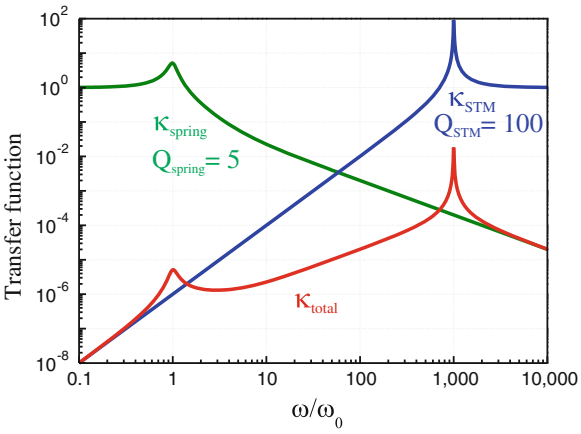

Fig. 3.20 Transfer function of the combined system κ_{total} which is the product of the individual transfer functions of the spring suspension system κ_{spring} and the STM itself κ_{STM}

vibrations with a frequency larger than ω_{STM}, while it damps the vibrations at lower frequencies. The total transfer function is the product of the transfer functions of the spring suspension and STM. In order to keep the total transfer function low at all frequencies, a low natural 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. 3.21a. An example of an eddy-current damping system is

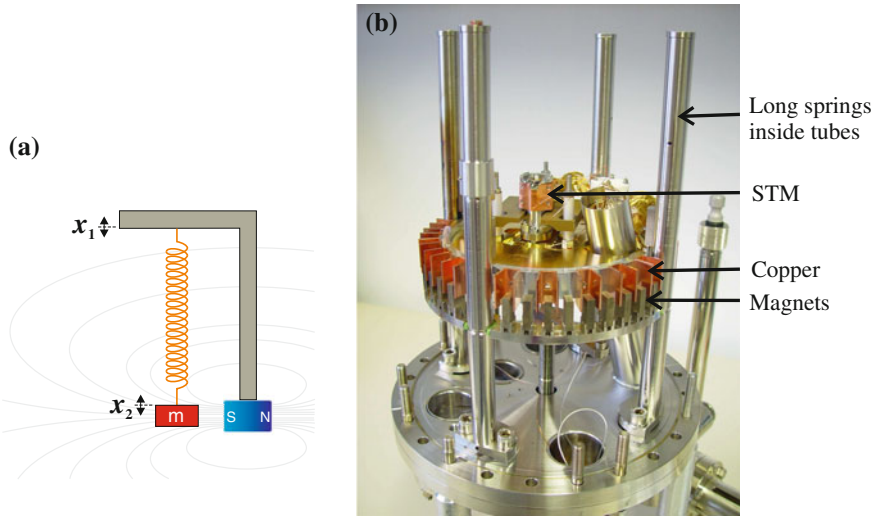

Fig. 3.21 **a** Principle of an eddy-current damping system with a magnet next to a conductor in which the energy is dissipated as eddy currents. **b** Photo of an eddy-current damping system with STM

shown in Fig. 3.21b. 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.

3.9 Building Vibrations

Building vibrations are most pronounced in the low frequency range below 10 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.

Geophones (accelerometers) are typically used to measure building vibrations. The quantity measured by these instruments is the velocity. In Fig. 3.22, the velocity 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 amplitudes are typically observed for low frequencies around 1–2 Hz. In Fig. 3.22 a value of $v_0 \approx 0.7\,\mu$m/s is observed at low frequencies. In order to convert the measured data from the velocity to oscillation amplitude or acceleration, we recall that

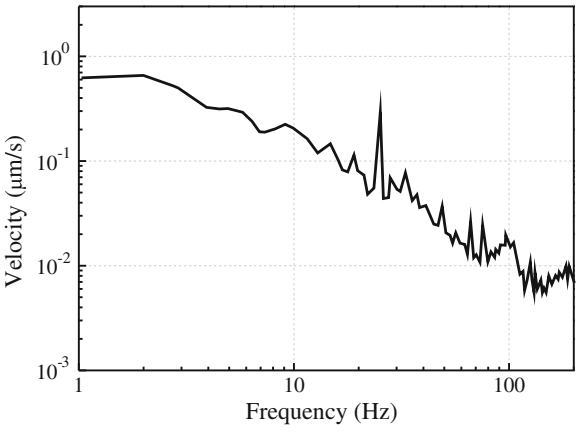

Fig. 3.22 Velocity of the building vibrations measured on the floor in a building at the Research Center in Jülich

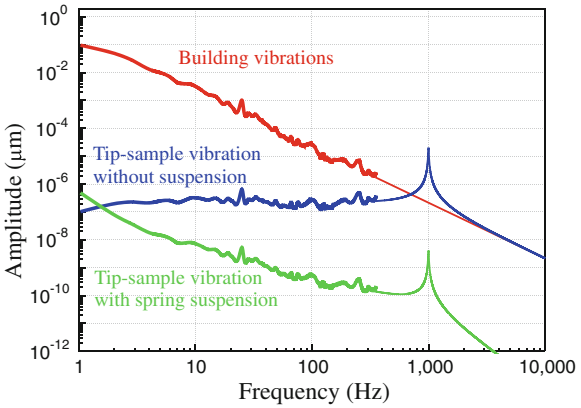

Fig. 3.23 Expected tip-sample vibrational amplitude as a function of frequency, calculated using the measured building vibrations and the appropriate transfer function from Fig. 3.20. The amplitude of the building vibrations is shown as a *red line*. The data are taken from Fig. 3.22 and extrapolated for higher frequencies. The *green* and *blue curves* show the behavior with and without a spring suspension system, respectively

$$x = x_0 \cos(\omega t), \tag{3.47}$$

$$v = \dot{x} = -x_0 \omega \sin(\omega t) := -v_0 \sin(\omega t), \tag{3.48}$$

$$a = \ddot{x} = -x_0 \omega^2 \cos(\omega t). \tag{3.49}$$

Therefore, the vibration amplitude at 2 Hz is $x_0 = v_0/\omega \approx 50\,\text{nm}$. The corresponding acceleration is $a_0 = \omega v_0 \approx 10^{-5}\,\text{m/s}^2 \approx 1\,\mu\text{g}$.[4]

The measured building vibrations $x_1^0(\omega)$ can be included in the vibration analysis performed previously. According to (3.44), the relevant tip-sample vibrational amplitude $x_3^0 - x_2^0$ can be expressed as a function of frequency as

$$x_3^0 - x_2^0 = \kappa_{\text{total}}(\omega) x_1^0(\omega). \tag{3.50}$$

If we multiply the total transfer function by the measured floor vibration amplitude (derived from Fig. 3.22), the expected tip-sample vibration amplitude arising due to the floor vibrations is shown in Fig. 3.23. The case where no spring suspension is invoked is shown as blue line, leading to a roughly constant tip-sample vibration amplitude of $10^{-4}\,\text{nm} = 0.1\,\text{pm}$. However, close to the resonance frequency of the STM the amplitude increases by the usually quite high quality factor of the STM. This disadvantageous resonance behavior (amplitude up to 0.1 nm) can be suppressed using a spring suspension system. The tip-sample vibrational amplitude including a spring suspension (green curve) suppresses the amplitude at STM resonance

[4] Sometimes a factor of $1/\sqrt{2}$ is included if the root mean square (RMS) amplitude instead of the peak amplitude is measured.

frequency, but also leads to a resonance at the eigenfrequency of the spring suspension system, which has to be suppressed by proper damping of the spring suspension system. In this case, the tip-sample vibrations are reduced to values below one picometer for all frequencies.

3.10 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 STM tips can be fabricated by self-adjusting electrochemical etching.
- The natural frequency of a spring suspension system depends only on the extension length Δl as $\omega_0 = \sqrt{\frac{g}{\Delta l}}$.
- 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 natural frequency of the spring suspension system ω_0 is combined with a high natural frequency of the STM assembly ω_{STM}, 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 $\kappa_{\mathrm{total}} \approx (\frac{\omega_0}{\omega_{\mathrm{STM}}})^2$ for $\omega_0 < \omega < \omega_{\mathrm{STM}}$.
- The expected tip-sample vibration amplitude can be calculated by multiplying the total transfer function by the (measured) building vibration amplitude.

Chapter 4
Scanning Probe Microscopy Designs

Due to the limited range of piezo actuator elements available of only one to several micrometers, it is necessary to use a coarse approach to bring tip and sample into such a close distance that the (tube) scanner can be used for the fine motion (up to several micrometers) during scanning. The task of coarse positioning largely determines the SPM design since nowadays almost all SPMs use a tube scanner for the fine motion. Here we concentrate on the general principles of SPM design and take the STM as an example. Specific aspects concerning atomic force microscopy designs will be discussed later.

Since the early days of STM, screw mechanisms are often used for the coarse positioning of the tip. Sometimes these mechanisms are combined with lever mechanisms in order to reduce the tip-sample motion relative to the screw motion. Nowadays, this macromechanical positioning is often replaced by micromechanical positioning using piezo electric actuators. We will describe several types of these most precise piezo electrically driven nanopositioners.

4.1 Nanoscope

The first commercial atomic force microscope, the nanoscope, which is still available in a modified form, works with a mechanical coarse approach which can be driven by a stepper motor [8]. Its scan head consists of an invar cylinder which houses a tube piezo (Fig. 4.1). Invar is used because it has a similar thermal expansion coefficient, to that of the piezo material. A tube scanner 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 tip (lower part) is used to control the vertical tip-sample motion (z-direction), while the upper part is segmented into four quadrants, which allows lateral motion (xy-scanning) (Fig. 4.1). The z-extension part of the tube piezo element acts as a lever to enhance the lateral motion. The tip is attached to the side of the tube scanner. The preamplifier for the tunneling current is located very close to the tip on top of the invar cylinder. The sample is mounted on a baseplate which supports the

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_4

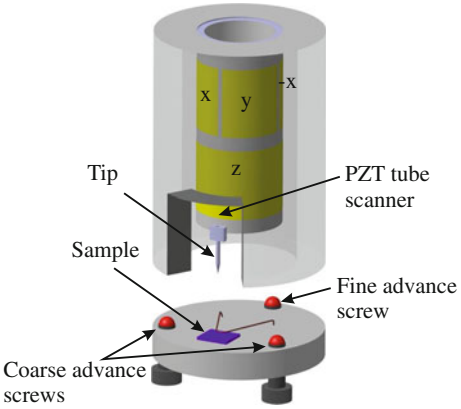

Fig. 4.1 Design principle of a Nanosope STM [8]

cylindrical head. The scan head rests on the three hemispherical ends of the screws. Two of these supporting points form a pivot close to the tip. The third support point is moved by a fine advance screw. In this way the movement of the tip relative to the sample is reduced substantially relative to the travel of the fine advance screw. The motion of the fine advance screw can be controlled by a stepper motor which allows the approach of the tip towards the sample to be automated. An additional xy-translation stage built into the base allows macroscopically different regions of the sample to be imaged.

4.2 Inertial Sliders

While the question of the best SPM design cannot be answered generally because it depends on the specific application, inertial sliders are very common in SPM designs. 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 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 its inertia 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. This very general configuration of an inertial slider is shown in Fig. 4.2a. 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.

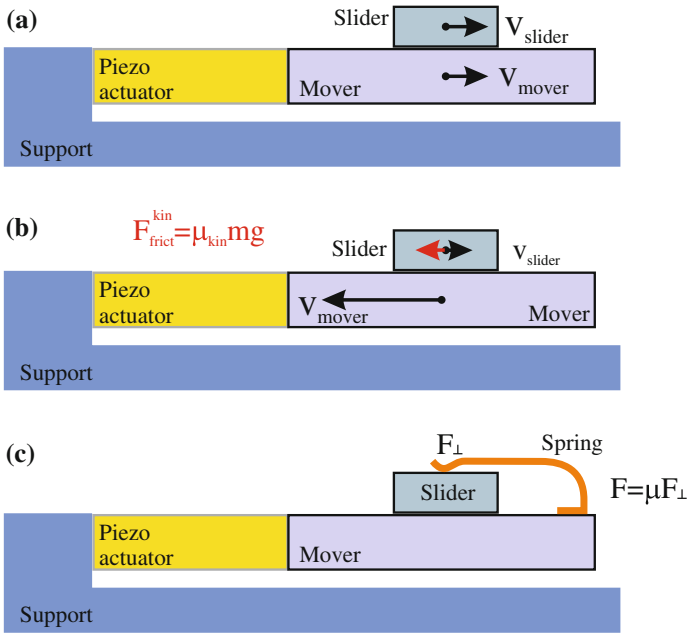

Fig. 4.2 Operating principle of an inertial slider. **a** Riding phase: $ma_{mover} < F^{stat}_{frict} = m_{stat}\, mg$. **b** Sliding phase: $ma_{mover} \geq F^{stat}_{frict}$. **c** Inertial slider with spring

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. 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^{stat}_{frict} is larger than the force on the slider due to its acceleration as

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

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^{stat}_{frict} are applied. If however, $m\, a_{mover} > F^{stat}_{frict}$ the frictional contact disengages, transforms to a sliding frictional contact and the slider will not move together with the mover (Fig. 4.2b). The necessary accelerations larger than

g can be 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

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

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

In Fig. 4.3 the position, the velocity and the acceleration of the mover and slider relative to an external reference are shown during the "riding phase" and "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 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 $a_{\text{frict}}^{\text{stat}}$, and the slider stays in frictional engagement with the mover. During the riding

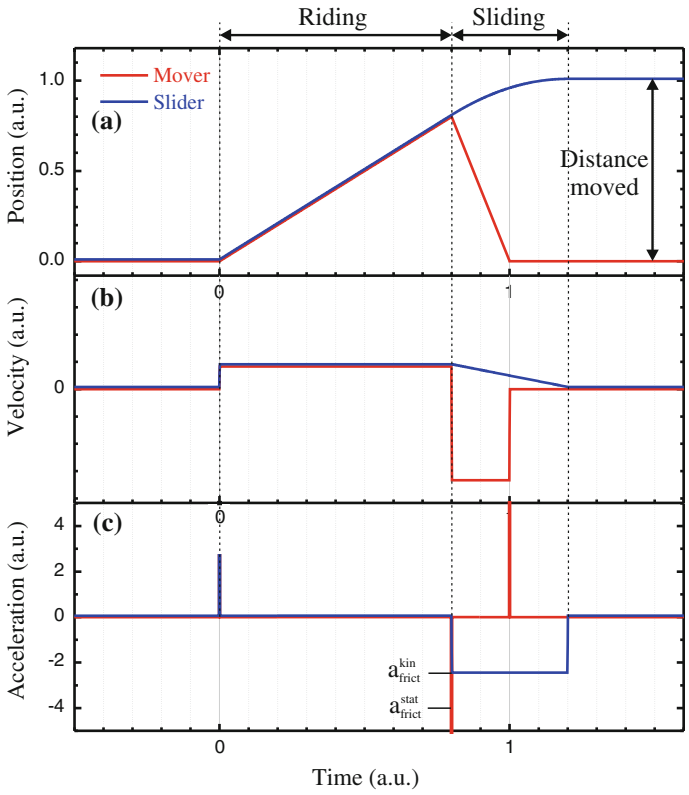

Fig. 4.3 Position, velocity and acceleration of the mover and the slider during inertial motion as a function of time relative to an external fixed reference system

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. 4.3c) 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 $a_{\text{frict}}^{\text{stat}}$, 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. 4.3a.

When the velocity of the mover stops (at time 1 in Fig. 4.3) there is another sharp (this time positive) spike in the acceleration of the mover. The slider continues to decelerate 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 slider has moved relative to the mover by a certain distance as indicated in Fig. 4.3a. In reality, the sliding phase occurs in a much shorter time relative to the riding phase than shown in Fig. 4.3. 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). 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 springs or magnets as shown in Fig. 4.2c. 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. 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 nanopositoner is shown in Fig. 4.4 [9]. 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, (hemi)spherical 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.

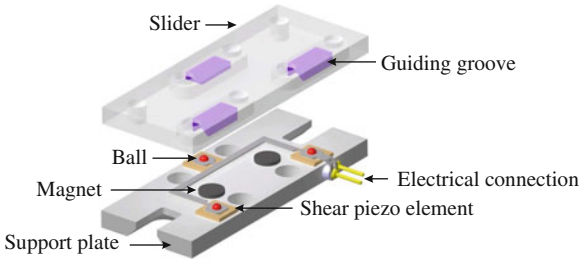

Fig. 4.4 Sketch of an inertial slider (length 35 mm)

4.3 Beetle STM

A widely used versatile STM design invented in Jülich by Karl Besocke is the beetle design [10]. The beetle STM as shown in Fig. 4.5a consists of a baseplate on which three piezoelectric tubes are mounted. Balls are fixed at the end of the tubes. A ramp ring divided into three helical sectors rests on the three balls. The sample is mounted in the middle of the ramp ring. A fourth piezoelectric tube is mounted at the center of the baseplate and acts as a tube scanner for xyz fine motion and a tip (holder) is fixed to the scanner piezo. By applying an appropriate synchronous saw-tooth-like voltage to the outer three segmented piezos, the ramp ring can be rotated by inertial motion. Changing the voltage quickly (saw-tooth signal) causes the balls to slip on the ramp due to the inertia of the ramp ring. Thus the sample can be moved towards the tip by rotation of the ramp ring. Also a horizontal motion of the ramp ring can be induced by an appropriate inertial motion.

The beetle design has several advantages. One is the compactness of the design which allows an SPM to be constructed with a very small mechanical loop from the sample via the outer three piezos to the baseplate and via the scanner tube and the tip back to the sample. This compactness results in high resonance frequencies and correspondingly small vibrational noise. Another advantage is that the symmetric design reduces thermal drift. A thermal expansion in the scanner piezo tube is compensated by a similar expansion in the outer piezos (same material). Also an easy xy-motion of the sample over several millimeters on the ramp rings is one feature

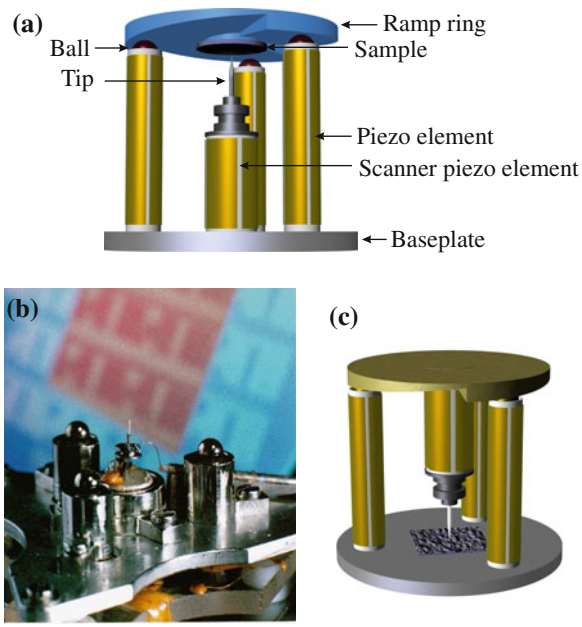

Fig. 4.5 **a** Sketch of the beetle STM design. The height can be as small as about 20 mm. **b** Photo of a beetle STM with shielded piezo tubes and without sample on *top*. **c** Variant of the beetle design with the scanner tube mounted to the ramp ring

of the beetle design. One disadvantage of the original beetle design was that heavy (metal) samples, which also need thick leads for heating are not so well accommodated on top of the SPM in the ramp ring. However, the beetle design is very flexible. It can be turned upside down, so that the ramp ring and the sample are fixed and the microscope "walks" on the ramp ring. This type is also called "Johnny walker" design. There are even more variants, for instance (starting from the original design according to Fig. 4.5a) the tube scanner can be turned upside down and fixed to the ramp ring (instead of the sample), as shown in Fig. 4.5c. The sample is mounted on (or below) the baseplate. One disadvantage of the beetle design is the coupling of xy-motion and z-motion. If the ramp ring is rotated downwards for the approach (z-motion) this motion is always accompanied by a slight unintentional shift also in the xy-direction and vice versa. Another disadvantage of the beetle design may be that due to the rotation of the ramp ring there is no fixed reference frame for the xy-directions.

4.4 Pan Slider

The Pan slider is an STM design with very high rigidity which is mainly used in vacuum and cryogenic environments [11]. This STM type is named after Shuheng Pan, who invented the design. The moving part is a sapphire prism containing a tube piezo scanner. The stepping is actuated by six shear piezo stacks, as shown in Fig. 4.6. 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 also 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.

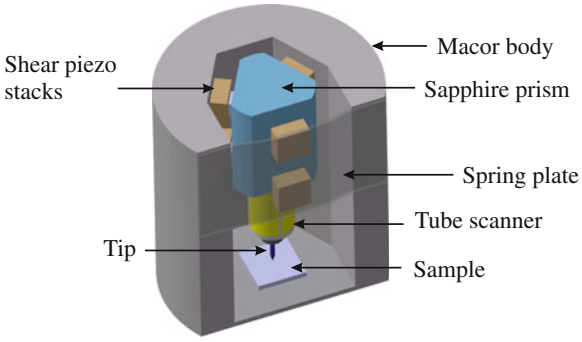

Fig. 4.6 Pan STM design using shear piezo elements in order to move a sapphire prism on which a tube scanner is mounted

4.5 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* the piezo tube (which is used as scanner in almost all SPMs). 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.

The task of the KoalaDrive nanopositioner is to move a rod along its axis, as shown in Fig. 4.7a. For use in an STM a tip (holder) is fixed to the end of the rod. The KoalaDrive consists of two tube piezo elements mounted in series, for instance, one after the other, as shown in Fig. 4.7a. At the ends and between the two tube piezos, three springs are mounted, holding a central rod. The upper two springs shown in Fig. 4.7a 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

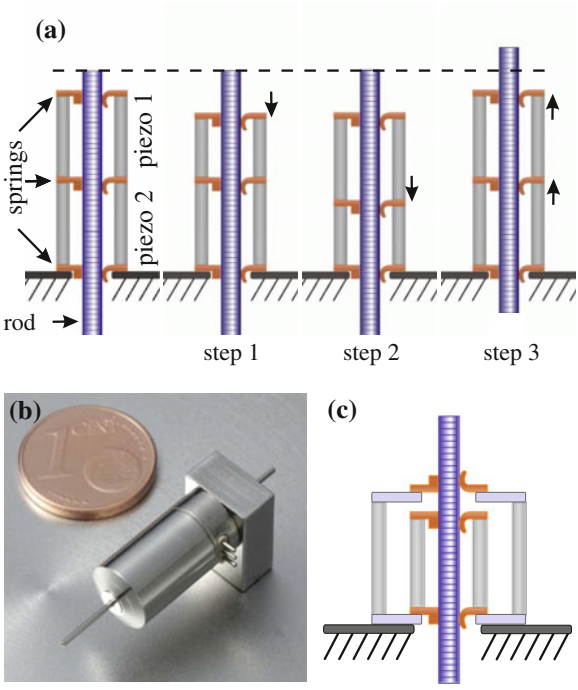

Fig. 4.7 **a** Working principle of the KoalaDrive: concerted interplay between static friction and sliding friction. If only one spring moves, the rod is held stationary by the other two (*step 1* and *step 2*). The motion of the springs during the different steps of a cycle is indicated by *arrows*. If two springs move simultaneously, the central rod moves together with them (*step 3*). **b** Photo of the KoalaDrive. **c** A variant of the KoalaDrive design where the two piezoelectric tubes are coaxially stacked into each other

moving spring will the frictional engagement be lifted and sliding friction will occur. One cycle of motion is shown in Fig. 4.7a. 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. Figure. 4.7b shows a photo of a KoalaDrive. The ultracompact KoalaDrive can have a diameter of less than 2.5 mm and a length smaller than 10 mm. Depending on the particular application, the design of the KoalaDrive can be modified. If, for instance, the length of the drive should be small, the two piezo tubes can alternatively be placed coaxially into each other instead of one after the other, as can be seen in Fig. 4.7c.

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 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. Avoiding steep slope signals also means fewer demands on the power supply (no high slew rate needed) and for the cabling (no high currents flow). 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 scanning tunneling microscope (STM) tip attached to the central rod. The KoalaDrive is ultra high vacuum compatible and works at cryogenic temperatures (down to liquid helium temperatures), as well as in magnetic fields.

In the next step, the KoalaDrive can be used to build an ultracompact STM. 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. 4.8a. The STM is completed by attaching a tip (plus tip holder) to the central rod and an outer frame, which holds the sample, as shown in Fig. 4.8a. 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 STM of minimal size: A complete STM scanner can be integrated inside a piezo tube of 6 mm outer diameter and 12 mm length. In Fig. 4.7b, a photograph of an actual KoalaDrive STM is shown. The use of the KoalaDrive makes the scanning probe microscopy design ultracompact and leads accordingly to high mechanical stability.

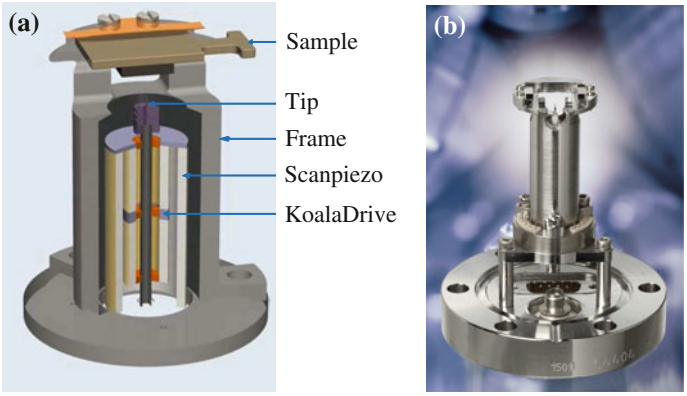

Fig. 4.8 **a** Design of an STM using the KoalaDrive leading to an STM of minimal size. **b** Photograph of an actual KoalaDrive STM

4.6 Tip Exchange

Unfortunately, an initially sharp STM tip degrades when used for some time. If the tip is used under ambient conditions it can be replaced straightforwardly by bending the tip wire slightly and inserting it into the cannula of a syringe. In the simplest case, this syringe needle is already part of the STM and attached in or at the scan piezo element. When working in vacuum the (tungsten) wire itself cannot be handled, but the tip wire is mounted in a tip holder, which can be transferred into vacuum and grabbed there, usually by a modified sample holder adapted to hold a tip holder. Finally, the tip holder (with tip) is inserted into the STM in vacuum using a wobble stick or another kind of manipulator. The easiest way of inserting a tip holder into an STM is if the receptacle at the STM includes a small magnet which guides the tip holder (made of magnetic material) to its desired position. Often a fork mechanism (or gripper mechanism) is used to release the tip holder from the manipulator when it is in position in the STM. Instead of magnetic forces also a spring mechanism can be used to fix a tip holder in the STM.

4.7 Summary

- Coarse approach is the approach between the tip and sample from the macroscopic range down to the range covered by the tube scanner. The coarse approach determines the SPM design.
- Inertial sliders are actuated by a saw-tooth 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.

- In the Beetle SPM, the rotation of an inclined ramp ring by inertial motion leads to an approach between tip and sample.
- In the Pan SPM design, the linear motion of a sapphire prism by inertial motion leads to an approach between tip and sample.
- The KoalaDrive nanopositioner avoids inertial motion, but uses alternating movements of three springs holding a central rod, which in turn holds the STM tip. The operating principle provides smooth travel and avoids the shaking which is intrinsically present if nanopositioners based on inertial motion with saw-tooth driving signals are used. The KoalaDrive used as a coarse approach can be integrated into an xyz-scanning piezo element and results in an ultracompact STM design of high mechanical stability.

Chapter 5
Electronics for Scanning Probe Microscopy

First we discuss some fundamental issues of electronics, such as voltage divider, low-pass filter, and operational amplifier. Then we continue to discuss topics more closely related to scanning probe microscopy such as the current amplifier in scanning tunneling microscopy and feedback electronics, which in SPM 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 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

$$\begin{aligned} V_{\text{in}} &= V_1 + V_2 = I(R_1 + R_2) && \text{(Kirchhoff's voltage law)} \\ V_2 &= R_2 I = V_{\text{out}} && \text{(Ohm's law)} \end{aligned} \tag{5.1}$$

These equations can be solved for

$$\frac{V_{\text{out}}}{V_{\text{in}}} = H = \frac{R_2}{R_1 + R_2}. \tag{5.2}$$

The output voltage divided by the input voltage is the called transfer function H. 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 $H \approx 1$ for $R_1 \ll R_2$ and $H \approx R_2/R_1$ for $R_1 \gg R_2$.

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_5

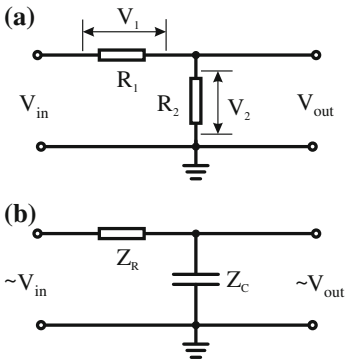

Fig. 5.1 **a** Circuit scheme of a voltage divider. The transfer function is given by $H = V_{\text{out}}/V_{\text{in}} = R_2/(R_1 + R_2)$. **b** This circuit is also a voltage divider, however, now R_2 is replaced by a capacitor and an AC input voltage is considered. Thus, we use the complex impedances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

5.2 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

$$V = V_0 e^{i(\omega t+\varphi_V)}, \quad \text{and} \quad I = I_0 e^{i(\omega t+\varphi_I)}. \tag{5.3}$$

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

$$\begin{aligned} Z_{\text{C}} &= \tfrac{1}{i\omega C} && \text{for a capacity } C, \text{ and} \\ Z_{\text{L}} &= i\omega L && \text{for a inductance } L, \text{ and of course} \\ Z_{\text{R}} &= R && \text{for a resistor } R. \end{aligned} \tag{5.4}$$

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 (5.2) as

$$H(\omega) = \frac{V_{\text{out}}}{V_{\text{in}}} = \frac{Z_C I}{(Z_R + Z_C)I} = \frac{\frac{1}{i\omega C}}{R + \frac{1}{i\omega C}} = \frac{1}{1 + i\omega RC}. \tag{5.5}$$

The transfer function is a complex quantity. In the Bode diagram, 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. 5.2a. The corresponding equations are

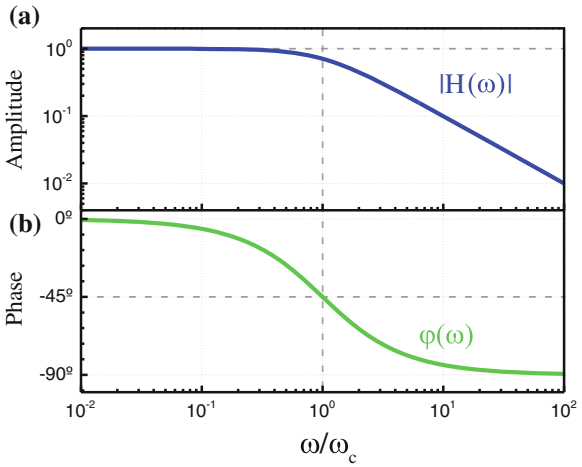

Fig. 5.2 The Bode plot shows the absolute value of the complex transfer function (gain) (**a**) and the phase shift of the output relative to the input signal (**b**). The figure shows the Bode plot of the circuit in Fig. 5.1b. The behavior of the absolute value of the transfer function (amplitude) approaches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

$$|H(\omega)| = \frac{|V_{\text{out}}|}{|V_{\text{in}}|} = \frac{1}{\sqrt{1+\omega^2 R^2 C^2}}, \quad \text{and} \quad \varphi_V = \arctan(-\omega RC). \tag{5.6}$$

For frequencies lower than the corner frequency $\omega_c = 1/(RC)$, the absolute value of the transfer function approaches unity, i.e. gain $|V_{\text{out}}| / |V_{\text{in}}|$ 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 $\omega_c = 1/(RC)$.

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 goes to $-90°$ 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 periodic functions via Fourier methods.

5.3 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 (5.2) the output voltage V_{out} can be written as

$$V_{\text{out}} = V_{\text{signal}} \frac{R_{\text{device}}}{R_{\text{signal}} + R_{\text{device}}}, \tag{5.7}$$

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

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 so called impedance converter is used, which we discuss in Sect. 5.5.1 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 $R_{\text{input}} \gg R_{\text{output}}$.

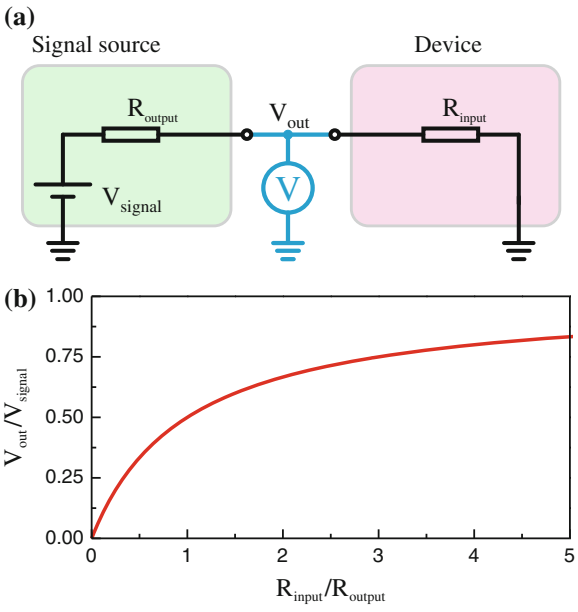

Fig. 5.3 **a** Signal source, consisting of an ideal voltaget source V_{signal} and an output resistance R_{output}, connected to a device with an input resistance, characterized by the resistance between the input and the ground R_{input}. **b** The output voltage for this circuit approaches V_{signal}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

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 later.

5.4 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

$$\langle V \rangle = \lim_{T\to\infty} \frac{1}{T} \int_0^T V(t)dt. \tag{5.8}$$

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

$$\sqrt{\langle \Delta V^2 \rangle} = \sqrt{\lim_{T\to\infty} \frac{1}{T} \int_0^T (V(t) - \langle V \rangle)^2 dt}. \tag{5.9}$$

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 $\Delta V(t)$. In the following, we will consider the frequency dependence of the noise. The frequency dependence of the noise can be characterized by the power spectral density (PSD)[1] $N_V^2(\omega)$. An important property of the power spectral density of the noise is that it relates to the mean square noise as

$$\left\langle \Delta V^2 \right\rangle = \int_0^\infty N_V^2(f)df = \frac{1}{2\pi} \int_0^\infty N_V^2(\omega)d\omega. \tag{5.10}$$

[1] The power spectral density of the noise $\Delta V(t)$ can be defined via the Fourier transform of the noise as $N_V^2(\omega) = \lim_{T\to\infty} \frac{1}{2\pi T} \left| \int_0^T \Delta V(t) e^{-i\omega t} dt \right|^2$.

If a detection scheme is used which measures the noise variable only within a certain (angular frequency) bandwidth $B_\omega = \omega_2 - \omega_1$ between ω_1 and ω_2, the mean square noise can be written as

$$\left\langle \Delta V^2 \right\rangle = \frac{1}{2\pi} \int_{\omega_1}^{\omega_2} N_V^2(\omega) d\omega. \tag{5.11}$$

This expression can also be considered as defining the noise power spectral density $N_V^2(\omega)$. The noise PSD indicates how much power the noise signal carries in a small region around ω. The noise amplitude spectral density is defined as $N_V = \sqrt{N_V^2}$. If the noise spectral density is constant between ω_1 and ω_2 and zero outside, (5.11) reduces to

$$\left\langle \Delta V^2 \right\rangle = \frac{1}{2\pi} (\omega_2 - \omega_1) N_V^2(\omega), \tag{5.12}$$

and we obtain

$$\sqrt{\left\langle \Delta V^2 \right\rangle} = N_V \frac{1}{\sqrt{2\pi}} \sqrt{B_\omega}. \tag{5.13}$$

The (constant) noise spectral amplitude density of the noise variable ΔV is expressed in the unit of the noise variable per $\sqrt{\text{rad} \cdot \text{Hz}}$, for instance volt/$\sqrt{\text{Hz}}$. The actual RMS value of the noise variable measured with a specific bandwidth is then given by the noise amplitude spectral density times the square root of the bandwidth. Note that the angular frequency bandwidth $B_\omega = \omega_2 - \omega_1$ is defined as angular frequency, i.e. in units of rad/s, not cycles/s. Similarly, the unit of the noise power spectral density $N_V^2(\omega)$ is volt^2/(rad $\cdot$ Hz). If the natural frequency f is considered, (5.13) reads

$$\sqrt{\left\langle \Delta V^2 \right\rangle} = N_V \sqrt{B}, \tag{5.14}$$

with $B = f_2 - f_1$ and N_V in volt/$\sqrt{\text{Hz}}$.

5.5 Operational Amplifiers

Since operational amplifiers are used in several parts of STM 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 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).

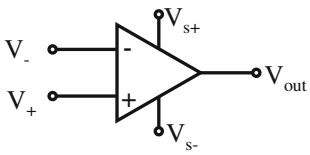

Fig. 5.4 Block diagram of an operational amplifier showing the supply voltages V_s, the input voltages $V_\pm$ and the output voltage V_{out}

We will show that if these properties of an operational amplifier are met the characteristics of the amplifier are determined by the feedback 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

$$V_{\text{out}} = G(V_+ - V_-). \tag{5.15}$$

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.

5.5.1 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 (5.15) 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 (5.15) reads

$$V_{\text{out}} = G(V_{\text{in}} - V_{\text{out}}), \tag{5.16}$$

which leads to

$$V_{\text{out}} = V_{\text{in}} \frac{G}{1+G}. \tag{5.17}$$

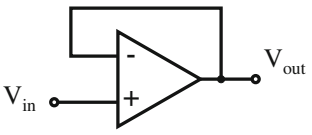

Fig. 5.5 Operational amplifier wired as a voltage follower. A negative feedback is realized by connecting the output to the negative (inverting) input

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 $= V_{-}$). One could think 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 "one-way" device for signals, 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.

5.5.2 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 (5.15) 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[2]) 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

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

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

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

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

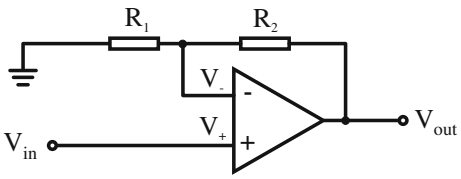

Fig. 5.6 Operation principle of non-inverting amplifier

If G is very large the gain becomes

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{1}{K} = 1 + \frac{R_2}{R_1}. \tag{5.20}$$

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 (5.15) 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. In the previous two circuits the difference between the inputs V_+ and V_- approached 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.

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

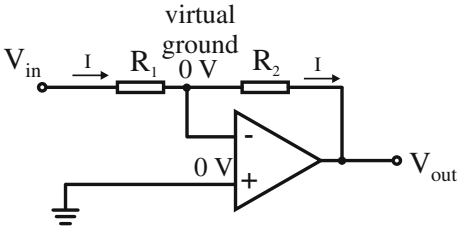

Fig. 5.7 Circuit of an inverting amplifier realized with an operational amplifier

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{-I\,R_2}{I\,R_1} = -\frac{R_2}{R_1}. \tag{5.21}$$

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.

5.6 Current Amplifier

The tunneling current in STM has very small values, typically 0.01–10 nA. The current amplifier is an essential element of an STM since it amplifies the current and converts it to a voltage. Such amplifiers are 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 $I_{\text{in}} = V_{\text{in}}/R_1$ 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, $I_{\text{in}} = I_{\text{FB}} = -V_{\text{out}}/R_{\text{FB}}$. Or

$$V_{\text{out}} = -I_{\text{in}}\,R_{\text{FB}}. \tag{5.22}$$

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 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 $Z_{\text{R}} = R$, and $Z_{\text{C}} = 1/(i\omega C)$, respectively. Since the two impedances in the feedback arm of

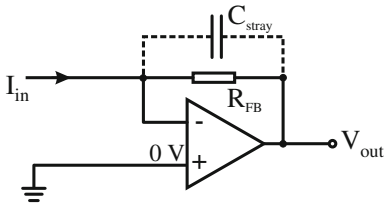

Fig. 5.8 Circuit used as current amplifier in STM. The gain (actually transconductance in V/A) is proportional to the resistance of the feedback resistor R_{FB}. The bandwidth of this current amplifier is limited by the stray capacitance C_{stray}

the operational amplifier are in parallel, the following expression results for the total (complex) impedance Z as

$$\frac{1}{Z} = \frac{1}{Z_R} + \frac{1}{Z_C} = \frac{1}{R} + i\omega C. \tag{5.23}$$

The absolute value of the complex impedance results as

$$|Z| = \frac{R}{\sqrt{1 + (\omega RC)^2}}. \tag{5.24}$$

Replacing according to (5.22) $V_{out} = -Z I_{in}$, and identifying R with R_{FB}, as well as $C = C_{stray}$ results in

$$V_{out} = \frac{-I_{in} R_{FB}}{\sqrt{1 + \left(\omega R_{FB} C_{stray}\right)^2}}. \tag{5.25}$$

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 $f_{corner} = 1/\left(2\pi R_{FB} C_{stray}\right)$. 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 without significant loss of the signal (i.e. from DC to $f_{corner} \sim 1/(2\pi R_{FB} C_{stray})$. It can be seen that the gain which is proportional to R_{FB} and the bandwidth proportional to $1/R_{FB}$ 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 [12, 13] and can be calculated as

$$I_{noise} = \sqrt{\frac{4 k_B T B}{R_{FB}}}. \tag{5.26}$$

with B being the bandwidth and k_B the Boltzmann constant. In Table 5.1 some numerical values are given.

Table 5.1 Gain, bandwidth and noise for a current amplifier with $R_{FB} = 100\,\text{MΩ}$ and $R_{FB} = 1\,\text{GΩ}$

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

5.7 Feedback Controller

In scanning probe microscopy, a feedback controller is used to follow the surface topography. Before we come to the application of a feedback controller to SPM, 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 is measured constantly by a sensor, and compared to the setpoint w by subtraction $w - x$. Depending on this error signal, 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 set-point value w. This whole operation of the controller acts in a closed feedback loop as shown in Fig. 5.9. The control loop fulfills the task of adapting the system output to the setpoint in the presence of disturbing external noise.

Before we turn to the feedback controller of the STM, let us consider (as an example) a simpler system: 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 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 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 smarter way to keep x as close as possible to w. There are two kinds of time delays in the feedback loop: First the time delay in the system itself (this delay is large for the case of the heating of the house and much smaller in the case of STM). For simplicity we will not consider this time delay in the following. Secondly, there is a time delay due to the controller, which we will consider in the following.

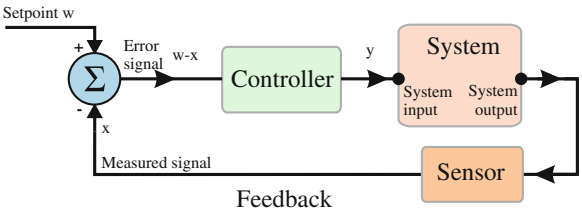

Fig. 5.9 A general model for a feedback loop

5.7.1 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 $w - x$, as

$$y = K_P(w - x). \tag{5.27}$$

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, which is only limited by maximum heating power of the heater). Since the output of the controller is instantaneously proportional to the error signal, the P controller is a fast reacting type of controller.

One problem with the proportional controller is that a pure proportional control will not settle at the set-point value w, but will retain a steady-state error, which is a function of the proportional gain. This can be qualitatively understood as follows. If in the example of our heating system we have continuous losses of heat (outside it is cooler than inside), therefore we need continuous 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 (5.27) the actuating variable y is zero for zero error signal $w - x$. This means that the pure proportional controller cannot reach the setpoint w. The higher the load (i.e. the cooler it is outside) the greater is the deviation from the set-point 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. The deviation between the output x and the setpoint w is proportional to the heat dissipation (load) and inversely proportional to the proportional gain K_P.

The time delay due to the controller is related to the proportional gain K_P. The greater K_P is, the shorter is the time delay of the controller, i.e. the controller can follow fast. However, a large value of K_P also leads to a larger overshoot.

An example of how a P controller can be implemented using an operational amplifier was shown in Fig. 5.7. The gain constant K_P can be modified by changing the resistances as $K_P = -R_2/R_1$.

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 set-point value.

5.7.1.1 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

$$y(t) = K_I \int_0^t (w - x(\tau))d\tau. \tag{5.28}$$

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 slow reaction to changes of the input signal, due to the integration. Of course also the I controller can be made faster (shorter time delay) by increasing K_I, however, this also increases the tendency towards overshooting and instable 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.

5.7.2 *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

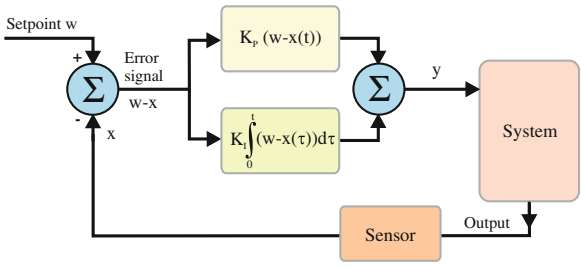

Fig. 5.10 Schematic of a PI controller

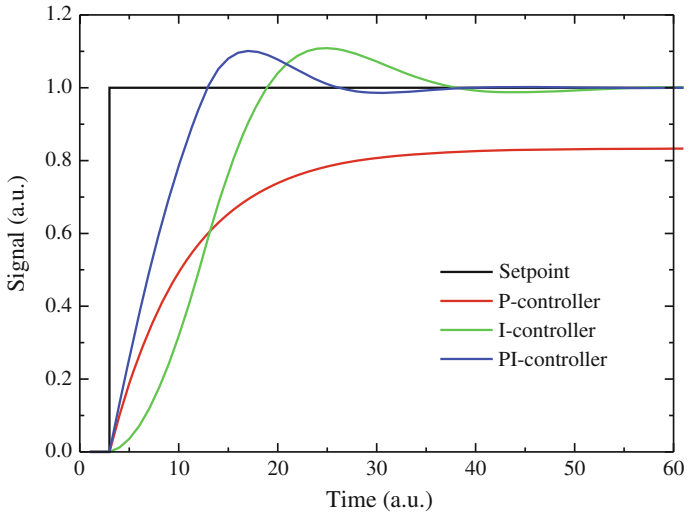

Fig. 5.11 Comparison of the step response of different controllers. The setpoint is a step function which changes from zero to one at time zero. Due to the steady-state error of the P controller the set point is never reached

$$y(t) = K_P(w - x(t)) + K_I \int_0^t (w - x(\tau))d\tau. \tag{5.29}$$

One way to show the performance of controllers is the step response. Step response means that the setpoint is changed instantaneously and the reaction of the controller (and the whole system) to reach the new setpoint is monitored. The step response of different controllers is compared in Fig. 5.11. The P controller does not reach the new set-point value, and the I controller alone is quite slow. The PI controller reaches the setpoint in a reasonable time for an appropriate choice of K_P, K_I.

5.8 Feedback Controller in STM

In STM or SPM in general the elements in the above-mentioned feedback loop have the following correspondence (Fig. 5.12).

- The system output x corresponds to the tunneling current, which is converted to a corresponding voltage by the current amplifier (sensor).
- The setpoint w corresponds to a voltage representing the desired tunneling current.
- The PI controller determines the system input (control variable) y, which is the voltage to be applied at the z-piezo element in order to change the tip-sample distance.

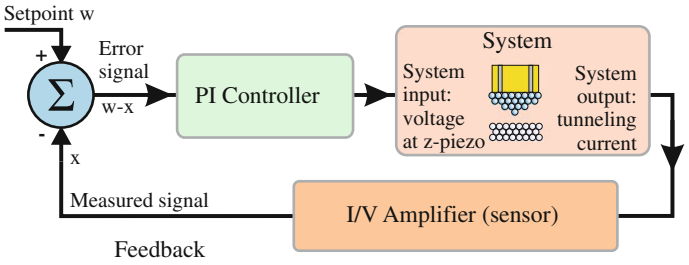

Fig. 5.12 Model of an STM feedback loop

- The most complex part of the feedback loop is the system itself. In the case of STM, it consists of DA converters, the high-voltage amplifiers (HVA) for the z-piezo voltage, the z-piezo element for the vertical positioning of the tip, and the tunneling contact.
- The noise of the z-signal arises due to external mechanical vibrations, the noise of the amplifiers, and the noise of DA and AD converters.

In STM, the P-part of the controller regulates fast deviations from the setpoint such as the atomic corrugation or atomic step edges (here the integrator helps to reach the final value, i.e. eliminates steady-state deviations).

In SPM, there is one effect which excretes the highest load to the feedback controller. Usually the sample is not oriented perfectly parallel to the xy-directions given by the scanner. This slope is usually the largest height signal in the original STM data and will be removed by appropriate background subtraction in the final image, as we will see later. However, the feedback has to follow this 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 by far the largest height signal compared to, for instance, a few atomic steps (3 Å high) in such an image.

The I controller has the advantage that it is less prone to noise. Depending on the conditions, the measured signal (tunneling current in STM) 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.

Now we consider the problem that a feedback loop may become instable and start to oscillate. If the controller parameters (the gains of the proportional and integral terms) are chosen incorrectly, the feedback loop can become instable, i.e. its output starts to oscillate. An important reason for the instability of the feedback loop is the time delay (reaction) of the system. In our simple example of the heating system of a house, it takes some time after a deviation of temperature is detected before the radiators and the air in the house become hot. In the case of the STM, the time delay of the system is given by the time lag between a change of the z-voltage by the controller and a corresponding change of the tunneling current. Also the speed of the controller itself (given by the gains of the proportional and integral terms) is a source

of time delay. It is intuitively clear that a large gain (heating power) and a long delay time of the system will give rise to large overshoots and result in an instability with oscillations of the controller system. A large part of controller theory is concerned with finding conditions for stability of a feedback loop. Here we will only provide a very qualitative intuitive discussion of the stability of a feedback loop.

A different way of characterizing the stability of a feedback loop than the analysis of the step response is to measure the output signal relative to a sinusoidal input signal (transfer function). The transfer function is the output signal divided by input signal. The knowledge of the (sinusoidal) output behavior as function of the sinusoidal input for all frequencies gives complete knowledge of the system response, since any input signal can be represented as a sum of the sinusoidal functions (Fourier theorem). The transfer function is a frequency dependent function and consists of an amplitude and a phase (complex number).

The transfer function of the whole feedback loop can be measured as shown schematically in Fig. 5.13. Initially the feedback loop is enabled and the STM is in tunneling operation. Then the (digital) feedback is switched off and the z-piezo voltage is modulated. The sinusoidal input signal is fed through all analogue components of the STM, HV amplifier, piezo actuator, tunneling junction, and current amplifier, as well as the controller. Then the output signal is measured (amplitude and phase), which results in the frequency dependent transfer function.

The measured transfer function (amplitude part) of the analogue components of a particular STM feedback loop is plotted in Fig. 5.14, as system output divided by system input amplitude as a function of frequency. The characteristics of this transfer function are the characteristics of a low-pass, and the amplitude drops significantly above 4 kHz. This corresponds to the bandwidth of the current amplifier, which is the bandwidth-limiting element of the analogue components in the system. The other elements of the system, HV amplifier (HVA), piezo actuator, and tunneling junction, do not limit the bandwidth of the system.

One very simplified condition for an instable feedback loop is the following: If for a certain frequency the output amplitude is larger than the input amplitude

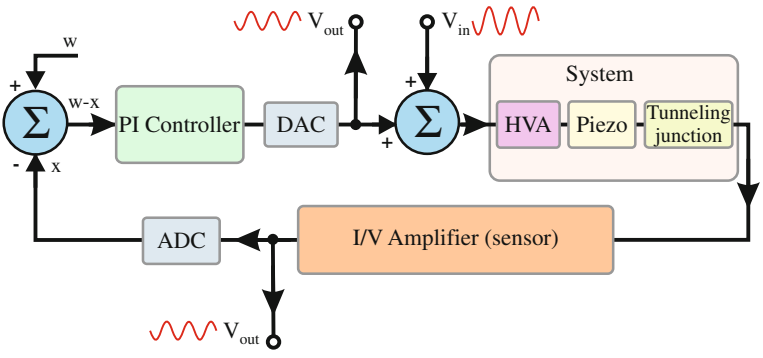

Fig. 5.13 Scheme of the measurement of the transfer function

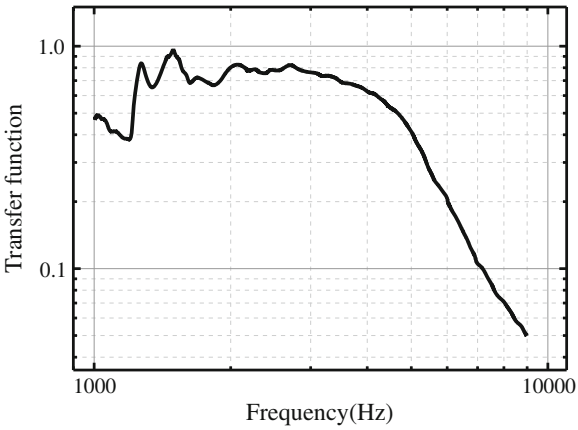

Fig. 5.14 Measured transfer function of the analogue components in the STM feedback loop: HV amplifier, piezo actuator, tunneling junction, and current amplifier

(amplitude of the transfer function ≥ 1) and the phase for this frequency is close to 0° the feedback loop will become instable. This means that small deviations from the setpoint will build up to an oscillation of large amplitude.

5.9 Implementation of an STM Feedback Controller

Feedback controllers are realized via a digital feedback loop nowadays. The tunneling current is measured by the current amplifier and then the corresponding voltage is digitized by analog digital converters (ADC), as shown in Fig. 5.15. These converters

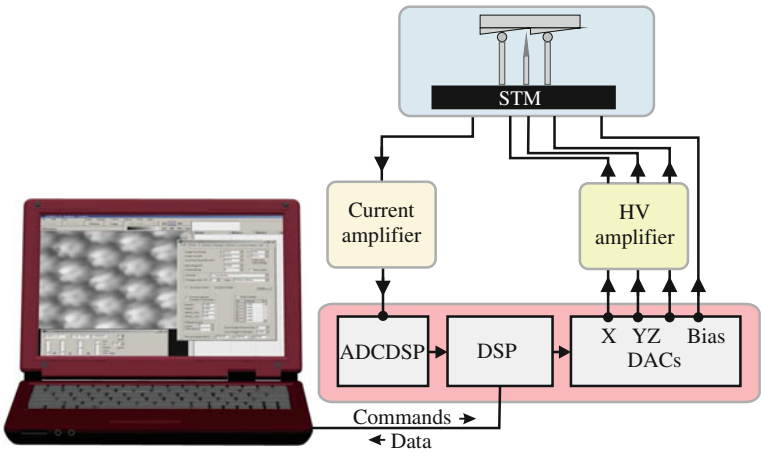

Fig. 5.15 Implementation of computer controlled STM electronics

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) (Fig. 5.15). A DSP is a own computer on which a single user single-task real-time program runs. From the measured (digitized) current and the current 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 (i.e. to run a tunneling voltage ramp or a z-ramp), 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 in the DSP program. Furthermore, non-linear algorithms for noise reduction can be implemented.

An example for a pseudocode implementation of a PI controller is given in the following.

```
start
read measured_signal  x(t)
error_signal = set_point - measured_signal  w − x(t)
integral = integral + error_signal * dt  ∫_0^t (w − x(τ))dτ
controller_output = KP * error_signal + KI * integral  y(t)
goto start
```

Once the controller output (new z-voltage) is calculated, this number is converted into an actual voltage by (for instance) 20 bit digital analogue converters (DAC). This z-voltage (range: ± 10 V) is then amplified by a high-voltage amplifier to a range of typically ± 200 V (Fig. 5.15). 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. 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. Also the tunneling bias voltage is supplied from the computer via a DAC in order to ramp this voltage in spectroscopic measurements.

When scanning an STM 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 STM tip above the surface and displays it as an image, i.e. in gray scale as a function of x and y.

The digital control of the STM also allows an automated procedure to be used during the coarse approach of the tip towards the sample. A flow chart for an automated control could be as shown in (Fig. 5.16). After the automatic coarse approach a desired current setpoint is chosen and scanning can be started.

The bias voltage between tip and sample (usually between a few millivolts and a few of volts) can be applied to the sample (sample bias). In this case, the tunneling

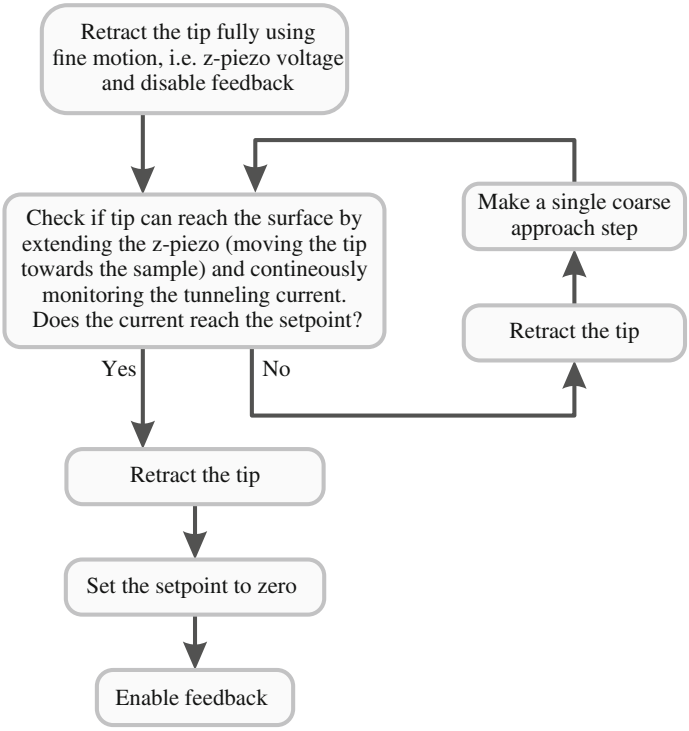

Fig. 5.16 Flow chart of the automatic approach procedure in scanning tunneling microscopy

current is measured relative to the ground. If the sample is grounded, the preamplifier has to float on a bias potential (tip bias) in order to apply a bias voltage between tip and sample.

5.10 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 scanning probe 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

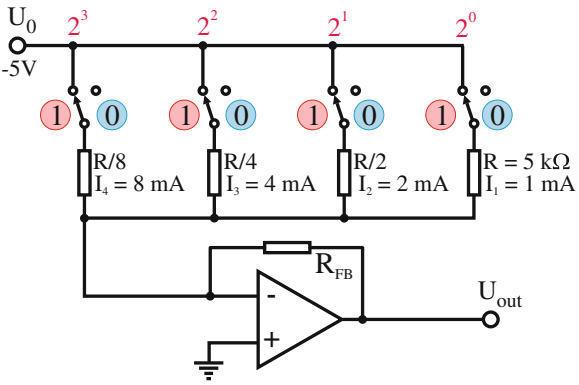

Fig. 5.17 Operating principle of a digital-to-analog converter

the following. In Fig. 5.17, 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 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 (5.22), the analog output voltage at the operational amplifier is

$$U_{\mathrm{out}} = -R_{\mathrm{FB}} U_0 \sum_{\mathrm{i=all\ closed\ switches}} \frac{1}{R_{\mathrm{i}}}. \tag{5.30}$$

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

5.11 Analog-to-Digital Converter

In scanning tunneling microscopy, the analog voltage at the output of the current preamplifier has to be converted to a number (e.g. 16-bit value) proportional to the analog voltage (tunneling current). For this task, an analog-to-digital converter (ADC) is used. An ADC can be realized by the comparison of the analog signal

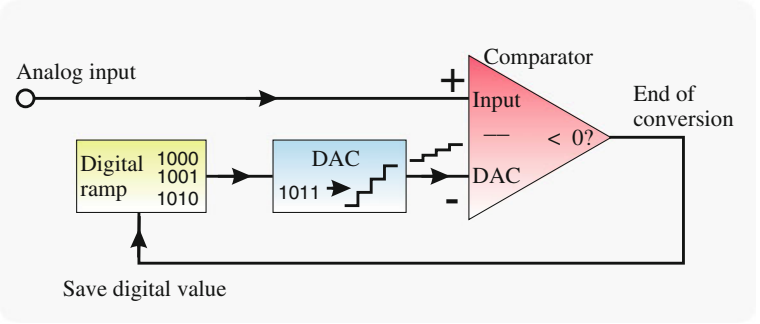

Fig. 5.18 Operating principle of an analog-to-digital converter

(to be digitized) to a voltage from a digitally generated voltage ramp. The principle of operation of one simple ADC is shown in Fig. 5.18. 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. 5.18). 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.

5.12 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 feedback loop (actually the current amplifier) 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. 4.2), 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.

5.13 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.

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.

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 (modulated) signal has to be known and, second, the phase of the signal has to be stable. If the signal to be measured is a DC signal the signal has to be modulated, i.e. multiplied by an AC reference signal to obtain a phase stable AC signal of a known frequency.

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

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

where A and B are the amplitudes of both harmonic functions and ω_1 and ω_2 are the corresponding angular frequencies, 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 term containing the frequency difference results in a DC component of the

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_6

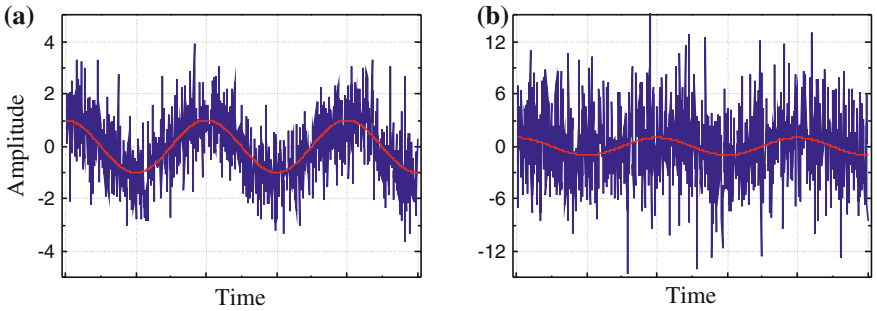

Fig. 6.1 Sinusoidal AC signal (*red*) and sinusoidal signal plus noise (*blue*). The noise increases from (**a**) to (**b**). The amplitude of the harmonic signal is always one

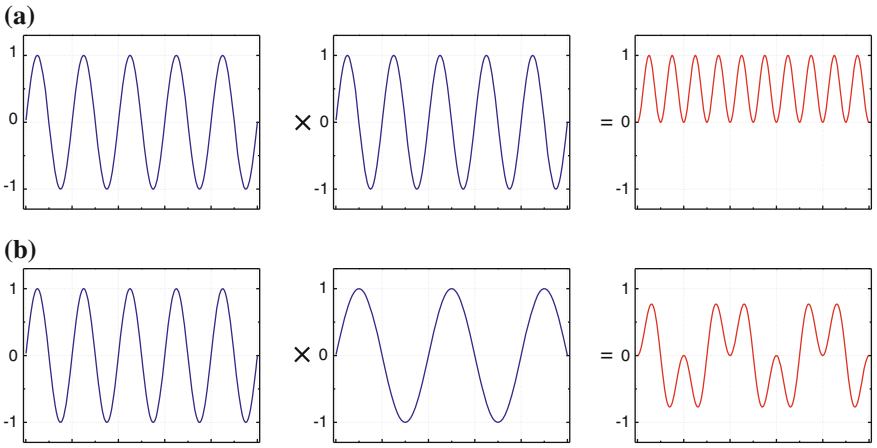

Fig. 6.2 **a** Product of two phase-coherent harmonic functions with identical frequency $\omega_1 = \omega_2$ results in a DC component plus a harmonic component. **b** Product of two phase-coherent harmonic functions with different frequencies $\omega_1 \neq \omega_2$ results in a harmonic signal without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

value $\frac{1}{2}AB\cos\varphi$. 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

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T A\cos(\omega_1 t+\varphi)\times B\cos(\omega_2 t)\,dt=\begin{cases}\frac{1}{2}AB\cos\varphi & \omega_1=\omega_2\\ 0 & \omega_1\neq\omega_2\end{cases} \tag{6.2}$$

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 $\frac{1}{2}AB\cos\varphi$, 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.

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 A (which is the signal amplitude to be measured modulated by the reference signal 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 $\sim(\omega_1 - \omega_2)$. Long integration times of about $2\pi/(\omega_1 - \omega_2)$ are required in order to average these low frequency components 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 $\varphi_{\text{noise}} \neq const.$ 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.

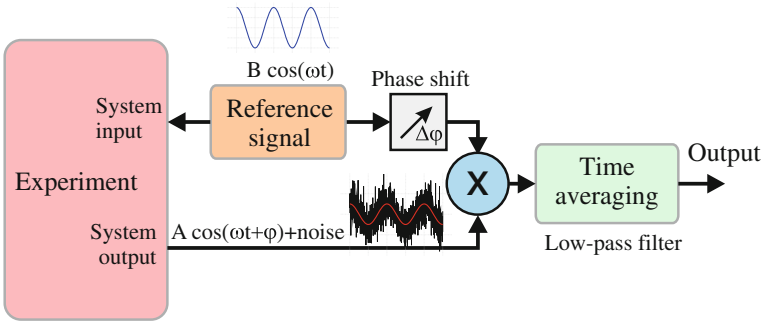

Fig. 6.3 Schematic of a lock-in amplifier consisting of a reference oscillator which modulates (via the experimental setup) the output signal of the system. This signal serves as input for the lock-in amplifier and is multiplied by the reference signal and then low-pass filtered. Due to this, only the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

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 $\cos\varphi$. 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.

The absolute value 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). If we neglect the constant factor $1/2\,B$ this results in $X = A\cos\varphi$ and $Y = A\cos(\varphi - \pi/2)$. Expanding this to complex variables $\tilde{X} = Ae^{i\varphi}$ and $\tilde{Y} = e^{i\varphi-\pi/2}$ 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 $\varphi = \arctan(Y/X)$. In digital lock-in amplifiers, the measured values X and Y are available as numbers and the computation can be performed arithmetically.

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.

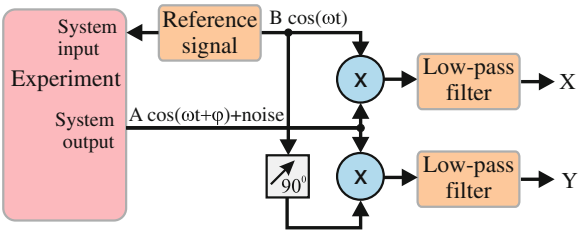

Fig. 6.4 Schematic of a two channel lock-in amplifier. Measuring X and Y and subsequently applying some arithmetic calculations leads to the simultaneous determination of the absolute value of the amplitude and the phase

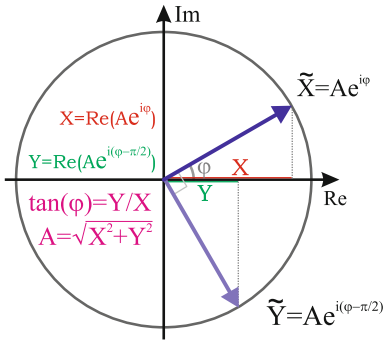

Fig. 6.5 Simultaneous determination of the amplitude A and the phase shift φ of the signal by a measurement with an additional phase shift of 90°, using a two-channel lock-in amplifier

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 tunneling current, or dI/dV is measured as a function of the lateral xy-position on the surface. Data representation is the task to map the measured heights (DAC values) 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.

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 acquired 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 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

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_7

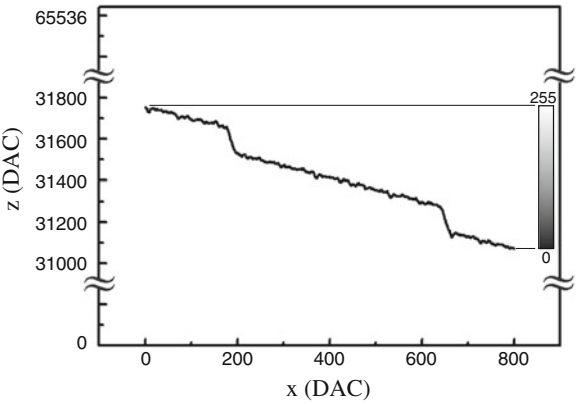

Fig. 7.1 For a good data representation the 256 *gray levels* have to be mapped to the 65,000 DAC levels in a proper way

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 measured height range 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 $\Delta h = \Delta x \tan\alpha \approx 500\,\text{Å}$. 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 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 background slope of 26 Å. If the atomic corrugation on a single atomic terrace is 1 Å the signal to background ratio is 1/26 in this case.

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.

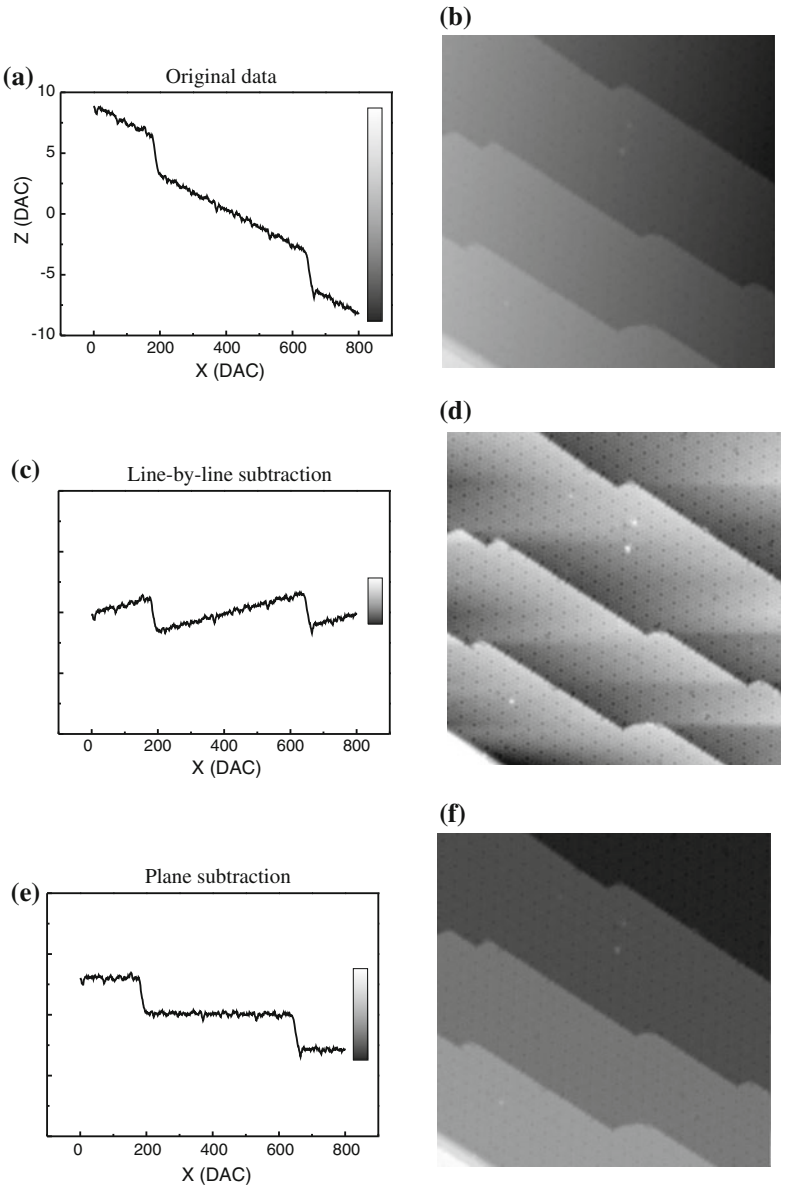

Fig. 7.2 STM data taken on a stepped Si(111) surface with the atomically resolved (7 × 7) reconstruction contained in the data. Comparison of different kinds of background subtraction for a single scan line (*left panel*) and a whole image (*right panel*). **a** and **b** Show the original data without 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 corresponds to an angle of 0.7° between sample and scanner

This background subtraction increases the contrast in the image, but also leads to artifacts like the black shadows which can arise 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 in large scans. This 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. 7.2f. 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.

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 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.

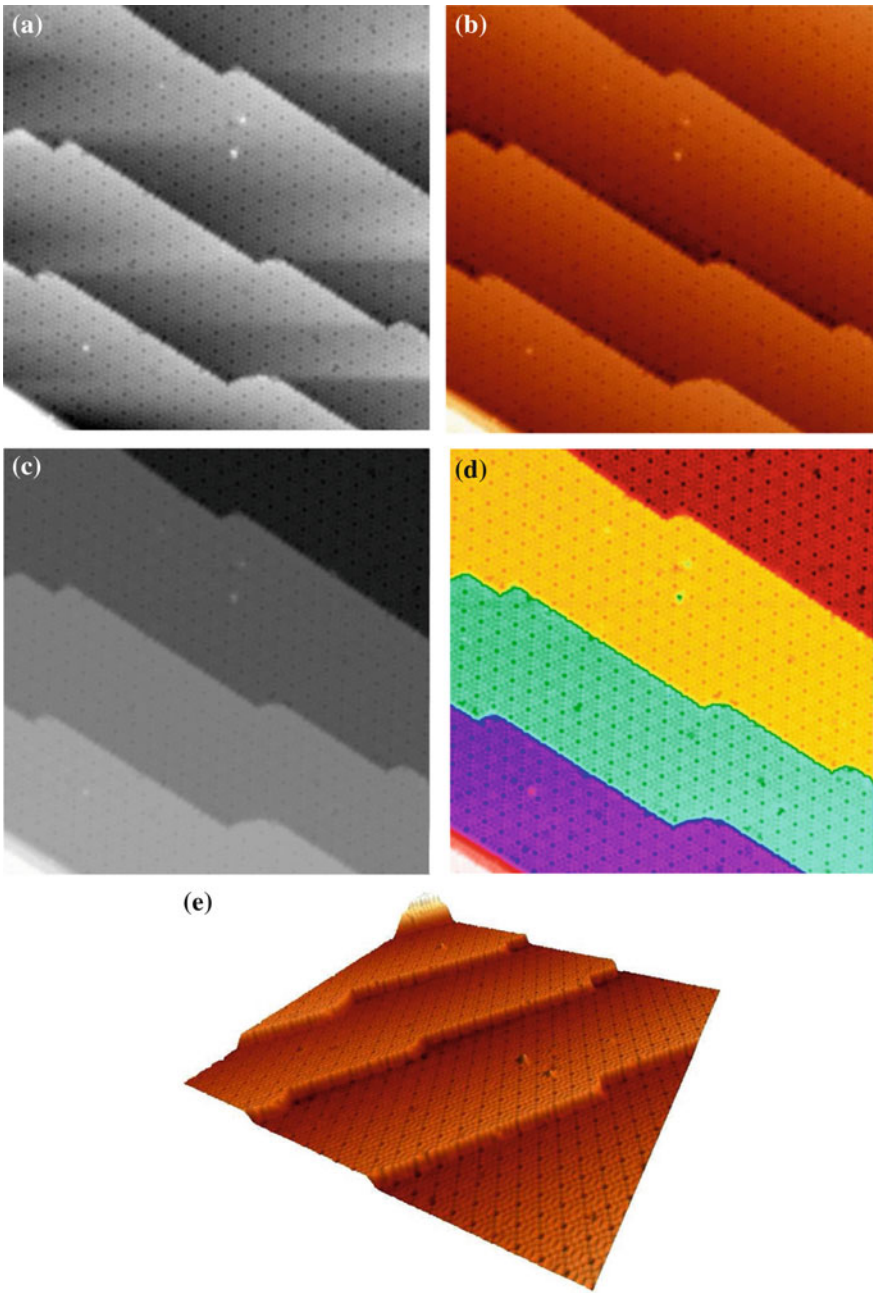

Fig. 7.3 STM image of a Si(111)-7 × 7 surface shown in different representations. Line-by-line background subtraction using **a** a *gray scale* palette and **b** a color palette. Plane background subtraction on one terrace **c** with *gray scale* palette and **d** a color palette with different colors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

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.

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 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

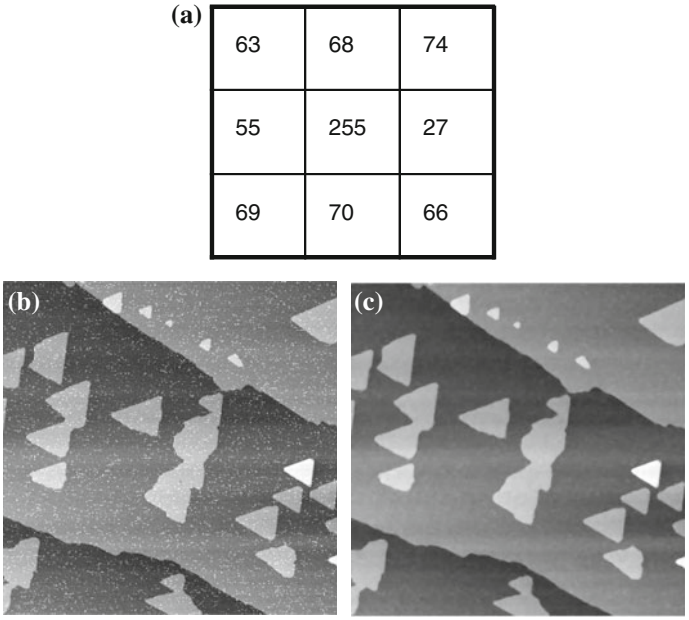

Fig. 7.4 **a** Example of the median filter showing gray values in a matrix of 8 pixels around a center pixel. When applying the median filter, the value of the center pixel is replaced by the fifth highest value (68 in the example). Thus the outlier value of 255 is replaced by the more reasonable value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

the 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 SPM 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.

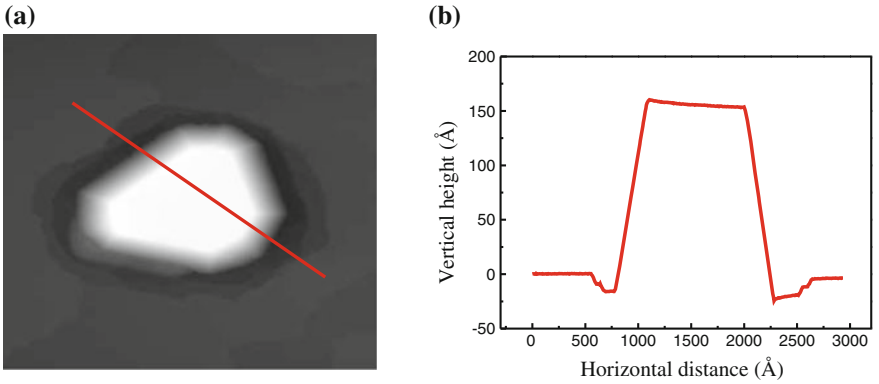

Fig. 7.5 **a** Gray scale STM image of a 3D Ge island. **b** Line scan across this island

A simple and general procedure for data analysis is the line scan. By interactive mouse clicking, 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 (atoms). Also the slopes of facets of surface features such as islands can be determined.

A second example of data analysis is the measurement of the roughness of a surface. The usual quantity characterizing the roughness of a surface is the RMS roughness defined as the standard deviation of the heights $h(x, y)$

$$\sigma = \sqrt{\langle (h(x, y) - \overline{h})^2 \rangle} = \sqrt{\frac{\sum_{x=1}^{L} \sum_{y=1}^{W} (h(x, y) - \overline{h})^2}{LW}}, \tag{7.2}$$

with L and W being the length and width of the image (number of pixels), and $\bar{h}$ the average height. A necessary requirement for a correct determination of the roughness is a good background subtraction of the scan slope.

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.

Chapter 8
Artifacts in SPM

The ideal tip is a sharp needle which can image surface features with high aspect ratios. If the tip has a broader shape artifacts occur due to a convolution of the tip shape with the surface features. Other kinds of artifacts in scanning probe microscopy include thermal drift, feedback overshoot, piezo creep, and electrical noise.

8.1 Tip-Related Artifacts

The most common artifacts in scanning probe microscopy occur due to the tip shape. Topographic features 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 are broadened and measured corrugation amplitudes can be reduced. In extreme cases, if sharp asperities are present on the surface the tip shape is imaged by the surface asperities. 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 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. Generally, if all (or many) features on the sample have the same shape, or if 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.

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_8

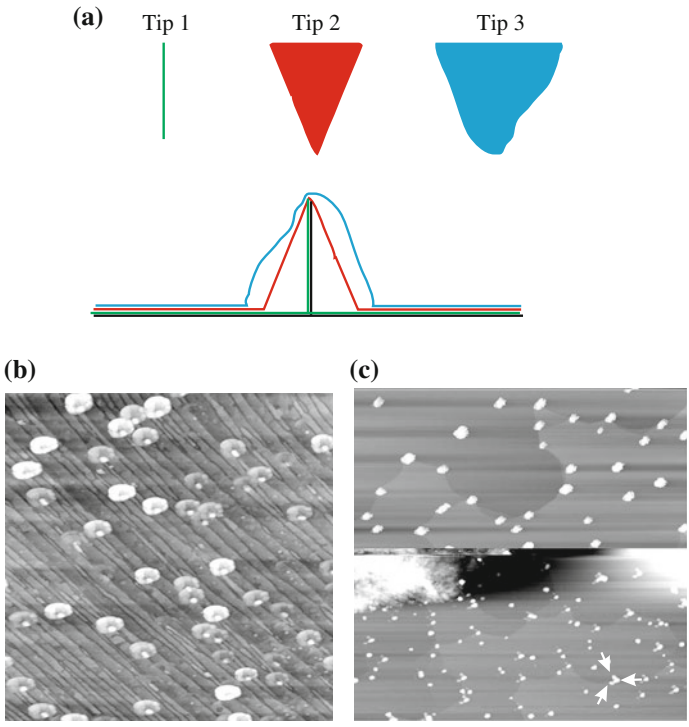

Fig. 8.1 **a** Sketch of the principle of how the tip shape influences the image of a sharp asperity present on the surface. **b** Example in which high aspect ratio carbide clusters are imaged by a blunt tip. All imaged clusters have a similar apparent shape: the tip shape. **c** Image of carbide clusters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

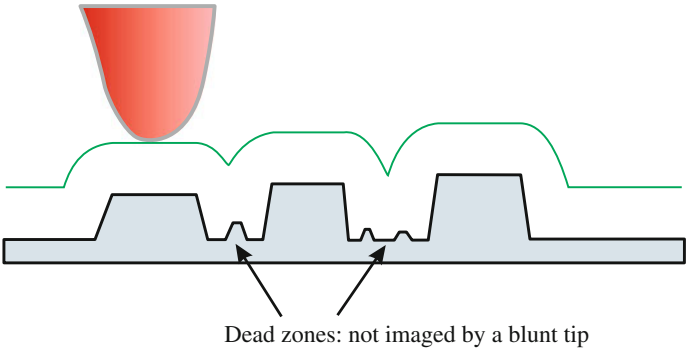

Fig. 8.2 Schematic showing the occurrence of "dead zones" due to the blunt shape of the tip

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 often not very useful for 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

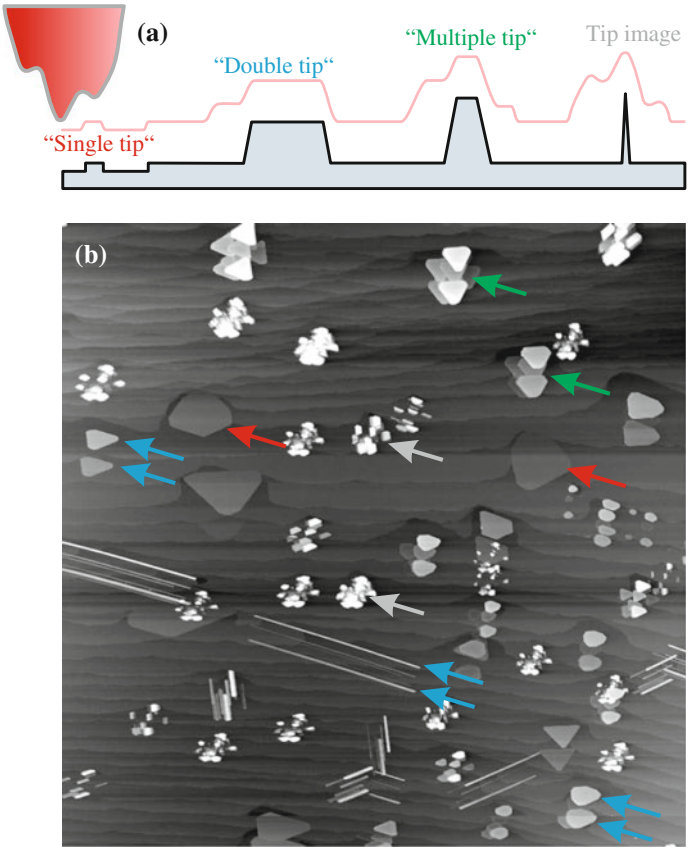

Fig. 8.3 **a** Sketch of a double (multiple) tip giving rise to doubled (multiple) imaging of surface features. The *light red line* shows the trace of the tip above the surface. **b** Example of silicide nano islands and nano wires imaged. The higher the structures imaged, the stronger is the tendency 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 if the tip structure instead of the surface feature (*gray arrows*)

needle-like structures on the surface is not practicable. (c) The tip shape changes quite often. Therefore, any tedious measurement of the tip shape does not last for long. Probably not until deconvolution is attempted.

One particular case of a blunt tip is a double 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 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 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 STM 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.

This does not mean that you should not believe any SPM images, but rather you should always critically reflect on your SPM measurements and to reproduce measurements with different tips in order to exclude tip artifacts as carefully as possible.

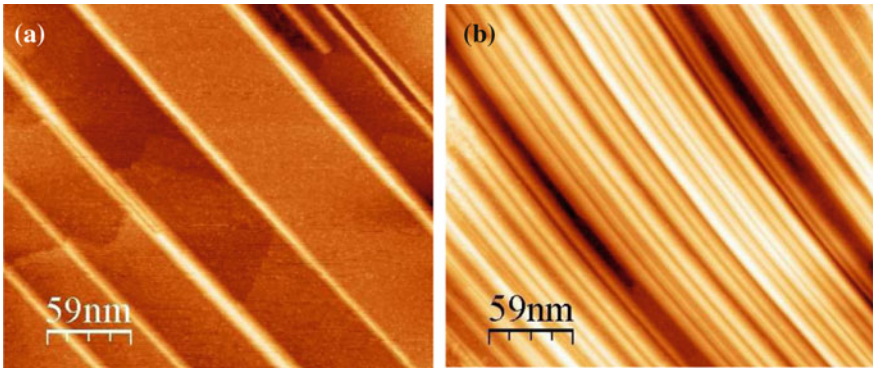

Fig. 8.4 STM image of 5 Å yttrium deposited on Si(110). **a** Silicide nanowires imaged with a sharp tip. **b** The same surface imaged with a blunt tip leads to much higher apparent coverage due to multiple images of the silicide nanowires

8.2 Other Artifacts

An artifact often appearing at the beginning of an image is a bending of all image structures, as seen in Fig. 8.5. This results due to piezo creep. Specifically if one moves to a new lateral position away from the previous one this effect is strong.

In discussing problems of piezo actuators Sect. 3.6, 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 SPM images. An example of image distortion due to creep or a non-linearity in the piezo extension is shown in Fig. 8.6. A silicide nanowire, which is known to be straight due to its crystallographic structure, is imaged as bent.

If the feedback parameters are not optimized this can lead to image artifacts. 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 such as a monoatomic step height or an other structure with high aspect ratio. In Fig. 8.7 the real signal (e.g. topography) changes from zero to one at $x = 50$ and back to zero at $x = 250$. The reaction of the AFM feedback signal to this is shown 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.

Different kinds of artifacts are induced by 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. 8.8. 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.

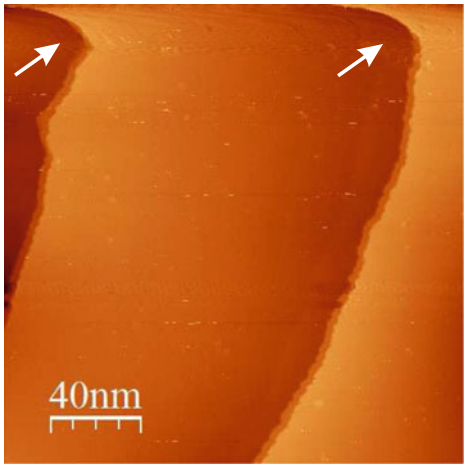

Fig. 8.5 Bending of atomic steps in the beginning of an image of a Si surface highlighted by *arrows*. Additionally to this artifact also an artifact due to a double tip is present in this image

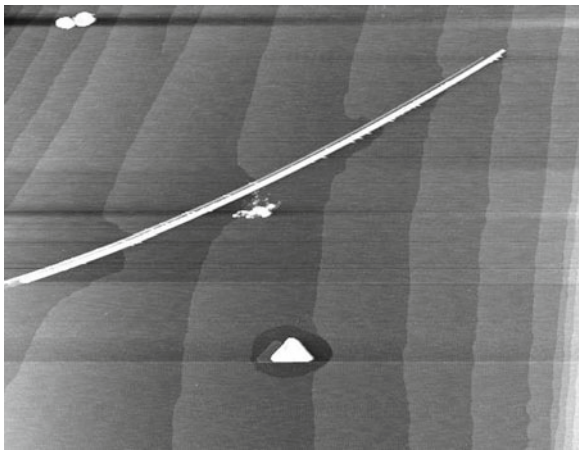

Fig. 8.6 Image of a straight silicide nano-wire, which appears bent in the STM image due to non-linearities in the piezoelectric actuators

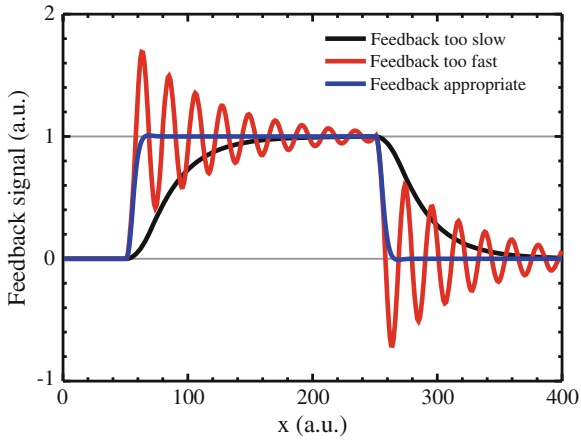

Fig. 8.7 Reaction of the AFM feedback signal to an abrupt change in the topography for too slow feedback settings (*black line*), too fast feedback settings (*red line*), and appropriate feedback settings (*blue line*)

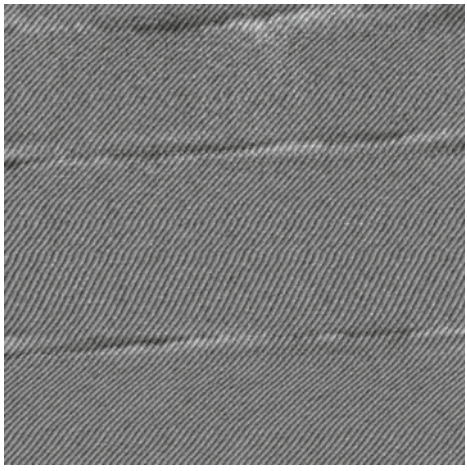

Fig. 8.8 Example of an image which is strongly influenced by 50 Hz noise. The three horizontal atomic step edges are hardly visible due to the strong 50 Hz noise

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 vibrational noise due to building vibrations. In the section on vibration isolation, we discussed how to combat this kind of noise.

8.3 Summary

- The shape of the tip influences the SPM 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 creep and non-linearity leads to distorted images.
- Power line noise and feedback overshoot are further sources of image artifacts.

Chapter 9
Work Function, Contact Potential, and Kelvin Probe Scanning Force Microscopy

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 scanning force microscopy 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, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_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 to the nuclei 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 [14]. 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

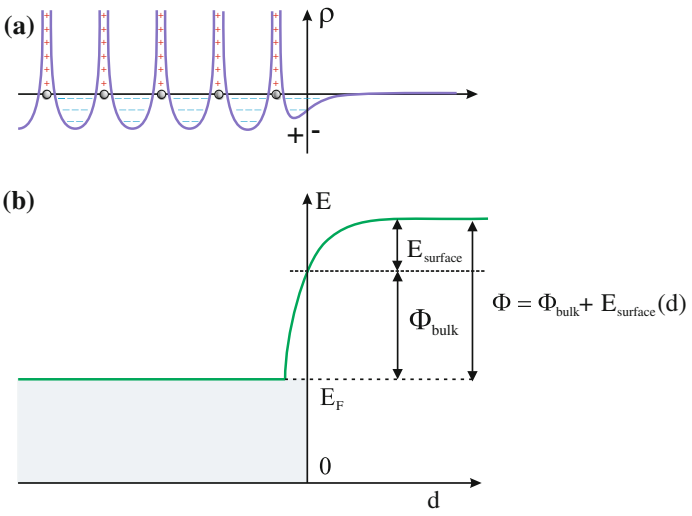

Fig. 9.1 **a** Charge density in a metal crystal which is modified close to the surface and spills out towards the vacuum. This behavior can be described qualitatively by a dipole layer of excess charge density close to the surface. **b** Energy of an electron as function of the distance d from the surface 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

$$E(L) = \frac{\hbar^2 \pi^2 n^2}{2 m_e L^2}. \tag{9.3}$$

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 different "spill out" of charge. This leads to different surface dipoles and therefore

[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. $-e^2/(4\pi\varepsilon_0 2d)$. 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

$$V_{\text{image}}(d) = \int_{\infty}^{d} \frac{e^2}{4\pi\varepsilon_0 2r} dr = \frac{-e^2}{4\pi\varepsilon_0} \frac{1}{4d}. \tag{9.5}$$

.

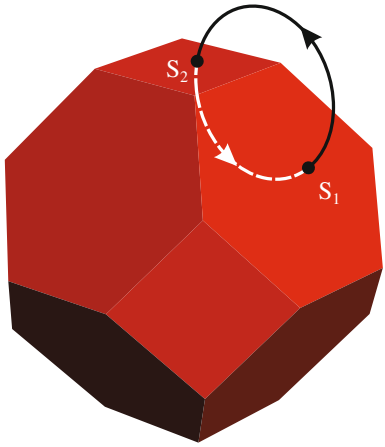

Fig. 9.2 Due to energy conservation, zero total work has to be done in moving an electron along the closed path from inside the metal crystal through surface S_1 and back through surface S_2. This argument shows that the two surfaces S_1 and S_2, which are assumed to have different work functions, have to be at different electrostatic potentials. This different potentials are built up by corresponding surface charges

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 charge, 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

$$e(\phi_1 - \phi_2) = E_1 - E_2 = \Phi_1 - \Phi_2. \tag{9.6}$$

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 to an external electric fields with a range corresponding to the size of the crystal. At larger distances from the crystal these fields vanish.

[4] All net charges are located at the surface of a metal, since the electric field vanishes in the interior of a metal.

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 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 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.

[5] It is always assumed that the electron is at rest, i.e. there is no kinetic energy contribution to the work.

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

$$eV_{\mathrm{contact}} = \Delta\Phi \tag{9.8}$$

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 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

$$V = V_{\mathrm{contact}} - V_{\mathrm{comp}}. \tag{9.9}$$

[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 $E_{\mathrm{vac}} = E_{\mathrm{vac}}^{\infty}$.

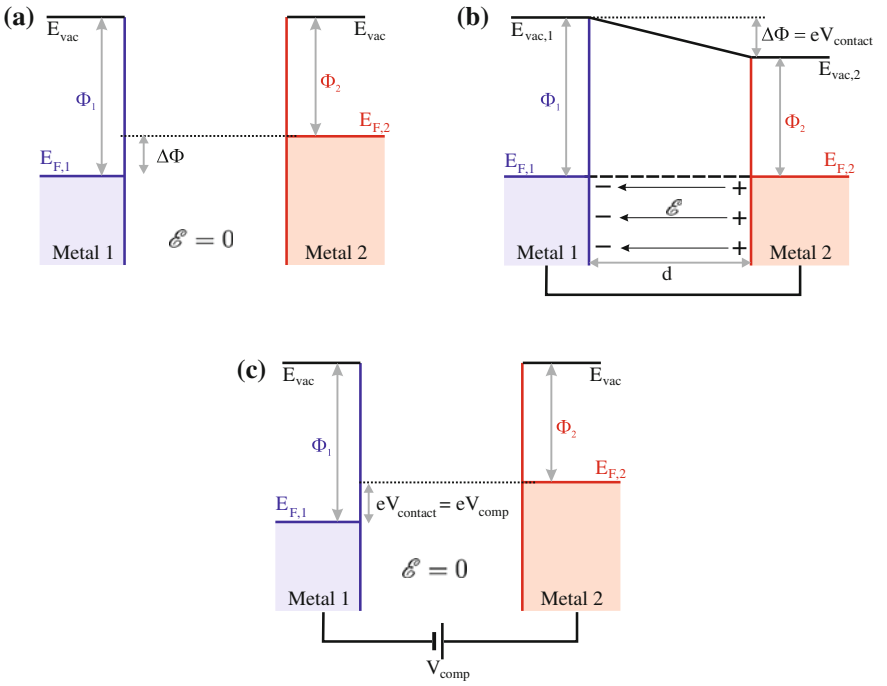

Fig. 9.3 **a** Potential energy diagram for two metals with work functions Φ_1 and Φ_2, which are initially not connected and share thus a common vacuum level. **b** If the two metals are connected by a conducting wire, the Fermi levels of the two metals align. A buildup of surface charge leads to a 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

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

$$I = \frac{dQ}{dt} = \frac{dC}{dt}\left(V_{\text{contact}} - V_{\text{comp}}\right), \tag{9.11}$$

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

$$V_{\text{comp}} = V_{\text{contact}} = \frac{1}{e}\Delta\Phi. \tag{9.12}$$

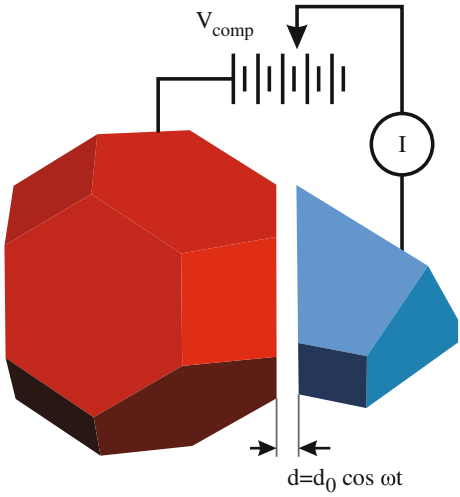

Fig. 9.4 The surfaces of two metals are brought together in a plate capacitor configuration. When the distance d between the plates is modulated a charge flow (capacitive current) can be measured. When an external bias potential just compensates the work function no current flows anymore

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 (KFM)

While Kelvin probe scanning force microscopy 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

$$F_{\mathrm{el}}(z, V) = -\frac{\partial E}{\partial z} = -\frac{1}{2}\frac{\partial C}{\partial z}V^2(t). \tag{9.13}$$

[7] This is done since the force (not the current) is measured in a scanning force microscopy setup.

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 $V_{\text{contact}} - V_{\text{comp}}$, and additionally a voltage component which is modulated at the modulation frequency ω_{mod} resulting in a total voltage between tip and sample as

$$V(t) = V_{\text{contact}} - V_{\text{comp}} + V_{\text{mod}} \cos(\omega_{\text{mod}} t) \tag{9.14}$$

Thus the tip-sample force which is proportional to the square of the tip-sample voltage $V(t)$ results as

$$\begin{aligned} F_{\text{el}}(V) &= -\frac{1}{2}\frac{\partial C}{\partial z}\left[V_{\text{contact}} - V_{\text{comp}} + V_{\text{mod}} \cos(\omega_{\text{mod}} t)\right]^2 \\ &= -\frac{1}{2}\frac{\partial C}{\partial z}\Big[\left(V_{\text{contact}} - V_{\text{comp}}\right)^2 + 2\left(V_{\text{contact}} - V_{\text{comp}}\right) V_{\text{mod}} \cos(\omega_{\text{mod}} t) \\ &\qquad + V_{\text{mod}}^2 \cos^2(\omega_{\text{mod}} t)\Big]. \end{aligned} \tag{9.15}$$

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 $V_{\text{contact}} - V_{\text{comp}} = 0$. 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 $\propto \frac{1}{e}\Delta\Phi - V_{\text{comp}}$, 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 in Chap. 14). 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.

Chapter 10
Surface States

When the electronic structure of (crystalline) materials is described, usually the bulk is considered. Since the STM probes the electronic states at the surface we will now consider also the electronic states at the surface, the surface states. We use the single electron approximation and start with a very brief review of the bulk electronic structure. Then the surface states are discussed in one dimension within the quasi-free electron model. We will see that solutions of the Schrödinger equation with complex wave vectors lead to surface states. While these solutions are not allowed in (infinite) bulk crystals, they are allowed if a surface is present. Finally, we transfer the one-dimensional model qualitatively to three dimensions and discuss the two-dimensional surface states of a three-dimensional solid.

10.1 Surface States in a One-Dimensional Crystal

The well-known parabolic bands are found in the one-dimensional model of a periodic solid [14]. At the Brillouin zone boundary, different bands cross each other. If a weak lattice periodic potential $\hat{V}$ is present this leads to a splitting of the bands at the zone boundary. Due to the presence of the potential $\hat{V}$ a band gap free of electron states occurs. According to the Bloch theorem, the wave function in a one-dimensional lattice periodic potential $\Psi_k(z)$ can be written as a plane wave modulated with a lattice periodic modulation factor $u_k(z)$:

$$\Psi_k(z) = u_k(z)e^{ik\cdot z}, \tag{10.1}$$

with the lattice periodic function $u_k(z) = u_k(z + z_n)$ and the translational lattice vector z_n. Of course, also a corresponding solution exists for $-k$.

This applies to the bulk electronic structure, but what happens at the surface? To answer this question we consider a one-dimensional model of a quasi-free electron in a periodic potential ending at the surface as shown in Fig. 10.1. Inside the solid ($z < 0$) the general one-dimensional bulk solution applies, which can be written as

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_10

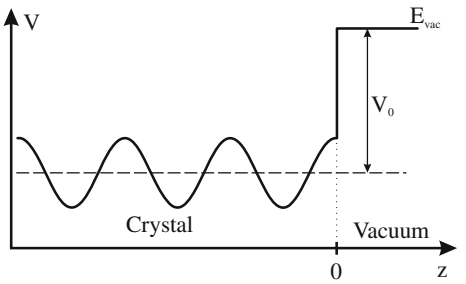

Fig. 10.1 Lattice periodic potential inside the one-dimensional crystal which ends at the surface ($z = 0$). In the vacuum region outside the crystal the potential has the constant value of E_{vac}

a linear combination of the solutions for k and $-k$ as

$$\Psi_{\text{bulk}}(z) = Au_k(z)e^{ik\cdot z} + Bu_{-k}(z)e^{-ik\cdot z}, \tag{10.2}$$

with real wave numbers k and $-k$ as well as energies in the allowed bands, i.e. outside the band gaps. However, any solution inside the crystal has to match the solution Ψ_{vac} for the region outside the crystal ($z > 0$). For the constant potential on the vacuum side, the solution is an exponentially decaying wave function (a wave function with a positive exponential cannot be normalized) as

$$\Psi_{\text{vac}} = D\ \exp\left[-\sqrt{\frac{2m}{\hbar^2}(V_0 - E)}\ z\right], \quad E < V_0. \tag{10.3}$$

The two solutions inside (10.2) and outside (10.3) the crystal and their derivatives have to be matched at the surface $z = 0$ as

$$\Psi_{\text{bulk}}(z = 0) = \Psi_{\text{vac}}(0) \ \text{ and } \ \Psi'_{\text{bulk}}(z = 0) = \Psi'_{\text{vac}}(0). \tag{10.4}$$

These two equations plus one equation from the normalization of the wave functions fix the three unknowns A, B, and D, and the matching condition can be fulfilled for any value of the coefficient k [15]. Thus, all energies which are allowed in the bulk crystal are also allowed for the surface problem. The resulting wave function is a bulk Bloch wave with an exponentially decaying tail into the vacuum (Fig. 10.2a). This solution is not really a surface state but it is a bulk electronic state up to the very surface where it is matched to an exponentially decaying tail.

Additionally to these bulk states decaying at the surface, there are solutions to the Schrödinger equation which are confined close to the surface and which are called surface states. We consider here the general concept of how surface states can arise from the presence of the surface. Usually the wave vector (wave number in the 1D case) k is considered to be real. However, Bloch's theorem does not require that the wave number k is real, it also allows Bloch functions with complex wave numbers. The (now) complex wave number k consists of a real part k' and an imaginary part κ as $k = k' + i\kappa$. Considering solutions with a complex wave number additional

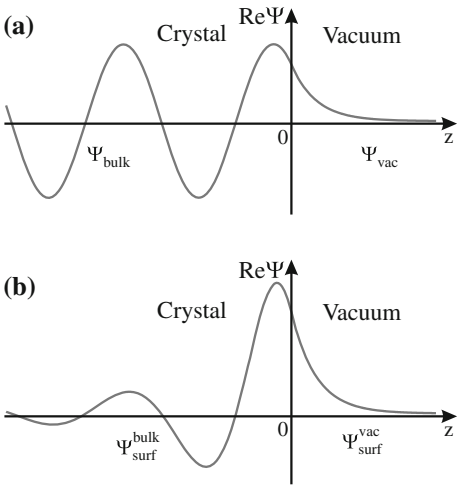

Fig. 10.2 Real part of the one dimensional wave function, for **a** a Bloch wave matched to an exponentially decaying tail in the vacuum and **b** a surface state wave function Ψ_{surf}. This surface state wave function consists of a part $\Psi_{\text{surf}}^{\text{bulk}}$ which is oscillatory inside of the crystal ($z < 0$) and exponentially increasing towards the surface. At the surface this wave function is matched to an exponentially decaying tail in the vacuum outside the crystal $\Psi_{\text{surf}}^{\text{vac}}$ for $z > 0$

(1D) solutions to the Schrödinger equation can be found inside the bulk which can be written as

$$\Psi(z) = u_k(z)e^{i(k'+i\kappa)\cdot z} = \left[u_k(z)e^{ik'\cdot z}\right]e^{-\kappa\cdot z}. \tag{10.5}$$

These wave functions grow without bound in one direction and decay exponentially in the opposite direction (depending on the sign of κ). Since the wave function has to be finite everywhere, such solutions have no relevance in the infinite crystal. However, this is no longer true at surfaces. Here the presence of the surface stops the exponential rise of the wave functions, at $z = 0$ the wave functions have to be matched to the exponentially decaying tail in the vacuum, $z > 0$. Wave functions may be obtained which are strongly localized at surfaces, have real energy eigenvalues, and can be normalized, as shown in Fig. 10.2b.

Thus we have grasped that the surface states arise due to wave functions with an imaginary part of the wave number. These wave functions can be normalized inside the crystal, since they grow to infinity only outside the finite crystal. Before we discuss the surface state wave functions further, we consider the range of wave numbers and energies for which surface states exist (a detailed treatment of this issue can be found in [15, 16]). If a complex wave number k is inserted into the expression for the energy as a function of k (dispersion relation) in the nearly-free electron approximation [15], the requirement that the energy has to be real leads to restrictions for the (complex) wave vector. For real wave numbers ($\kappa = 0$), of course, the usual solutions exist with k-values from $k = 0$ to $k = \pm\pi/a$, and subsequent higher bands (solid lines in Fig. 10.3) [16]. Furthermore, for complex wave numbers real energies are obtained if, and only if $k' = \pm\pi/a$ (as shown in [15]). This means that complex wave numbers occur only at the zone boundaries of the real part of the wave number ($k = \pm\pi/a + i\kappa$). In Fig. 10.3 also these solutions with an imaginary part of the wave number κ are shown as dashed lines at the zone boundary of k'.

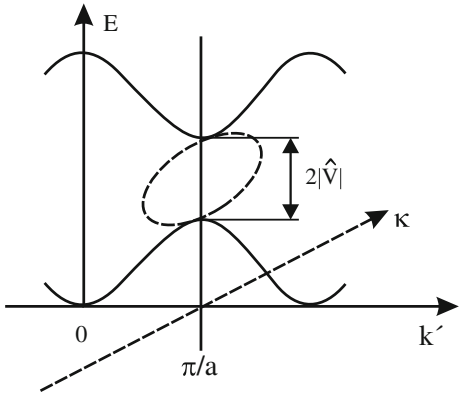

Fig. 10.3 Sketch of the electronic band structure for a 1D semi-infinite chain of atoms in the nearly-free electron model with an interaction potential $\widehat{V}$. Bulk states which decay exponentially at the surface into the vacuum give rise to energy bands shown as *solid lines*. Genuine surface states such as the ones shown in Fig. 10.2b are found in the band gap of the bulk states. They have complex wave vectors with a real part at the zone boundary $k' = \pm\pi/a$ and an imaginary part of the wave vector κ. These surface states with real energies are shown as *dashed lines*. The wave function matching condition restricts this continuous range of surface state wave functions to one particular surface state per bulk band

These solutions are the surface states and have real energies in the forbidden bulk band gap.

Due to the restriction of the real part of the wave number k to the zone boundary ($k' = \pm\pi/a$), the surface state wave functions can be written as

$$\Psi_{\text{surf}}^{\text{bulk}}(z) = Au_k(z)e^{i(\frac{\pi}{a}+i\kappa)\cdot z} + Bu_{-k}(z)e^{-i(\frac{\pi}{a}+i\kappa)\cdot z}. \tag{10.6}$$

For $\kappa > 0$ the first term will grow without bond inside the crystal $z < 0$, because it is proportional to $e^{-\kappa z}$. Since this would violate the finiteness of the wave function, A has to be zero. Correspondingly for $\kappa < 0$ B has to be zero.

Up to now we have only considered the part of the wave function inside the crystal. In the next step, the solution inside the crystal, (10.6) i.e. for $z < 0$ will be matched to an exponential tail in the vacuum. The two equations from the wave function matching conditions (wave functions and derivative at $z = 0$) plus the condition for the normalization of the wave function fix the three parameters B, D, and k. The wave function matching condition picks one k value out of the continuous range of values within the forbidden bulk energy gap (Fig. 10.3). Only one particular energy within the bulk band gap is compatible with the wave function matching conditions. The present consideration for a semi-infinite chain therefore yields one single electronic surface state per bulk band, which is located somewhere in the gap of the bulk states. Electrons in these states are, localized within a few Å of the surface plane (Fig. 10.2b).

10.2 Surface States in 3D Crystals

Now we consider the generalization of the results for the one-dimensional semi-infinite chain to a 2D surface of a 3D crystal. We now call the real part of the wave vector (perpendicular to the surface) named k' in the 1D model, $k_\perp$. As we have already discussed, the value of $k_\perp$ for surface states is at the Brillouin zone boundary. In the 3D case we have an additional wave vector parallel to the surface $k_\parallel$. Because of the 2D translational symmetry parallel to the surface, the general form of a surface state wave function is of the Bloch type with coordinates parallel to the surface. The energy eigenvalues of the surface states become functions of the wave vector $k_\parallel$ parallel to the surface as shown by the dashed lines at $k_\perp = \pi/a$ in Fig. 10.4. We thus arrive at a 2D band structure for the energies E_{surf} of the electronic surface states. A surface state is described by its energy level E_{surf} and its wave vector $k_\parallel$ parallel to the surface, $k_\perp$ is in any case at the Brillouin zone boundary. The plane in the reciprocal space given by the two components of the wave vector $k_\parallel$ parallel to the surface forms the surface Brillouin zone. In the dispersion relation in Fig. 10.4, only one of the two directions parallel to the surface is shown.

For the bulk states both $k_\parallel$ and $k_\perp$ components are allowed ranging from zero to the zone boundary, i.e. $k_\perp$ is not restricted to the value at the zone boundary ($k_\perp = \pi/a$). Therefore, a 3D band structure results as shown in Fig. 10.4. If all the bulk states are projected along $k_\perp$ onto the plane $k_\perp = \pi/a$ the shaded area in Fig. 10.4 results. If the surface state bands lie in this band gap of the projected bulk band structure, they are true surface states. Surface resonances lie inside the region of the projected bulk

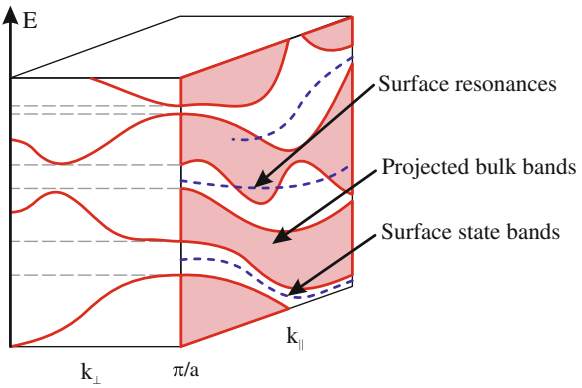

Fig. 10.4 Schematic of an electronic band structure of a three-dimensional crystal. The *shaded areas* arise by projecting the bulk band structure along $k_\perp$ onto the plane at $k_\perp = \pi/a$. The *dashed lines* in this plane indicate surface state bands. They can either lie in a gap of the projected bulk band structure (true surface states) or in areas inside the projected bulk band structure. In the latter case, these states are called surface resonances

band structure. Surface resonances have wave functions present throughout the bulk and, additionally, a large amplitude at the surface. Fig. 14.4 is an example in which no bulk band gap exists (energy range without states in the bulk), however, gaps in the projected bulk band structure exist.

10.3 Surface States Within the Tight Binding Model

So far we have discussed surface states within the picture of the quasi-free electron, which are sometimes, for historic reasons, called Shockley surface states. However, the surface states can also be described within the tight binding approximation; in this case they are called Tamm states. We will discuss this only qualitatively here. For the topmost surface atoms the bonding partners on one side are missing completely, which means that the wave functions have less overlap with the wave functions of the neighboring atoms. The shift of the atomic energy levels (into the electronic bands of the crystal) is thus smaller at the surface than in the bulk. The surface states split off from the bulk bands as shown in Fig. 10.5. Every atomic orbital leading to one of the bulk electronic bands should also give rise to one surface state level. The stronger the perturbation induced by the surface, the greater is the deviation of the surface level from the bulk electronic bands. When a particular orbital is responsible for chemical bonding, e.g. the sp^3 hybrid in Si or Ge, it is strongly affected by the presence of the surface. Bonds are broken and the remaining lobes of the orbital sick out from the surface. They are called dangling bonds. The energy levels of such states are expected to be significantly shifted from the bulk values.

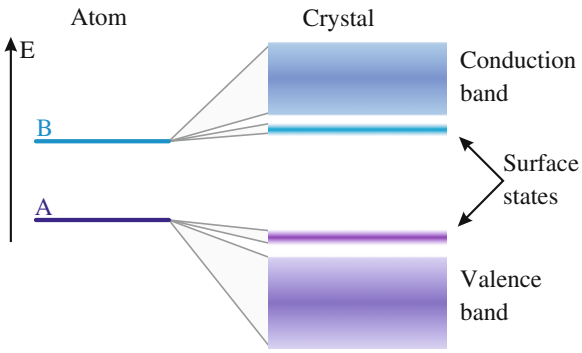

Fig. 10.5 Qualitative origin of surface states in the tight-binding picture. The surface atoms have less bonds to neighboring atoms. Therefore the energy shift from the atom levels is less than for the bulk atoms

10.4 Summary

- When a surface is present, the bulk states of the infinite crystal are still solutions of the semi-infinite solid if the bulk wave functions are matched to an exponentially decaying tail in the vacuum.
- In the semi-infinite solid (terminated by a surface), additional solutions exist which have a complex wave vector. This leads to an exponential increase of the wave function in the direction of the surface. Unlike the case of the infinite crystal, this is no problem since the crystal is finite and the exponential increase of the wave function stops at the surface.
- The energies of the (1D) surface states lie in the band gap of the bulk band structure and the real part of the surface state wave vector is at the zone boundary, i.e. π/a.
- True surface state bands are characterized by energy levels E_{surf}, which are not degenerate with bulk bands. They lie in the gaps of the projected bulk band structure.
- Surface resonances lie in parts of the surface Brillouin zone, where projected bulk states exist.

Part II
Atomic Force Microscopy (AFM)

Chapter 11
Forces Between Tip and Sample

One disadvantage of the STM technique is that it cannot image insulating samples since a tunneling current between tip and sample is needed. The idea behind the atomic force microscope (AFM) is to measure the force(s) 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.

11.1 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/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.

For the simple case of two noble gas atoms the dipole interaction of the first atom acting on the second can be treated analytically using some approximations [17], resulting in an interaction potential of

$$U_{\mathrm{vdW}}(r) = -\frac{C}{r^6}. \tag{11.1}$$

The distance dependence with the minus sixth power corresponds to a long-range interaction. The van der Waals interaction is (in a first approximation) non-directional (isotropic) and additive, which means that for two groups of atoms the total force between these groups is the sum of each pair between the two groups. Taking a

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_11

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 elementary volume element of the tip dV_A and an elementary volume element dV_B of the sample can be written as

$$dU_{\mathrm{vdW}} = -\frac{C\rho_A\rho_B}{|r_A - r_B|^6} dV_A dV_B, \tag{11.2}$$

with ρ_A and ρ_B being the atom densities of tip and sample, respectively. Approximating the tip by a sphere of radius R and the sample by a semi-infinite solid results in a van der Waals interaction energy [17] of

$$U_{\mathrm{vdW}} = -\frac{HR}{6D}, \tag{11.3}$$

where R 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. It is defined as $H = \pi^2 C\rho_A\rho_B$, with C being the coefficient in the atom-atom pair potential in (11.1). 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

$$F_{\mathrm{vdW}} = -\frac{HR}{6D^2}. \tag{11.4}$$

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 with 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 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. From now on we will use this model potential to describe tip-sample interactions. 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

$$U_{\mathrm{LJ}}(r) = 4U_0 \left[\left(\frac{R_a}{r} \right)^{12} - \left(\frac{R_a}{r} \right)^{6} \right], \tag{11.5}$$

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. 11.1a 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. 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. 11.1. 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. 11.1 and occurs where the force changes its sign, or correspondingly at the minimum of the potential.

If the tip and sample come into elastic contact, not only the corresponding wave functions intrude into each other (as considered using the Lennard-Jones potential), but the positions of the atoms change due to the elasticity of the tip and sample materials. This effect is described by the Hertzian theory of the elastic contact of two bodies. If the elastic modulus of the sample material is much smaller than the elastic modulus of the tip material, this results in a less steep repulsive distance dependence than the Lennard-Jones potential. This is the case if, for instance, soft materials like polymers or organic materials are imaged by a hard silicon tip [17]. Throughout this

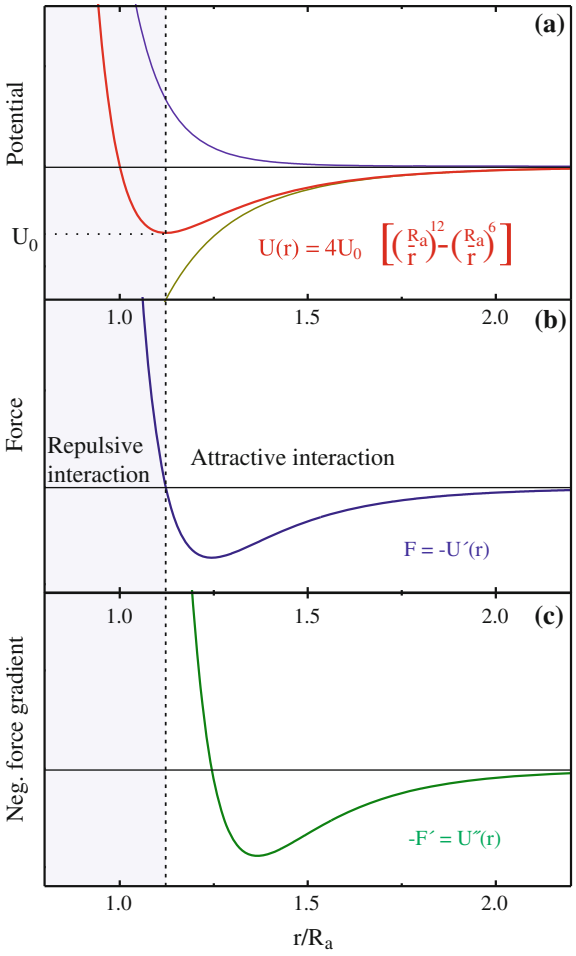

Fig. 11.1 **a** The Lennard-Jones potential will be used in the following as a model potential for a tip-sample interaction. The *green* and the *blue lines* show the attractive and the repulsive parts of the potential, respectively. The corresponding force is shown in (**b**) and the (negative) force gradient in (**c**). The border between attractive and repulsive forces (interactions) is indicated by the *vertical dashed line*

text we will however use the Lennard-Jones potential as a model potential for the tip-sample interaction.

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

$$F_{\mathrm{el}}(z, \Delta V) = -\frac{\partial E_{\mathrm{el}}(z)}{\partial z} = -\frac{1}{2}\frac{\partial C(z)}{\partial z}\Delta V^2. \tag{11.6}$$

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

$$C_{\text{plate}}(z) = \epsilon_0 \frac{A}{z}, \tag{11.7}$$

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 [18], and the sample by a semi-infinite solid, the electrostatic force between tip and sample results as

$$F_{\text{el}} \approx -\pi\epsilon_0 \frac{R}{z} \Delta V^2. \tag{11.8}$$

For a tip radius $R = 50\,\text{nm}$, 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 tip (or cantilever) than the final sphere of radius R might be important.

While the electrostatic force can have considerable values, it vanishes according to (11.6) 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 a (negative) parabola. Measuring the 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.

11.2 Snap-to-Contact

For a soft cantilever, atomic force microscopy is accompanied by so-called "snap-to-contact". To introduce this effect let us discuss an 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

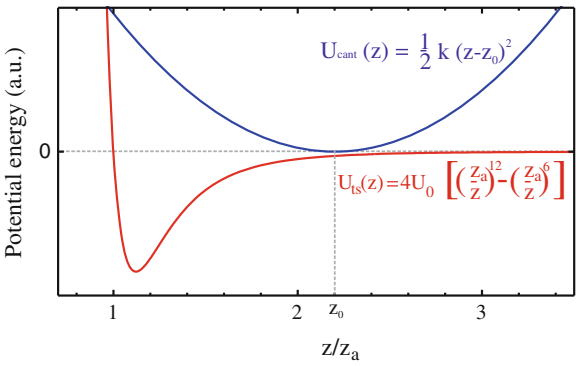

Fig. 11.2 Graphic representation of the two potentials acting on the cantilever: the tip-sample potential modeled by a Lennard-Jones potential and the parabolic potential arising due to the cantilever spring constant

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 now analyze the stability of a (cantilever) spring system if an outer (tip-sample potential) potential is added. The total potential energy of the cantilever system consists of two parts, as shown in Fig. 11.2: (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), and (b) the parabolic potential U_{cant} arising due to the spring constant of the cantilever.[1]

The total potential energy of the cantilever-tip-sample system can be written as

$$U_{\mathrm{tot}}(z) = U_{\mathrm{ts}}(z) + U_{\mathrm{cant}}(z) = 4U_0 \left[\left(\frac{z_a}{z}\right)^{12} - \left(\frac{z_a}{z}\right)^{6}\right] + \frac{1}{2}k\,(z - z_0)^2\,. \quad (11.9)$$

The variable z is the distance between the origin of the Lennard-Jones (tip-sample) potential and of the tip. The parameter z_0 is the distance from the sample 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. The actual tip-sample distance (including the contribution due to the bending of the cantilever) 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 adapting the tip-sample distance z, which corresponds to the lowest U_{tot}. However, 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.

[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.

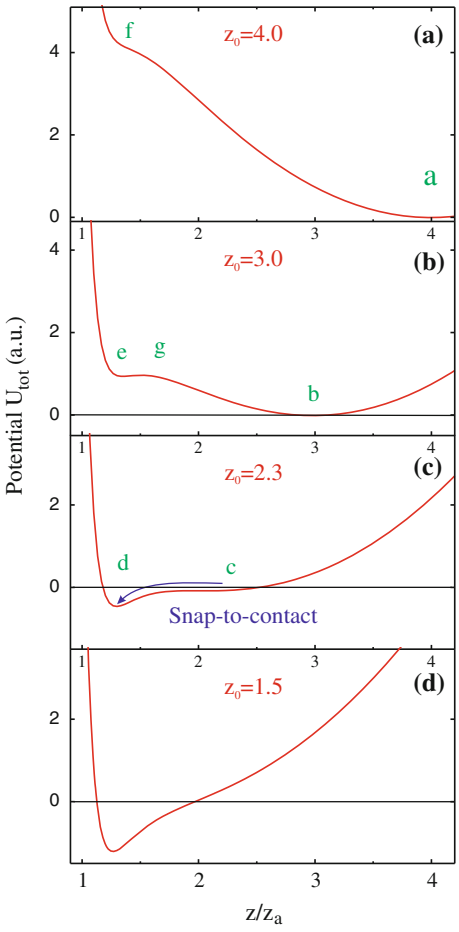

Fig. 11.3 Graphic representation of the total potential (tip-sample plus cantilever spring potential according to (11.9)) as a function of the tip-sample distance z. The potential is shown for different values of z_0, decreasing from (**a**) to (**d**), as the tip approaches the surface. The parameter 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**)

A graphic representation of the total potential of the cantilever (sum of the tip-sample potential and spring potential) is given in Fig. 11.3 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. 11.3a, 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, the potential minimum close to z_0 vanishes (converts to a saddle point) due to the increased interaction strength of the tip-sample potential for smaller tip-sample distances (Fig. 11.3c). Correspondingly, the cantilever tip will find a new stable minimum not close to z_0 but closer to the sample surface (Fig. 11.3d). 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 $U_{\text{tot}}(z)$ 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 $U_{\text{tot}}(z)$ is that the first derivative of the potential with respect to z has to be zero ($\frac{\partial U_{\text{tot}}}{\partial z} = 0$), which means that

$$\frac{\partial U_{\text{ts}}}{\partial z} + k(z - z_0) = 0. \tag{11.10}$$

Since $-\frac{\partial U}{\partial z} = F$, the above condition is actually a condition of force balance

$$F_{\text{ts}}(z) + F_{\text{cant}}(z, z_0) = 0. \tag{11.11}$$

This balance of forces is graphically represented in Fig. 11.4, with the force due to the cantilever bending F_{cant} represented by straight blue lines (Hooke's law: $F_{\text{cant}} = -k(z - z_0)$) 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 occurs if the red line corresponding to the Lennard-Jones force crosses one of the straight (blue) lines representing the cantilever spring force. It can be seen from Fig. 11.4 that for each position of z_0 one (or more) distances z can be found for which the force balance (11.11) 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 ($\frac{\partial^2 U_{\text{tot}}}{\partial z^2} > 0$, positive curvature). This second condition can be written as

$$\frac{\partial^2 U_{\text{ts}}}{\partial z^2} + k > 0. \tag{11.12}$$

Since $F_{\text{ts}} = -\frac{\partial U_{\text{ts}}}{\partial z}$, this condition can be expressed in terms of the force gradient as

$$k > \frac{\partial F_{\text{ts}}}{\partial z}. \tag{11.13}$$

[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.

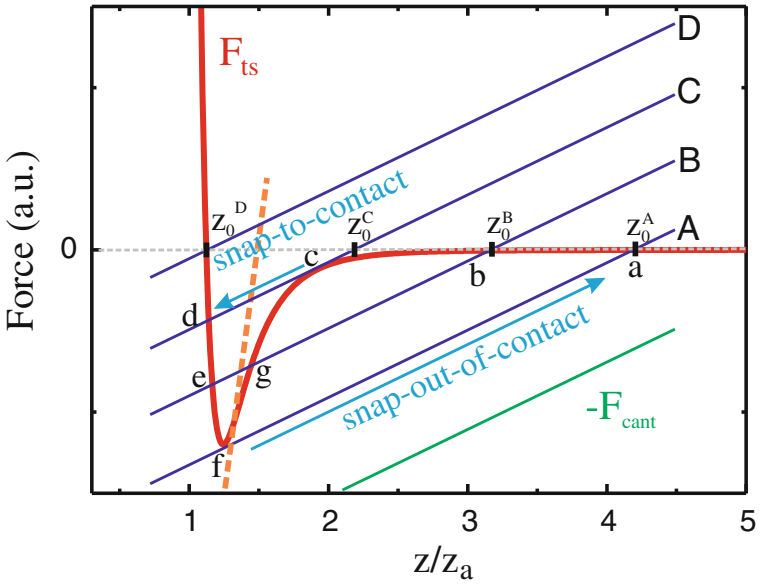

Fig. 11.4 Comparison of the tip-sample force (approximated by a Lennard-Jones potential) to the negative cantilever spring force (straight lines for different z_0, i.e. for different externally set tip-sample distances). If the two forces are the same (point(s) of intersection), a minimum, maximum or saddle point is present in the potential curve (compare Fig. 11.3). The cantilever spring constant k corresponds to the slope of the straight lines. If when the tip and sample approach each other, the gradient of the tip-sample force exceeds k, 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

If the tip and sample are still far from each other, the cantilever position z is at the minimum of the potential close to z_0 (Fig. 11.3) and the condition (11.13) is fulfilled, since the force gradient is very small for large z. If the tip and sample approach each other, this condition of stability holds until the force gradient becomes larger than the spring constant (which is the gradient of the cantilever force $k = -\frac{\partial F_{\text{cant}}}{\partial z}$). If (11.13) is no longer fulfilled the spring system becomes instable and snaps to contact (Fig. 11.3c). In the graphic representation in Fig. 11.4 this stability condition holds, if the slope of the tip-sample force F_{ts} (red curve in Fig. 11.4) is smaller than the slope (gradient) of the cantilever force (straight blue lines in Fig. 11.4).

After considering the equations governing snap-to-contact, we will now follow the snap-to-contact effect step by step using Figs. 11.3 and 11.4. Approaching the tip from large values of z_0 towards smaller ones, the tip-sample force can be neglected close to the point of equilibrium, which is very close to z_0 (point a in Figs. 11.3a and 11.4). When the cantilever approaches the surface (line B in Fig. 11.4), the cantilever spring force compensates the tip-sample force at the three intersection points b, g, and e in Fig. 11.4. The points b and e correspond to the two minima indicated in

Fig. 11.3b 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. 11.4, compare also Fig. 11.3c). 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. 11.4.

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 on line B and Fig. 11.3b). 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. 11.4, compare also Fig. 11.3a). 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, (11.13) is always fulfilled, i.e. for any value of z_0. This corresponds in Fig. 11.4 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. 11.4. 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 lower turnaround point of the oscillation. The red and the green lines in Fig. 11.4 will never cross,[3] as is the case for the blue lines. Due to the large amplitude for tip positions close to the surface the cantilever force is always stronger than the attractive tip-sample force and thus snap-to-contact is prevented. Thus the conditions of a large oscillation amplitude and a large z_0 prevent snap-to-contact and maintain the condition of stability, also for the case of small cantilever force constants k.

[3] Apart from a point very close to z_0.

11.3 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 electrostatic interaction between tip and sample can be suppressed by using an appropriate tip-sample bias voltage.
- 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.

Chapter 12
Technical Aspects of Atomic Force Microscopy (AFM)

The design of AFM instruments is in most aspects similar to that used in STM, as discussed in Chap. 4. We will mention here the aspects which are different from the STM. We start with 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. At the end of this chapter 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.

12.1 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

$$\omega_{\mathrm{vib}} = \sqrt{\frac{k}{m}} \tag{12.1}$$

can be applied. Thus the spring constant for the bonds of atoms in a solid results as

$$k = \omega_{\mathrm{vib}}^2 m \approx 10\,\mathrm{N/m} \tag{12.2}$$

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_12

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 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 10 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 ångström 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, cantilevers with a small force constant should be used. In summary, a first condition for a cantilever in 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. Let us assume that we scan a surface with a sinusoidal height profile in the static AFM mode, resulting in a cantilever oscillation with a frequency of 1 or 10 kHz for a fast scan. Thus, by following the topography, the cantilever is excited at a frequency of say 10 kHz. When discussing the harmonic oscillator (Chap. 2), we have seen that the harmonic oscillator follows an external excitation (with gain one and without a phase shift) only if the resonance frequency of the oscillator is much larger than the excitation frequency. Thus the cantilever should have a high resonance frequency, preferably $\gg$10 kHz.

While the 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. 14, the measured signal in the dynamic mode is proportional to ω_0/k. Thus for a large measured 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. 3.8 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 construction 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

$$\omega_{\text{cant}} = \sqrt{\frac{k}{m}}, \tag{12.3}$$

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 (12.3) 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).

12.2 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 prevents 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, which have some advantages over the thin-film cantilevers described above:

- The monolithic cantilevers from one material (Si) avoid strain due to thermal mismatch of the different materials bonded together in thin film cantilevers.
- All parts are made of single crystal material leading to a high internal Q-factor, which is important in dynamic AFM.

In the process described in the following, all parts of the cantilever are made of bulk silicon. The key ingredient is anisotropic wet etching, which means that different crystal directions are etched at different rates using anisotropic etchants like KOH. The (100) direction is etched much faster than the (111) direction. A simplified sketch of the fabrication process of a Si cantilever is shown in Fig. 12.1. The starting point is a Si(100)-oriented wafer on which a structured SiO_2 layer is formed as shown in Fig. 12.1a. 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. 12.1b. The formation of the tip is finished when the small oxide pad on top of the tip falls

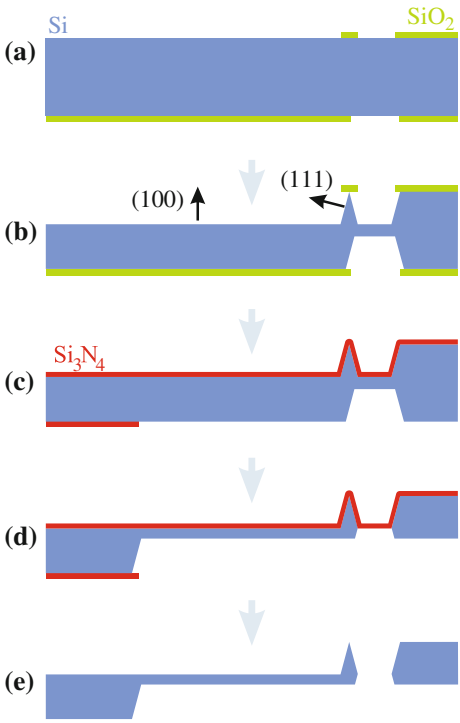

Fig. 12.1 Fabrication of a Si cantilever using alternating lithographic patterning and wet chemical etching as described in the text

off. 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. 12.1c). 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. 12.1d). Tip, cantilever and cantilever base are finished after removal of the protective silicon nitride film by a wet etching step (Fig. 12.1e). Electron microscopy images of a finished cantilever of this type are shown in Fig. 12.2. As the cantilever beam itself is too small to handle, it is connected to a solid silicon base with dimensions of several millimeters, seen partly in the left of the images in Fig. 12.2.

At the apex tip, 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. This can lead to the formation of a sharp metal whisker at the end of the tip.

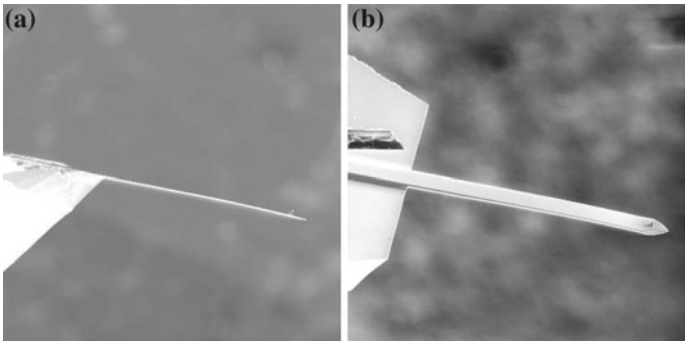

Fig. 12.2 Scanning electron microscopy images of a Si cantilever (length 450 μm) with a Si tip integrated at its end. A *side view* of the cantilever is shown in (**a**) and a *tilted view* in (**b**)

12.3 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 atomic force microscopes the beam deflection method is used. The basic setup of the beam deflection method is shown in Fig. 12.3. 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

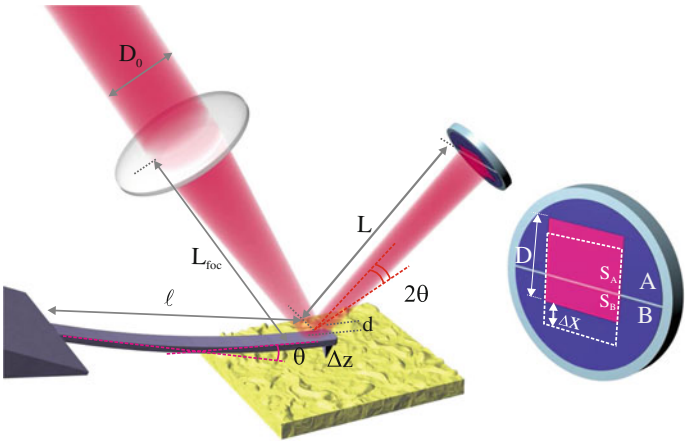

Fig. 12.3 Schematic of the beam deflection AFM method including the relevant distances necessary to calculate the sensitivity

absolute intensity detected by the photodiode can vary due to fluctuations of the laser intensity and depending 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 cantilever must be large 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 the tip, as normally done in STM. This is done because when scanning the cantilever the laser spot on the back of the cantilever would (in part) no longer focus on the cantilever.

12.3.1 Sensitivity of the Beam Deflection Method

In the following, the sensitivity for optical beam detection 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 usually converted to a (proportional) voltage at the output of the photodiode preamplifier electronics (transimpedance amplifier).

In the following, we will estimate the signal (current I) in the photodiode. We assume a total optical power of the laser diode of $S_0 = S_A + S_B$ and estimate (following Fig. 12.3) 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 [1]

$$\theta = \frac{3}{2}\frac{\Delta z}{l}. \tag{12.4}$$

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 point d is considered to be smaller than the cantilever width.

If we consider ray optics to determine 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 are related by

$$D = D_0\frac{L}{L_{\text{foc}}}. \tag{12.5}$$

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 diode D is given by diffraction ($\lambda = d \sin\alpha \approx dD/L$) as

$$D \approx \frac{\lambda L}{d}, \tag{12.6}$$

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 (12.6) 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 of the photodiode is proportional to the difference of the light intensities on both parts A and B of the split photodiode. The difference of the optical signals on both areas of the photodiode $S_A - S_B$ can be written according to the inset in Fig. 12.3 and using $\Delta x = 2\theta L$ as

$$S_A - S_B = S_{\text{area}} 2\Delta x D = \frac{S_0}{D^2} 4\theta L D. \tag{12.7}$$

If we insert now θ and D according to (12.4) and (12.6), the difference of the optical intensities at the photodiode results in

$$S_A - S_B = 6 S_0 \frac{\Delta z}{l} \frac{d}{\lambda}. \tag{12.8}$$

The electric current I in the photodiode 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. With this, the output current at the photodiode output as a function of deflection Δz can be written as

$$I = \frac{6 R S_0 d}{\lambda l} \Delta z. \tag{12.9}$$

The ratio of Δz and I is also called the detection sensitivity and is 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.

12.3.2 Detection Limit of the Beam Deflection Method

Up to now we have analyzed the magnitude of the photocurrent as a function of the external conditions such as 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 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 $R = I/(S_A - S_B)$ defined above as an equivalent quantity.

In the following, we estimate the fundamental limit in the noise of the photocurrent 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 a single measurement of the current is one second, the bandwidth is one hertz. If the measurement bandwidth is defined as $B = 1/\Delta t$, the measured current can be written as

$$I = eN\frac{1}{\Delta t} = eBN. \tag{12.10}$$

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

$$\Delta I_{\text{shot}} = eB\sqrt{N} = eB\sqrt{\frac{I}{eB}} = \sqrt{eBI}. \tag{12.11}$$

In our simplified explanation, a numerical factor of $\sqrt{2}$ is missing. In a statistically more rigorous derivation the following equation for the shot noise results

$$\Delta I_{\text{shot}} = \sqrt{2eIB}, \tag{12.12}$$

with I being the average signal current.

Identifying the photocurrent estimated above in (12.9) as signal S and the shot noise from (12.12) as the corresponding noise N, the signal-to-noise ratio is given by

$$\frac{S}{N} = \frac{I}{\Delta I_{\text{shot}}} = \frac{6d}{l\lambda} S_0 R \Delta z \frac{1}{\sqrt{2eS_0RB}}. \tag{12.13}$$

The smallest detectable cantilever displacement results as

$$\Delta z = \frac{l\lambda}{6d}\frac{S}{N}\sqrt{\frac{2eB}{S_0 R}}, \tag{12.14}$$

Now we discuss the dependence of the smallest detectable cantilever displacement on the different quantities involved. A laser beam with higher intensity S_0 will improve the detection sensitivity towards smaller Δz, 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 will increase by one, two, or three times. In this sense, the smallest detectable cantilever displacement is proportional to the signal-to-noise ratio required 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 $l/d \approx 10$, at a signal-to-noise ratio $S/N = 1$, the detection limit Δz of about 0.2 pm results. This shows that the detection limit is quite small. The simple beam deflection technique has a very high detection sensitivity.

In Chap. 18 we will also discuss other sources of noise in the measurement, such as the amplitude of the cantilever due to thermal excitation.

12.4 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. 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. 12.4 different methods used to measure the cantilever deflection are compared. The most widely used technique is the beam deflection method discussed in detail in the previous section. An advantage of this method is that it is easy to

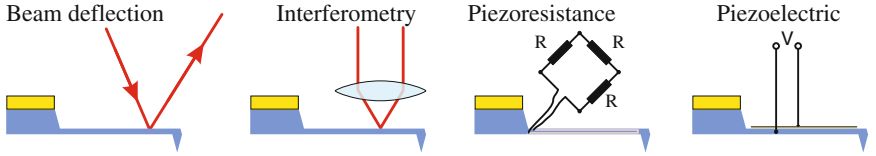

Fig. 12.4 Different kinds of deflection sensors in AFM

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. 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. One advantage of this technique is the easy absolute calibration of the cantilever motion by the wavelength of the light.

The piezoelectric detection method operates completely electrically and require minimal experimental effort for detection. Piezoresistive cantilevers are commercially available. They are realized by producing a piezoresistive layer on a cantilever. 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 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 the ease of use.

Piezoelectric cantilevers 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 cantilever oscillates close to the resonance frequency. The piezoelectric cantilever has two electrodes. One electrode can be used to excite the cantilever 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. 19.

12.5 Calibration of AFM Measurements

While the relation between the cantilever deflection Δz and the tip-sample force is easily given by Hooke' law[1] 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 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.

12.5.1 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 preamplifier. 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 (12.9). 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 up and down. The z-position corresponding to the voltage at the z-piezo element is obtained by multiplying this voltage by the corresponding piezo constant. Such a sensor voltage versus position curve is shown schematically in Fig. 12.5 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

[1] 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 [19] 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.

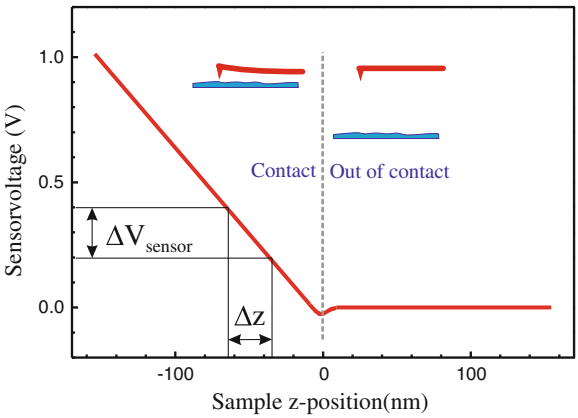

Fig. 12.5 Schematic of a typical sensor voltage versus sample z-position curve used to determine the sensitivity in AFM. The inverse slope measured in the contact regime gives the sensitivity as $S_{\text{sensor}} = \Delta z / \Delta V_{\text{sensor}}$

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.[2] 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 is concerned about the integrity of the tip, this calibration procedure should be done after the actual measurements have been completed. In Sect. 12.5.5 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. 19.

12.5.2 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. 12.6. The bending of the cantilever is out of the plane of the paper in Fig. 12.6a, while in Fig. 12.6b a side view is shown. The spring constant of a rectangular cantilever beam for the bending direction used in AFM is given by [1]

[2] Specific effects occurring at the kink between the two regions are discussed in Chap. 13.

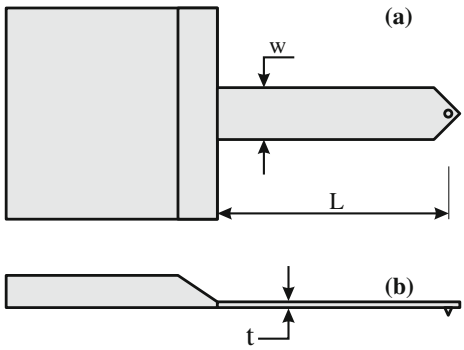

Fig. 12.6 Sketch of a rectangular cantilever together with the carrier chip on the left. **a** *Top view*, **b** *side view* including the dimensions of the cantilever

$$k = \frac{E w t^3}{4L^3}, \tag{12.15}$$

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 not easily measured. Unfortunately, this parameter enters with the third power into expression (12.15) 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 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 (12.15). Considering the effective mass of the rectangular cantilever $m_{\text{eff}} = 0.2357\,\text{m}$ from (2.46), the resonance frequency is written as

$$\omega_0 = \sqrt{\frac{k}{m_{\text{eff}}}} = \sqrt{\frac{k}{0.2357 \rho L w t}}. \tag{12.16}$$

Combining (12.16) and (12.15), t can be eliminated and the following expression for the spring constant is obtained

$$k = 0.239 w L^3 \omega_0^3 \sqrt{\frac{\rho^3}{E}}. \tag{12.17}$$

This approach can be extended to eliminate quantities which are not precisely known by other given or measured quantities 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.

12.5.3 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[3] ω_0 and Q, while excluding t and E [20]. The spring constant results as

$$k = 0.19\rho_f w^2 L Q_f \Gamma_i(Re)\omega_{0,f}^2. \tag{12.18}$$

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 [20]. 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.[4] There is also a relevant smartphone app (title: Sader method) 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 [21, 22].

12.5.4 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 [23]. Unlike the Sader method, it is not named after the developers, 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 an ideal harmonic oscillator is related to its static spring constant k by

$$\frac{1}{2}k\left\langle \Delta z_{\text{th}}^2 \right\rangle = \frac{1}{2}k_B T, \tag{12.19}$$

with $\left\langle \Delta z_{\text{th}}^2 \right\rangle$ 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

[3] 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 first resonance). Alternatively, the thermal noise spectrum can be measured, as described in the next section.

[4] The density and viscosity for the most frequently used fluids (air and water) are: $\rho_{\text{air}} = 1.2\,\text{kg/m}^3$, $\eta_{\text{air}} = 1.85 \times 10^{-5}\,\text{kg/(m s)}$, and $\rho_{\text{water}} = 1 \times 10^3\,\text{kg/m}^3$, $\eta_{\text{water}} = 8.9 \times 10^{-4}\,\text{kg/(m s)}$, respectively, under ambient conditions and at sea level [24].

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. How such measurements are performed in practice will be shown below. An advantage of the thermal method is that it cannot only be applied to cantilever-type sensors, but also to other types of sensors such as quartz sensors.

In the following, we will present several (more or less small) corrections which have to be applied if the determination of the force constant is not only done in principle but in reality. If you are not interested in the details, you can skip this part. We consider the most important case of rectangular cantilevers.

For an ideal harmonic oscillator represented in Fig. 12.7a by a mass and a spring, the expression (12.19) holds. However, a real rectangular cantilever beam (Fig. 12.7b) also has higher modes of oscillation. The first four modes of a cantilever beam are shown in Fig. 12.7c. For each higher mode one more node appears in the shape of the

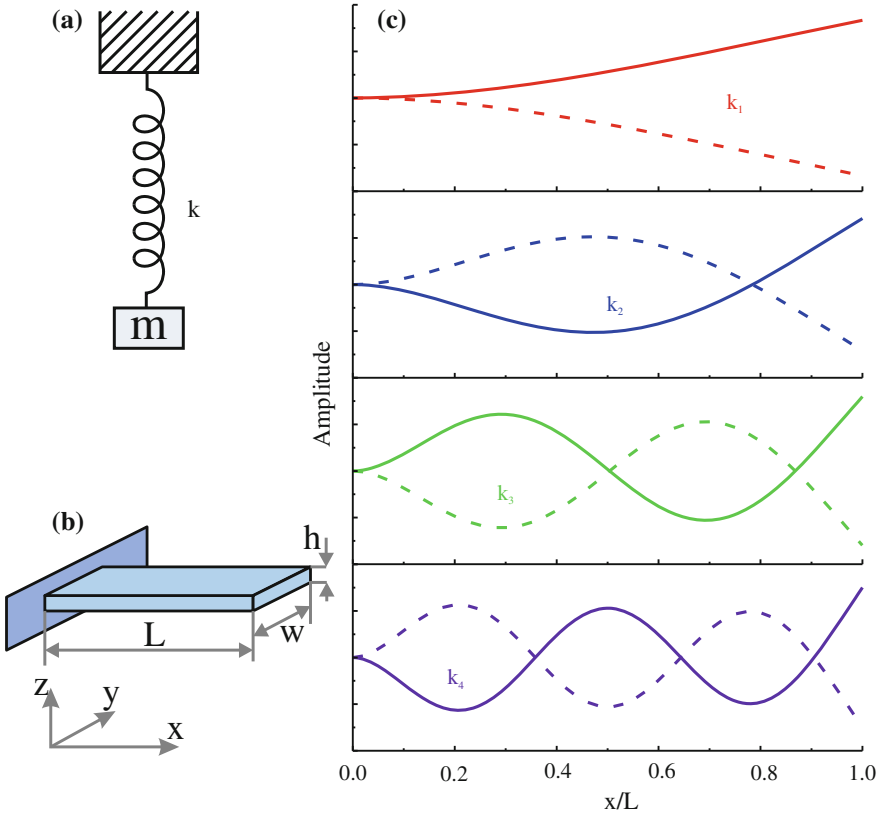

Fig. 12.7 **a** Ideal one-dimensional harmonic oscillator represented by a mass m on a spring with spring constant k. **b** Sketch of a cantilever-type beam. **c** The first four modes of a rectangular cantilever. A (dynamic) spring constant k_i can be assigned to each mode

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 $k_B T$. Thus in analogy to the ideal harmonic oscillator a (dynamic) spring constant k_i of the mode i can be defined by the relation

$$\frac{1}{2} k_i \left\langle \Delta z_{\mathrm{th},i}^2 \right\rangle = \frac{1}{2} k_B T, \tag{12.20}$$

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 [25] as

$$\left\langle \Delta z_{\mathrm{th},i}^2 \right\rangle = \frac{k_B T}{k} \frac{12}{\alpha_i} = \frac{k_B T}{k_i}, \tag{12.21}$$

with the values of α_i and correspondingly the (dynamic) spring constant k_i for each mode given in [25]. 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 $k_B T$, the spring constants for the higher modes increase significantly. Thus the thermally exited deflection for higher modes becomes very small for higher modes. Since the thermal excitation of the different modes are independent events, the total mean square thermal amplitude is the sum over the mean square amplitudes of all modes[5] $\left\langle \Delta z_{\mathrm{th}}^2 \right\rangle = \sum_0^\infty \left\langle \Delta z_{\mathrm{th},i}^2 \right\rangle$. 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 (12.19) is also recovered for a rectangular cantilever beam with the "static" spring constant k for a rectangular beam from (12.15) [25]. From (12.19) and (12.20) 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 in the harmonic oscillator.

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 this signal. The Fourier transformation of the square of the time-dependent noise signal is the power noise spectral density $N_{\mathrm{z,th}}^2(\omega)$, as introduced in Chap. 5. The noise spectral density is $N_{\mathrm{z,th}}(\omega) = \sqrt{N_{\mathrm{z,th}}^2(\omega)}$. 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.[6] 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

[5] 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.

[6] Details of how to extract the noise power spectral density from the time signal without using a spectrum analyzer are given in [23].

strength of the deflection noise of the first mode. In Chap. 18 it will be 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

$$N^2_{z,th,1} = N^2_{z,th,exc} G^2(\omega) = \frac{N^2_{z,th,exc}}{\left(1 - \frac{\omega^2}{\omega^2_{0,1}}\right)^2 + \frac{1}{Q_1^2}\frac{\omega^2}{\omega^2_{0,1}}}, \tag{12.22}$$

with $N^2_{z,th,exc}$ 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 $N^2_{z,th,exc}$, Q_1, and $\omega_{0,1}$ can be determined. The integral over $G^2(\omega)$ can be calculated and results as $\pi Q_1 \omega_{0,1}/2$ (compare Sect. 18.1). Thus using (12.20) the following additional relation results

$$\left\langle \Delta z_1^2 \right\rangle = \frac{1}{2\pi} \int_0^\infty N^2_{z,th,1}(\omega) d\omega = \frac{1}{2\pi} N^2_{z,th,exc} \frac{\pi Q_1 \omega_{0,1}}{2} = \frac{k_B T}{k_1}. \tag{12.23}$$

With this, the spring constant of the first mode results as

$$k_1 = 2\pi \frac{2k_B T}{\pi N^2_{z,th,exc} Q_1 \omega_{0,1}}. \tag{12.24}$$

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, which will be discussed in Sect. 19.3. If the cantilever spring constant is known from other sources, (12.23) can be used to determine the thermal oscillation amplitude $\left\langle \Delta z_1^2 \right\rangle$.

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. 12.5). 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 [23, 25, 26]. Moreover, for the case of the laser 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 $S_{sensor,end}$ has to be replaced by a corrected sensitivity $\chi\ S_{sensor,end}$ with the correction factor

$$\chi = \frac{S_{sensor,free,calc}}{S_{sensor,end,calc}} = \frac{\Delta z'_{free,calc}}{\Delta z'_{end,calc}}. \tag{12.25}$$

Thus the desired sensitivity factor for the free cantilever needed for the thermal method is given by

$$S_{\text{sensor,free}} = S_{\text{sensor,end,measured}} \frac{S_{\text{sensor,free,calc}}}{S_{\text{sensor,end,calc}}} = \chi \, S_{\text{sensor,end,measured}}. \qquad (12.26)$$

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 [26].[7]

12.5.5 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 [27]. 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 in order to obtain the sensitivity factor. The deflection noise density $N_{\text{z,th}}(\omega)$ given in (12.22) is related to the actually measured deflection *voltage* noise density $N_{\text{V,th}}(\omega)$ by $N_{\text{z,th}}(\omega) = N_{\text{V,th}}(\omega) S_{\text{sensor}}$. The thermal noise power spectral density $N^2_{\text{V,th}}(\omega)$ of the first mode can be measured using a spectrum analyzer. This experimentally measured noise density can be fitted (similar to (12.22)) by the function

$$N^2_{\text{V,th}}(\omega) = \frac{N^2_{\text{V,th,exc}}}{\left(1 - \frac{\omega^2}{\omega_0^2}\right)^2 + \frac{1}{Q^2}\frac{\omega^2}{\omega_0^2}}, \qquad (12.27)$$

[7] The sensitivity factor which we term S is called *InvOLS* in [26].

in order to determine ω_0, Q, and $N_{\mathrm{V,th,exc}}$. The deflection noise density $N_{\mathrm{z,th}}(\omega)$ can be obtained by multiplication by the still unknown sensitivity factor S_{sensor}. The procedure outlined for the thermal method can then be followed arriving at (12.24) with the only difference that $N_{\mathrm{z,th,exc}}$ should be replaced by $S_{\mathrm{sensor}}\ N_{\mathrm{V,th,exc}}$ (we skip the index 1 for the first mode). Thus we arrive at the following expression for the sensitivity factor[8]

$$S_{\mathrm{sensor}} = \sqrt{2\pi \frac{2k_B T}{\pi N_{\mathrm{V,th,exc}}^2 k Q \omega_0}}. \tag{12.28}$$

12.6 Summary

- Cantilever force sensors for atomic force microscopy should have a small spring constant in order to obtain a high force sensitivity and 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 micrometer dimensions.
- Cantilevers for atomic force microscopy are fabricated from silicon using wet etching technology 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).

[8] We keep the separate factor 2π in order to facilitate the conversion from the angular frequency ω to the natural frequency f, as $\omega = 2\pi f$.

Chapter 13
Static Atomic Force Microscopy

In static atomic force microscopy the force between the tip and 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. 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. 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.

13.1 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. 1.7. 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.

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_13

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 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. 13.1. In equilibrium, the total force on the cantilever has to vanish as

$$F_{\mathrm{tot}} = 0 = F_{\mathrm{ts}} + F_{\mathrm{cant}}, \tag{13.1}$$

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. For the force between individual pairs of tip and sample atoms, we consider the Lennard-Jones potential plotted once more in Fig. 13.2b. The direction of the force on individual tip atoms resulting from the interaction with the sample is shown by arrows in Fig. 13.2a. 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. 13.2), while the atoms slightly farther from the sample experience only an attractive interaction (blue in Fig. 13.2). The total tip-sample force is obtained by integration.

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,

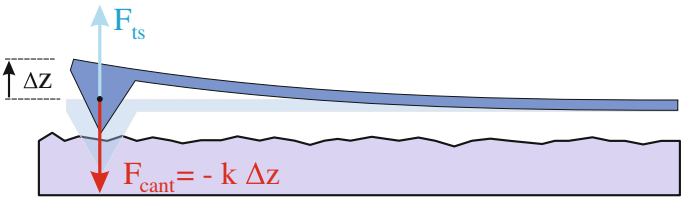

Fig. 13.1 Force equilibrium in static mode. The tip-sample force F_{ts} and the spring force of the cantilever compensate to a net vanishing force

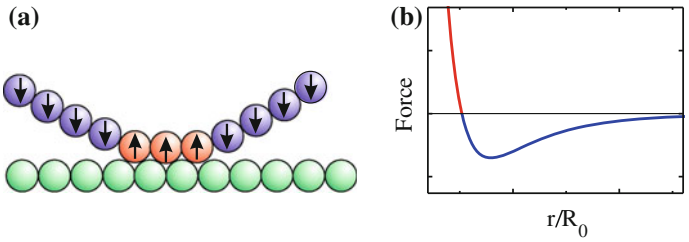

Fig. 13.2 Forces on the tip atoms due to interaction with the sample. For the atoms close to the surface the net interaction with the sample is repulsive (indicated in *red*). For larger distances to the sample, the interaction is attractive (indicated in *blue*)

if, for instance, the tip-sample distance changes by 1 Å. However, 1 Å change in tip-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. 13.2b) two tip-sample distances exist for a certain force. As discussed in Sect. 17.3 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.

13.2 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 (force gradient) of the sample atomic bonds. As discussed in Sect. 11.2, 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 does 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. 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 never 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 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.

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. Therefore, the tip atoms are not in registry with the sample surface structure and hence a superposition of tip and sample interactions, leading to an atomic resolution, is not expected. The usual understanding is that the atoms of the tip lock into the atomic lattice of the sample, so the atomic lattice of the sample is imaged. However, 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 destruction of soft (organic or biological) samples.

13.3 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 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 distance) and the change of the cantilever deflection is measured. In this case, no feedback is

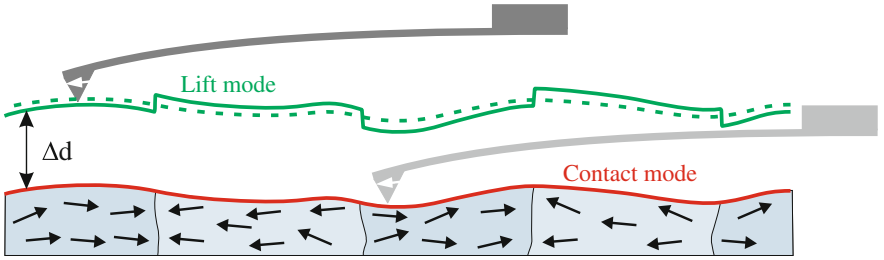

Fig. 13.3 Principle of the lift mode. In a first scan line, the topography is measured (contact mode). In a second scan line, the topography is retraced with an offset Δd (*dashed line*). The deflection due to the long-range magnetic interaction is measured relative to this retraced height (*solid line*)

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. 13.3. 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. 13.3 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 line in Fig. 13.3. 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.

13.4 Friction Force Microscopy (FFM)

When the tip is moved over the surface 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. 13.4. While a two electrode split photodiode was used in order to detect the vertical bending of the cantilever, quadrant photodiodes are used in order to measure also this torsional bending of the cantilever. 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 wear has taken place in the course of the experiment by subsequent imaging of the relevant area.

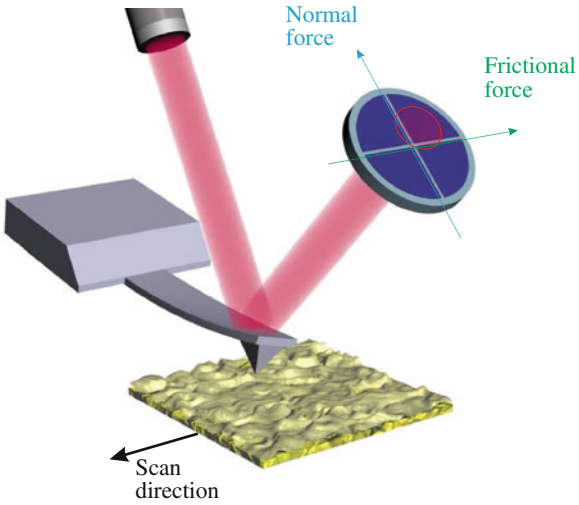

Fig. 13.4 Principle of the detection of frictional forces by the beam deflection method using a quadrant photodiode

13.5 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. 12.5. (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.

The aim is to obtain the tip-sample force $F_{\mathrm{ts}}(d)$ as a function of the tip-sample distance d, as indicated in Fig. 13.5. 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}. 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 as shown in Fig. 13.5. However, d can always be recovered as $d = z_{\mathrm{tip}} - z_{\mathrm{sample}}$. With the coordinate system in Fig. 13.5, 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

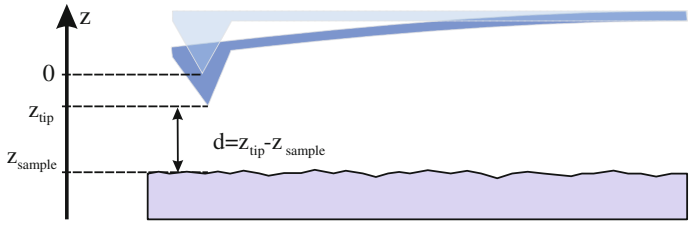

Fig. 13.5 Nomenclature for the coordinates used in force-distance curves

z_{tip} and z_{sample}, we choose the equilibrium position of the cantilever tip with the sample far away.

Figure 13.6a 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. 11.2). 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 tip-sample force vanishes. At this position, the cantilever is unbent ($z_{tip} = 0$). If the sample is pushed further towards the tip, the regime of repulsive tip-sample interaction is entered. In the repulsive regime the sample bends the cantilever upwards. As the repulsive force rises very sharply with decreasing tip-sample distance, both tip and sample move together ($\Delta z_{sample} \approx \Delta z_{tip}$ and $\Delta d \approx 0$) Specifically for a stiff sample with a high elastic modulus, the $z_{tip}(z_{sample})$ curve is a straight line with a slope of one, as shown in the left part of Fig. 13.6a. 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. 16).

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 to the position at which the maximum attractive force (adhesion force) between tip and sample acts. Generally, for elastic samples

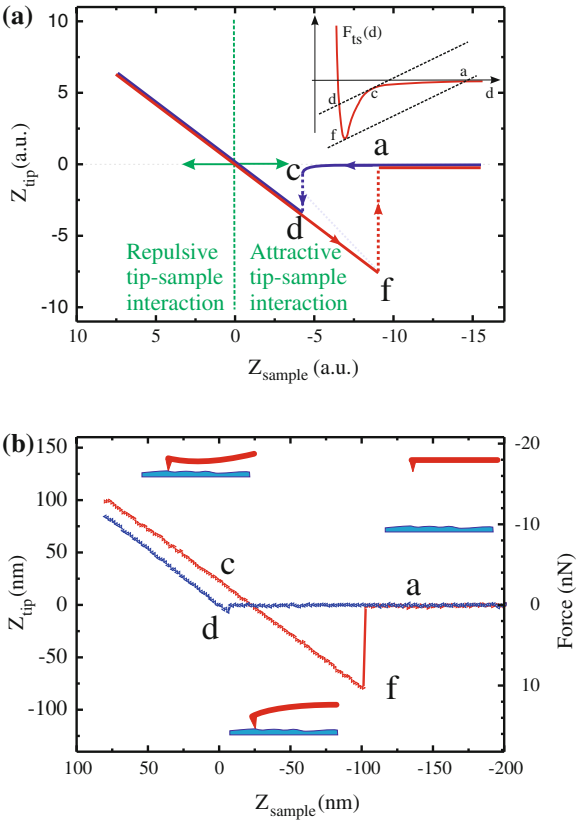

Fig. 13.6 **a** Schematic of a $z_{\text{tip}}(z_{\text{sample}})$ plot with the *blue curves* corresponding to an approach of the sample toward the tip, while the *red curves* correspond to a retraction of the sample. The nomenclature for the variables is the same as in Fig. 13.5. At points c and f, 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)

the retraction curve and the approach curve are the same in the repulsive regime, while the retraction curve lies below the approach curve for a plastic deformation of the sample.

In Fig. 13.6b 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 $F_{\text{ts}} = -kz_{\text{tip}}$) to a corresponding force, which is shown on the right axis in Fig. 13.6b. 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 about 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 of the force-distance curve can be extracted directly from the force-distance curve. In one respect, the measured force-distance curve does not follow the idealized expectation shown in Fig. 13.6a. The approach curve (blue) and the retract curve (red) do not coincide for positive sample distances in Fig. 13.6b. This effect arises due to hysteresis and creep effects of the piezoelectric actuators. For a quantitative analysis of the force-distance curves, those effects have to be carefully excluded.

In principle, the measured $z_{\text{tip}}(z_{\text{sample}})$ curve or the $F_{\text{ts}}(z_{\text{sample}})$ curve (right axis in Fig. 13.6b) 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. 13.6a, 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 larger 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 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. 13.6a, 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.[1] Due to these limitations, force-distance curves are not used to measure the fundamental forces.

[1] The influence of capillary forces can in principle be estimated by comparing $z_{\text{tip}}(z_{\text{sample}})$ plots in air and water. If the cantilever is fully immersed in water, capillary forces can be excluded.

13.6 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.
- Frictional 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.

Chapter 14
Amplitude Modulation (AM) Mode in Dynamic Atomic Force Microscopy

In dynamic atomic force microscopy the cantilever is excited using a piezo actuator which oscillates the cantilever base. The driving frequency is usually close to the resonance frequency of the cantilever. 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 so called frequency modulation mode (FM) of atomic force microscopy, as described in Chap. 17. In this chapter, we describe the amplitude modulation mode (AM) of AFM. Here the cantilever is driven at a fixed frequency with a fixed driving amplitude. The change of the resonance frequency leads to a change of the vibration 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 complicated case is then easier to comprehend.

14.1 Parameters of Dynamic Atomic Force Microscopy

Compared to STM which has only two parameters, the tunneling current and the tunneling voltage, 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, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_14

- 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 amplitude of the oscillation A_{drive}
- The oscillation amplitude A
- The phase ϕ between driving and oscillation
- The driving frequency ω_{drive}
- 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} and A_{drive} are set.
- In FM detection the shift of the resonance frequency $\Delta\omega$ is measured.

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.

14.2 Principles of Dynamic Atomic Force Microscopy I (Amplitude Modulation)

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. 2.3 including the influence of a time-independent external 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} varies only slowly in the range of the oscillation amplitude A. In this case, F_{ts} will be approximated as linear in the following. We use this limit here because this idealized scenario can be solved analytically. For the usual vibration amplitudes (several nanometers) the small amplitude limit does not hold.

The definition of the coordinates of the cantilever-tip-sample system is given in Fig. 14.1. For the tip oscillation, we use the coordinate z. For the tip-sample force $F_{\text{ts}}(d+z)$, we use the coordinate $d+z$ (tip-sample distance), with the offset d being the average tip-sample distance during an oscillation cycle.

Due to the small amplitude assumption, we can expand the force $F_{\text{ts}}(d+z)$ around the equilibrium position of the tip ($z = 0$, corresponding to a tip-sample distance d) as

$$F_{\text{ts}}(d+z) = F_{\text{ts}}(d) + \left.\frac{\partial F_{\text{ts}}}{\partial z}\right|_{z=0} z + \cdots . \tag{14.1}$$

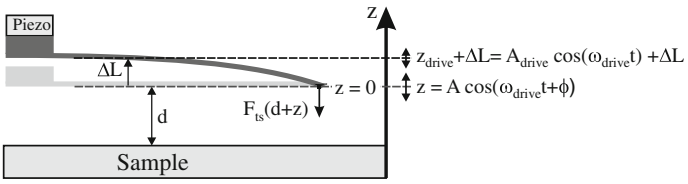

Fig. 14.1 Definition of the coordinates for a driven damped harmonic oscillator under the influence of a tip-sample force

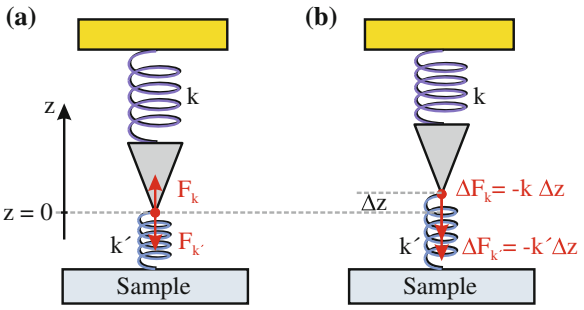

Fig. 14.2 **a** For the case of small amplitudes, the cantilever-tip-sample system can be effectively described by two springs, one representing the cantilever with force constant k and one representing the tip-sample interaction with force constant k'. **b** This system is equivalent to a system with an effective spring constant of $k_{\text{eff}} = k + k'$

In this approximation 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

$$k' = -\left.\frac{\partial F_{\text{ts}}}{\partial z}\right|_{z=0}. \tag{14.2}$$

The tip-sample interaction can be represented by adding a small spring with spring constant $k' \ll k$, as shown in Fig. 14.2a. The two spring constants add up[2] to an effective spring constant $k_{\text{eff}} = k + k'$. However, this 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

[2] 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 up. As indicated in Fig. 14.2 the cantilever spring under the influence of a tip-sample force can be replaced by a cantilever effective mass held by two springs. 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. 14.2b 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 $k_{\text{eff}} = k + k'$.

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.

Before we analyze the harmonic oscillator with the spring constant k_{eff}, we consider the static case (i.e. all oscillatory amplitudes in Fig. 14.1 are zero). Without a sample being present, the tip is at its zero position $z = 0$ and the cantilever is unbent (shown in light gray in Fig. 14.1). In this case the static bending ΔL is zero.[3] If the sample is now brought close to the tip, the tip-sample interaction will change the tip position. Since we would like to probe the sample 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. 14.1. In static equilibrium with the cantilever bent, the tip-sample force and the static bending force balance at $z = 0$ as

$$F_{\text{ts}}(d) = -k_{\text{eff}} \Delta L, \tag{14.3}$$

with ΔL being the static (offset) deflection of the cantilever as indicated in Fig. 14.1.

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 as $z_{\text{drive}} = A_{\text{drive}} \cos(\omega_{\text{drive}} t)$. As a result of this excitation, the tip will oscillate in the steady-state around its equilibrium position as $z = A \cos(\omega_{\text{drive}} t + \phi)$. This case corresponds to the driven damped harmonic oscillator discussed in Sect. 2.3 and using (2.17) the equation of motion can be written as

$$\ddot{z} + \sqrt{\frac{k_{\text{eff}}}{m}} \frac{1}{Q_{\text{cant}}} \dot{z} + \frac{k_{\text{eff}}}{m} (z - z_{\text{drive}}) = 0. \tag{14.4}$$

The tip-sample force is included by replacing the spring constant k by k_{eff}. As the force $F_{\text{ts}}(d)$ cancels out the force due to the static bending of the cantilever $-k_{\text{eff}} \Delta L$, according to (14.3), these terms have already be removed from the equation of motion. The equation of motion (14.4) was solved in Sect. 2.3 with the result that a resonance occurs at $\omega_0 = \sqrt{k/m}$. Since we replaced k by the effective spring constant k_{eff} in order to include the effect of a tip-sample force, the resonance frequency will shift from ω_0 for the case without tip-sample interaction to $\omega_0' = \sqrt{k_{\text{eff}}/m}$. Thus

$$\omega_0' = \sqrt{\frac{k_{\text{eff}}}{m}} = \sqrt{\frac{k + k'}{m}} = \sqrt{\frac{k}{m}\left(1 + \frac{k'}{k}\right)} = \omega_0 \sqrt{1 + \frac{k'}{k}}. \tag{14.5}$$

In the following, we assume that $|k'| \ll k$. For small x the approximation $\sqrt{1+x} \approx 1 + \frac{1}{2}x$ holds. Therefore, the new resonance frequency of the cantilever can be written as

$$\omega_0' \approx \omega_0 \left(1 + \frac{k'}{2k}\right). \tag{14.6}$$

[3] The tip length is set to zero in order to avoid an additional offset length.

The shift of the resonance frequency results in

$$\Delta\omega = \omega_0' - \omega_0 = \omega_0 \frac{k'}{2k} = -\frac{\omega_0}{2k} \frac{\partial F_{\mathrm{ts}}}{\partial z}\bigg|_{z=0}. \tag{14.7}$$

This result can be easily related to the experimentally observed frequency shift Δf as

$$\Delta f = \frac{\omega_0' - \omega_0}{2\pi} = f_0 \frac{k'}{2k} = -\frac{f_0}{2k} \frac{\partial F_{\mathrm{ts}}}{\partial z}\bigg|_{z=0}. \tag{14.8}$$

Together with the resonance frequency (maximum of the resonance curve) 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 ($F_{\mathrm{ts}}'(d) = \partial F_{\mathrm{ts}}(d+z)/\partial z|_{z=0}$) 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 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. 14.3. 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 $F_{\mathrm{cant}} = -kz$ is shown as a green line. The tip-sample force is approximated locally around $z = 0$ as linear $\Delta F_{\mathrm{ts}} = -k'z = \partial F_{\mathrm{ts}}/\partial z|_{z=0}\, z$, which is indicated by the dashed blue line. The resulting total force is shown as a red line with a slope of $k_{\mathrm{eff}} = k + k'$. Since $k' < 0$ and $|k'| \ll k$, the spring constant of the cantilever spring constant k is reduced by $|k'|$ comparing the green and red lines.

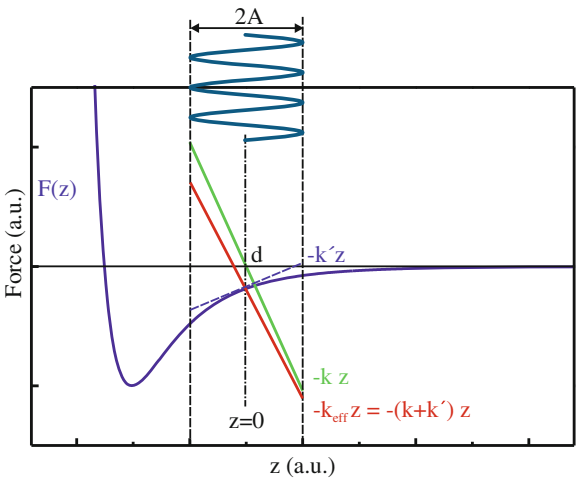

Fig. 14.3 In the small amplitude limit, the tip-sample force is approximated as linear within the range of the oscillation proportional to $-k'$. In this figure $k' < 0$ at the tip-sample distance d and the cantilever spring constant k is always positive. Thus the total effective force constant is the cantilever spring constant k reduced by the tip-sample force gradient proportional to k'

For a positive tip-sample force gradient $\partial F_{\mathrm{ts}}/\partial z = -k' > 0$ the resonance frequency will shift to lower values $\Delta f < 0$, while for a negative force gradient $\partial F_{\mathrm{ts}}/\partial z = -k' < 0$ the resonance frequency will shift to higher values $\Delta f > 0$. The frequency shift does not depend on the constant static offset force $F_{\mathrm{ts}}(d)$. 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 (14.3).

Often it is stated slightly imprecisely that the frequency shift Δf is positive (towards higher frequencies relative to ω_0) for repulsive forces and negative for attractive forces. We can understand this if we have a closer look at Fig. 14.4, 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. 14.4). 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. 14.4, respectively) have the same sign. Only for a small range of distances (shaded gray in Fig. 14.4) 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 (14.8). 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

$$\frac{\Delta f}{f_0} = \frac{k' A^2}{2kA^2} = \frac{E_{\mathrm{interaction}}}{2E_{\mathrm{cantilever}}}. \tag{14.9}$$

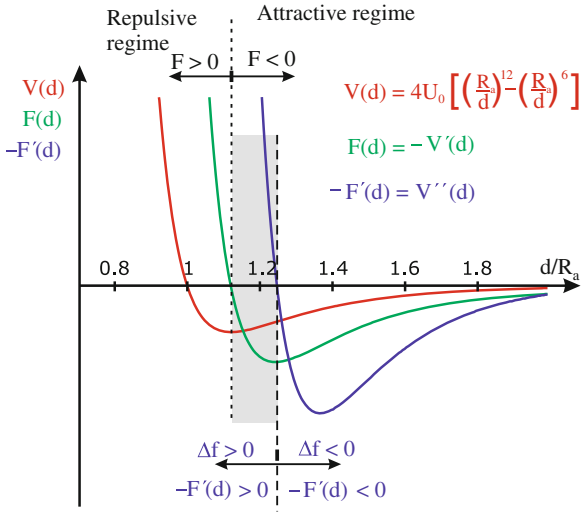

Fig. 14.4 Potential, force and negative force gradient for the Lennard-Jones model potential shown as a function of the average tip-sample distance d. As the frequency shift Δf is proportional to the negative force gradient it can be stated: For distances outside the *shaded region* the frequency shift Δf is positive (towards higher frequencies relative to ω_0) for repulsive forces, and negative for attractive forces

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).

14.3 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. 14.5b.

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. 14.5, this amplitude depends indirectly on the tip-sample distance. The amplitude 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 $A(\Delta\omega(F'_{\text{ts}}(d)))$.

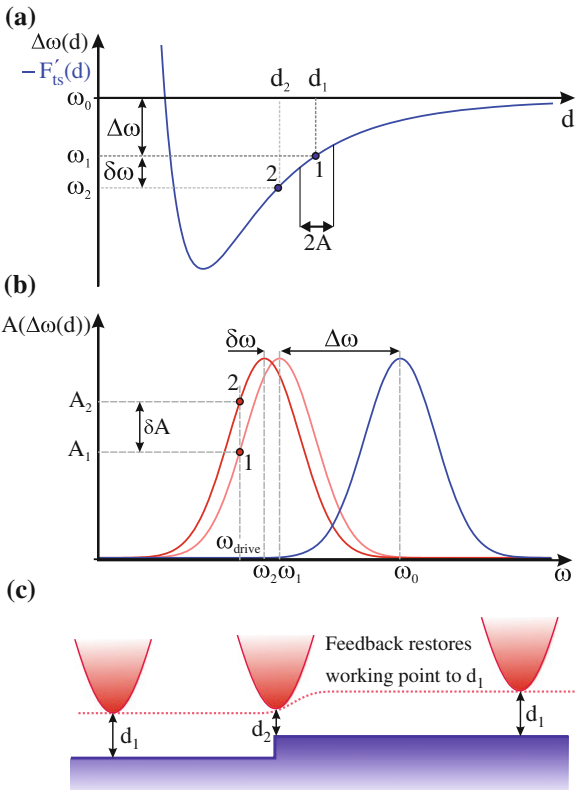

Fig. 14.5 In dynamic AFM the measured signal depends indirectly on the tip-sample distance. **a** Primarily, the force gradient and therefore also the resonance frequency (shift) depend on the tip-sample distance (here a Lennard-Jones potential is assumed). **b** Secondly, the measured amplitude depends on the frequency shift. For clarity $\Delta\omega_0$ has been chosen to be large compared to the width of the resonance curve. **c** When scanning over a step edge, the tip-sample distance changes until the feedback restores the old tip-sample distance

In the following, we go through these dependence step by step. The dependence of the force gradient on the tip-sample distance $F'_{ts}(d)$ based on the Lennard-Jones model potential is shown in Fig. 14.5a. 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. 14.5a. In Fig. 14.5b 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 is the measurement signal. In the feedback loop for the amplitude signal, a setpoint amplitude is selected, e.g. A_1 in Fig. 14.5b. The feedback loop controls the measured amplitude to the setpoint value by changing the z-position of the 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. If the feedback loop maintains a constant oscillation amplitude throughout a scan, this corresponds to a height profile taken at constant force gradient. Due to the dependence of the amplitude on the slope of the resonance curve the AM detection scheme is also called slope detection. 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 $\omega_{\text{drive}} < \omega_0$, corresponding to a negative force gradient (roughly: 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 force gradient (negative 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. 14.5c. 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. 14.5a, 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 (working point 2 in Fig. 14.5). This shift of the resonance frequency by $\delta\omega$ leads to an increase of the amplitude by δA to A_2 at ω_{drive}, as shown in Fig. 14.5b. 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 $A(\Delta\omega(F'_{ts}(d)))$. Therefore, keeping the feedback loop at a constant oscillation amplitude corresponds to establishing a constant tip-sample distance. An image scanned at constant tip-sample distance is called the topography. However, this assignment is only true if the *same* frequency shift-distance relation (Fig. 14.5a) is present all over the surface.

Let us now consider scanning over a border with two different dependences of the frequency shift as a function of tip-sample distance as shown in Fig. 14.6. This will lead to an apparent height contrast even if the actual height of the atoms in both areas is the same. Initially the tip is in region A with the corresponding force gradient dependence shown in Fig. 14.6b. The setpoint frequency ω_1 stabilizes the tip-sample distance to d_A (working point 1). If by lateral scanning the tip crosses

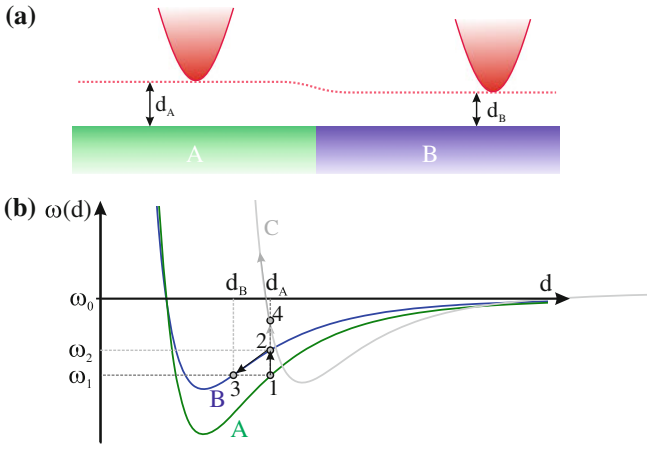

Fig. 14.6 **a** A scan from a region with material A to a region with material B can lead to a different apparent tip height in atomic force microscopy. **b** This arises due to the different force gradient-distance curves present in the two regions. For another force gradient-distance curve *C* an instability will occur due to the different sign of the slope of the force gradient, i.e. due to the non-monotonous character of the force gradient on the tip-sample distance

the border from A to B, the force gradient curve B in Fig. 14.6b 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 to d_B (working point 3). This leads to a reduced apparent height d_B as shown in Fig. 14.6a. It is similar to the electronic effects in scanning tunneling microscopy.

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. 14.6b. 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 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.

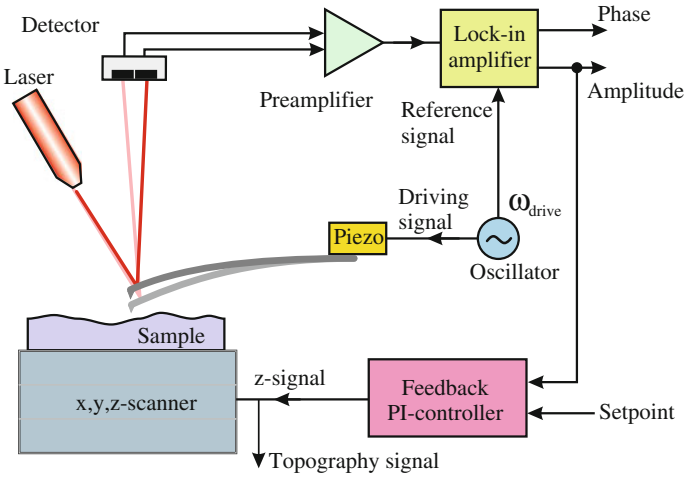

Fig. 14.7 Experimental setup for the AM detection scheme using a lock-in amplifier to detect the deviation of the oscillation amplitude from the setpoint value

14.4 Experimental Realization of the AM Detection Mode

A scheme of the experimental setup for the amplitude modulation AFM detection is shown in Fig. 14.7. The sinusoidal driving signal at ω_{drive} is generated by an oscillator. This signal excites the piezoelectric actuator driving the cantilever base.

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 a tube piezo element, 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. The cantilever excitation results in a cantilever oscillation of 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. 14.7. 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 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.[4]

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 operation of the feedback in more detail by considering the example of a scan over a step edge. As a starting condition, we assume that before scanning over a step edge the amplitude is nicely kept closely to the amplitude setpoint value. When the step is approached laterally, the tip-sample distance will decrease. This leads, as discussed in the last section, to a deviation of the oscillation amplitude (from the setpoint amplitude) which is measured at the output of the lock-in amplifier. Thus this quasi-DC amplitude signal contains the deviations from the setpoint amplitude (e.g. due to the topography of the surface) before they are compensated by the feedback. Subsequently, this measured amplitude enters the feedback controller and deviations from the setpoint are compensated by changing the z-signal to a value equivalent to the step height. After this, the setpoint oscillation amplitude (corresponding to a certain tip-sample distance) 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 structure 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 later in Chap. 15. Less frequently, the phase signal is used as a feedback signal and the oscillation amplitude is recorded as a free signal.

The setup shown in Fig. 14.7 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 parameters like the resonance frequency ω_0, the Q-factor, and the amplitude at the resonance frequency $A(\omega_0) = A_{\text{free}}$ to be determined. 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

[4] 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). Then the task of the lock-in amplifier is the demodulation of the low frequency amplitude signal. The term demodulation is traditionally used in connection with signal detection in AM radio receivers. This is the reason why the term AM detection is used for this detection scheme.

minimal detectable amplitude change. The slope of the resonance curve increases with increasing Q-factor. Thus, in AM detection the sensitivity with which a frequency shift can be detected 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 bandwith). Due to this the AM detection scheme is not used for cantilevers with Q-factors larger than about 500.

14.5 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 (2.17). 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. 2.4 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

$$z(t > 0) = A' \cos(\omega_{\text{drive}} t + \phi') + G e^{-\omega_0' t/(2Q)} \cos(\omega_{\text{hom}} t + \phi). \tag{14.10}$$

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 (2.25) and (2.28), respectively, replacing ω_0 by ω_0'. The second term in (14.10) corresponds to an exponentially decreasing transient. G and ϕ are determined by the initial conditions and ω_{hom} is introduced in Sect. 2.4.

In Fig. 14.8a 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.[5] The envelope of the cantilever deflection $z(t)$ is plotted, since a single oscillation is not 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 oscillations, corresponding to a time $\tau \approx Q/(f_0\pi) = 0.2\,\text{ms}$ (cf. (2.36)). This time constant still allows for fast scanning speeds in AFM scanning.

In Fig. 14.8b 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 oscillations.

[5] This value for the frequency shift was chosen as it leads to half of the original amplitude in the steady-state.

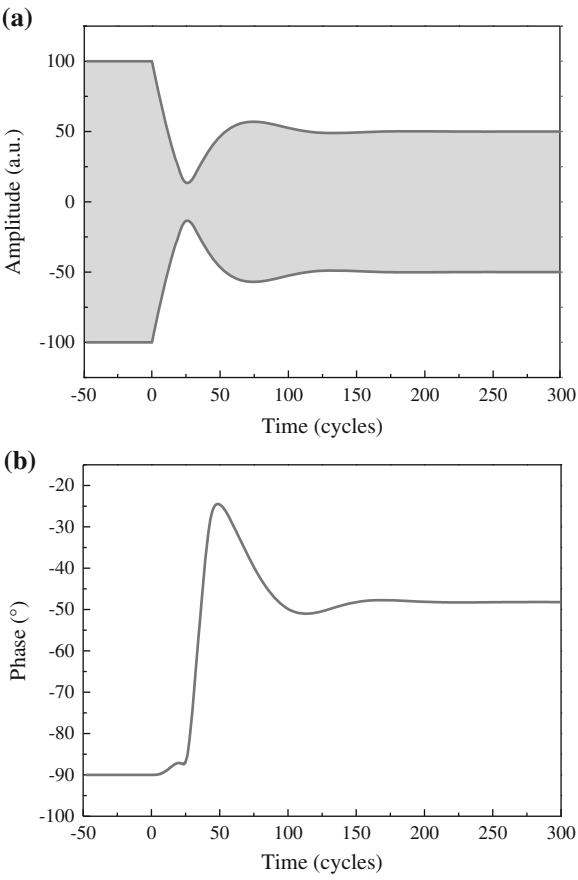

Fig. 14.8 The envelope of the oscillation amplitude (**a**) and the phase (**b**) in reaction to a change of the resonance frequency from ω_0 to ω_0' at time $t = 0$. The amplitude and phase response show that, after a transient, the new steady-state amplitude and phase are reached after about Q oscillations

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 several times 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 can 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. 17. The AM detection scheme is used for cantilevers under ambient conditions, where the quality factor is less than several hundred due to dissipative damping in air.

14.6 Dissipative Interactions in Non-contact AFM in the Small Amplitude Limit

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 very 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 simple harmonic oscillator, dissipation was included by the Q-factor. The types of dissipative forces included via the Q-factor are: energy losses (damping) if the cantilever oscillates in air or a liquid, as well as internal energy losses in the cantilever material (i.e. the cantilever itself is not 100 % elastic). This cantilever dissipation energy $E_{\text{cant}}^{\text{diss}}$ leads according to (2.41) to a corresponding Q-factor $Q_{\text{cant}} \propto 1/E_{\text{cant}}^{\text{diss}}$. An additional dissipative tip-sample interaction leads to a dissipated energy per cycle of $E_{\text{ts}}^{\text{diss}}$ and a corresponding Q-factor Q_{ts}. As the dissipation energies add up to a total dissipation energy, the inverse Q-factors add up to an effective Q-factor as

$$\frac{1}{Q_{\text{eff}}} = \frac{1}{Q_{\text{cant}}} + \frac{1}{Q_{\text{ts}}}. \tag{14.11}$$

This is not the proper way to include tip-sample dissipation, as the Q-factor takes into account only the continuous damping of the cantilever in a fluid (2.17). This damping force was considered proportional to the velocity, having its maximal value at zero amplitude of the oscillation, while the dissipative tip-sample interaction should be maximal at the lower turnaround point of the tip, i.e. closest to the sample. Nevertheless, we will now consider the damping via the effective Q-factor, since in this case we can still use the previously derived equations for the amplitude and the phase (2.25) and (2.27) of a driven damped harmonic oscillator. We use the effective quality factor and replace the resonance frequency of the free cantilever ω_0 by the shifted resonance frequency $\omega_0' = \omega_0 + \omega_0 k'/(2k)$, according to (14.6). 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 as

$$A^2 = \frac{A_{\text{drive}}^2}{\left(1 - \frac{\omega^2}{\omega_0'^2}\right)^2 + \frac{1}{Q_{\text{eff}}^2}\frac{\omega^2}{\omega_0'^2}}. \tag{14.12}$$

and

$$\tan\phi = \frac{-\omega_0'\omega}{Q_{\text{eff}}\left(\omega_0'^2 - \omega^2\right)}, \tag{14.13}$$

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 the two limiting cases of only conservative interaction or only dissipative interaction.

In Fig. 14.9a the amplitude and phase for a free cantilever (blue curve: ω_0, Q) are compared to the case in which a conservative tip-sample interaction is included (red curve: ω_0', Q). In this case, the conservative tip-sample interaction leads to a shift of the whole resonance curve.[6] Due to the constant quality factor, the amplitude and shape of the resonance curve and phase 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. 14.9a. In this figure, the driving frequency was selected to be somewhat larger than ω_0.

The opposite assumption is that only the damping changes and the resonance frequency stays constant (ω_0, Q'). 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. 14.9b, 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 line in Fig. 14.9b. This leads to a reduced amplitude and also to a change of the phase shift at the driving frequency (vertical line in Fig. 14.9b).

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) or due to a dissipative interaction (change of the Q-factor). Both lead to a change of the amplitude at the driving frequency. It is not known whether an initial change of A during a scan (later balanced by the feedback loop) arises due to a change of $\Delta\omega$ or Q.

The dependence of the amplitude on Q 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 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 (14.12) and (14.13) for ω_0' and Q_{eff}. Since (14.12) and (14.13) are a rather complicated to solve, alternatively the complete resonance curves of amplitude and phase (like the ones shown in Fig. 14.9) 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 $-90°$ so that the force gradient can be determined. The damping Q_{eff} can be determined from the width of the resonance curve in amplitude or phase.

[6] The curves in Fig. 14.9 are plotted using (14.12) and (14.13). The resonance curves for two different resonance frequencies do not exactly correspond to a shift of the resonance curve. However, Fig. 14.9a shows that these curves correspond to a very good approximation to a shift.

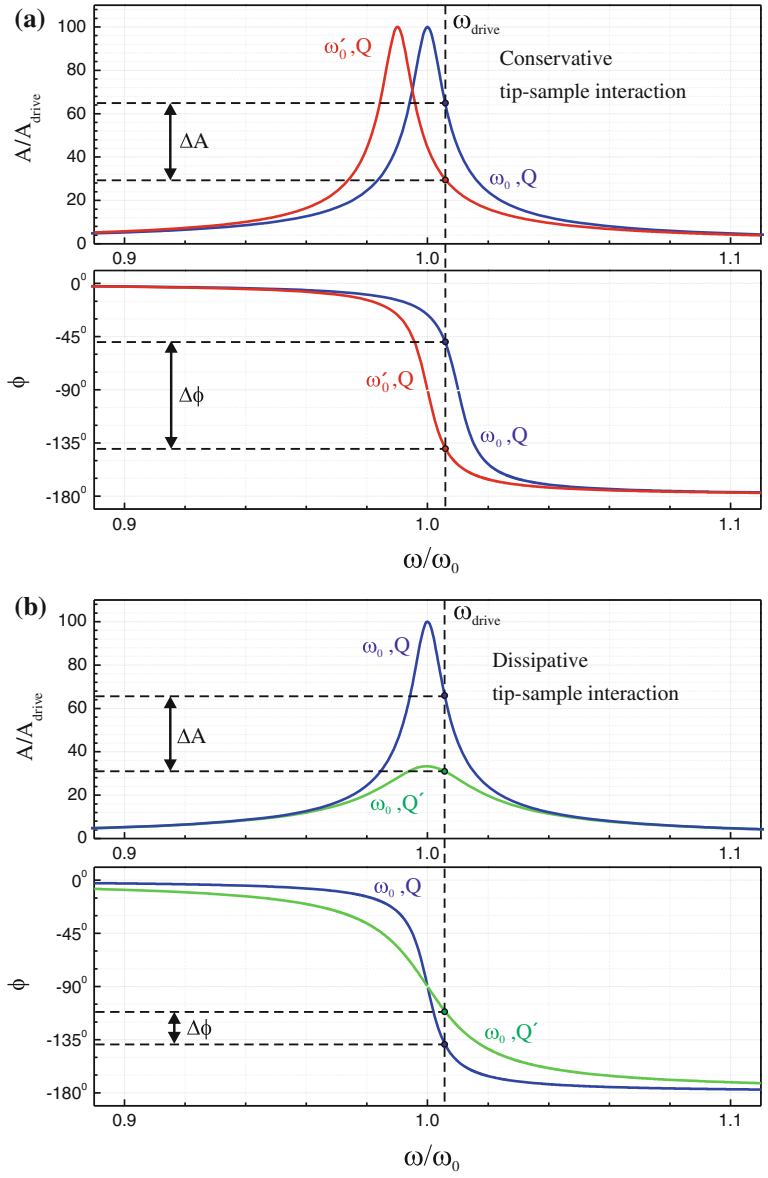

Fig. 14.9 **a** Amplitude and phase for a free cantilever (*blue curve*) compared to the case with a conservative tip-sample interaction included (*red curve*). The two resonance curves as well as the phase curves are shifted with respect to each other by $\Delta\omega$. **b** Amplitude and phase for a free cantilever compared to the case with a dissipative tip-sample interaction included (*green line*), i.e. the effective quality factor is different, while the frequency shift stays constant. In both cases (**a**) and (**b**) the oscillation amplitude at ω_{drive} is reduced, which makes it impossible to distinguish between a conservative and a dissipative interaction during scanning in the AM detection mode

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.

14.7 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 (14.13). From Fig. 14.9, we can see that the dependence of the phase as function of frequency can be approximated as linear close to the resonance at $\omega = \omega_0$ or $\phi = -90°$. In the following, we will derive this linear relation between phase and frequency. Using in the nominator of (14.13), the approximation $\omega_0' \approx \omega_0$ and in the denominator the approximation $\omega_0' + \omega_0 \approx 2\omega_0$, as well as subsequently the relation $\Delta\omega = \omega_0' - \omega$, results in

$$\tan\phi = \frac{-\omega\omega_0'}{Q_{\text{eff}}\left(\omega_0'^2 - \omega^2\right)} \approx \frac{-\omega_0^2}{Q_{\text{eff}}\left(\omega_0' + \omega\right)\left(\omega_0' - \omega\right)} \approx \frac{\omega_0}{2Q_{\text{eff}}\Delta\omega} = \frac{k}{Q_{\text{eff}}k'}. \tag{14.14}$$

Close to the resonance, the phase will be close $-\pi/2$ 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

$$\phi = -\frac{\pi}{2} + \Delta\phi = \arctan\left(\frac{\omega_0}{2Q_{\text{eff}}\Delta\omega}\right) \approx -\frac{\pi}{2} - \frac{2Q_{\text{eff}}}{\omega_0}\Delta\omega. \tag{14.15}$$

Thus the phase shift $\Delta\phi$ relative to the phase $-90°$ results as

$$\Delta\phi = -\frac{2Q_{\text{eff}}}{\omega_0}\Delta\omega = -\frac{Q_{\text{eff}}k'}{k} = \frac{Q_{\text{eff}}}{k}\frac{\partial F_{\text{ts}}}{\partial z}\bigg|_{z=0}. \tag{14.16}$$

This equation can be used for conversion between the frequency shift and the phase shift close to resonance. 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, 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 (14.16), the sign of the force gradient determines the sign of the phase shift, 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 (more precisely, negative force gradients) the relation $\phi > -90°$ holds for the phase.

14.8 Summary

- If the tip oscillation amplitude is small, the tip-sample interaction can be described by a second small spring 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

$$\Delta\omega = \omega_0 \frac{k'}{2k} = -\frac{\omega_0}{2k} \frac{\partial F_{\mathrm{ts}}}{\partial z}\bigg|_{z=0}. \tag{14.17}$$

  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 tip-sample interaction. 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 $\Delta\phi = -\frac{2Q_{\mathrm{eff}}}{\omega_0}\Delta\omega$. Thus the phase shift depends linearly on Q_{eff} and the force gradient.

Chapter 15
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 the mode that is used most frequently under ambient conditions. In the intermittent contact mode the oscillation amplitude is large compared to the range of the 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, an analytical solution of the equation of motion becomes difficult and we derive general dependences (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 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. 14. 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. Finally, we discuss how the dissipative tip-sample interactions are related to the phase of the oscillation in the intermittent contact mode.

15.1 Atomic Force Microscopy with Large Oscillation Amplitudes

In the intermittent contact mode, the oscillation amplitude is quite large (typically 50 nm) and cantilever force constants of typically 50 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 contributions close to the sample. In tapping mode, the constant driving frequency is usually selected at or very

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_15

close to the resonance frequency of the free cantilever (not at maximum slope, as in the slope detection mode). 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. 11.2.

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 15.1 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 spring force is linear with z. However, close to the lower turnaround point 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 last chapter.

In the following, we first describe the tapping mode qualitatively and subsequently discuss the results of an analytical or numerical treatment of the equation of motion. We now consider bringing a tip from a large tip-sample distance, where it oscillates at its free resonance frequency ω_0 with its (large) free amplitude A_{free}, towards the surface. The tip will eventually reach the repulsive interaction and is hindered

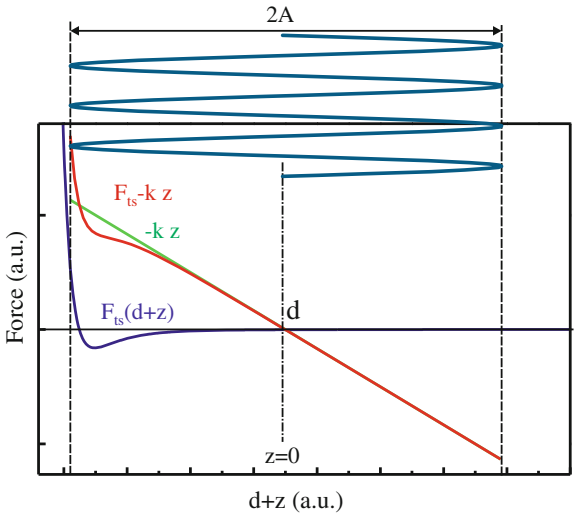

Fig. 15.1 Force-distance dependence of the cantilever force (*straight green line*), the tip-sample force (*blue line*), and the total force (*red line*) as a function of the momentary tip-sample distance $d + z$. In tapping mode, the range of the amplitude $2A$ is so large that it extends from almost 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 sinusoidal

from further indenting into the sample (tapping to the surface). It might be assumed that the trajectory of an oscillation for a very steep tip-sample force should deviate strongly from a sinusoidal shape. However, it appears (experimentally [28] and from simulations [29]) 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.

As an example, the oscillation traces for two different average tip-sample distances d are shown in Fig. 15.2a, when operation is performed in constant height mode, i.e. without feedback, restoring an amplitude setpoint. It was found experimentally and also from simulations that the oscillation amplitude as reduced approximately linearly when decreasing the average tip-sample distance d, once the oscillation path reaches the repulsive regime, as shown in Fig. 15.2b. In tapping mode detection, a certain amplitude A (corresponding to a certain average tip-sample distance d) is chosen as the amplitude setpoint for the z-feedback.

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. 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. Sect. 17.3).

In the following, we will provide a qualitative and a semiquantitative explanation for the amplitude reduction if the oscillation enters the regime of strong (repulsive) interaction. We will not invoke the concept of frequency shift of the whole resonance curve, since this applies only to the case of a linear force-distance dependence around d, which does not hold in the tapping mode. Further, it is important to understand that no dissipative tip-sample interaction is needed in order to reduce the oscillation amplitude. The amplitude reduction can be understood within the model of a driven oscillator (not harmonic).

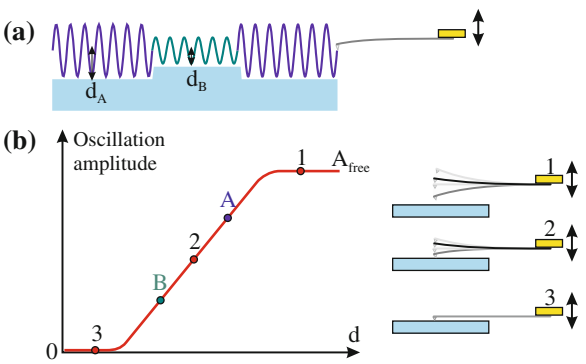

Fig. 15.2 **a** Schematic of the tip oscillation for two different average tip-sample distances d_A and d_B. The oscillation remains sinusoidal also at reduced distances d. **b** The vibration amplitude (being the free amplitude A_{free} for large tip-sample distances) decreases with decreasing tip-sample distance d, once the oscillation path reaches the repulsive range, i.e. $d \approx A_{\text{free}}$

We will discuss the energy flow into the oscillator supplied 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 harmonic 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 where the cantilever base (or the driving piezo) moves 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 via bending of the cantilever spring can be written as

$$\langle P_{\text{drive}} \rangle = \langle F \dot{z}_{\text{drive}} \rangle = \frac{1}{T} \int_0^T k \left[z_{\text{drive}}(t) - z(t) \right] \dot{z}_{\text{drive}}(t)\, dt. \tag{15.1}$$

Since all the functions in the integral are simple harmonic functions, the integral can be solved analytically, resulting in

$$\langle P_{\text{drive}} \rangle = -\frac{1}{2} k A_{\text{drive}} A \omega \sin\phi. \tag{15.2}$$

This expression is valid very generally, it is not necessary to assume that the driving frequency is close to the resonance frequency. It can be seen that the maximum power is delivered if the phase is $-90°$, corresponding to the resonance condition of the free cantilever. 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 $E_{\text{osc}} \approx 1/2\, kA^2$, and the energy supplied by the driving and then dissipated during one cycle is $E_{\text{drive}} = \langle P_{\text{drive}} \rangle T$, Q_{cant} can be written as

$$Q_{\text{cant}} = 2\pi \frac{E_{\text{osc}}}{E_{\text{drive}}} = \frac{-A}{A_{\text{drive}} \sin\phi}. \tag{15.3}$$

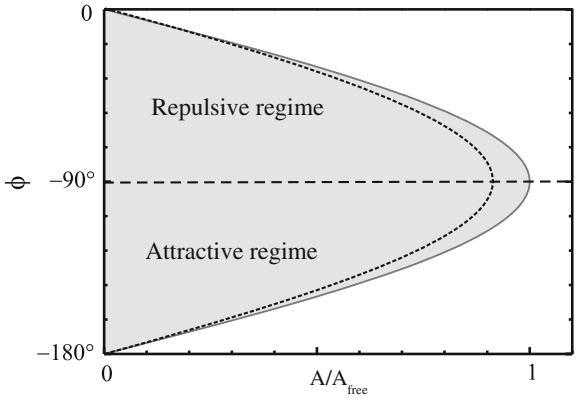

Fig. 15.3 Dependence of the phase on the amplitude according to (15.4). This expression is obtained from energy conservation and the assumption of a sinusoidal oscillation. For a given amplitude A/A_{free}, the phase can have two different values. $\phi < -90°$ corresponds to a net attractive tip-sample 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

If we identify the oscillation amplitude of the free cantilever (without any tip-sample interaction) as $A_{\text{free}} = Q_{\text{cant}} A_{\text{drive}}$, the oscillation amplitude can be written as

$$A = A_{\text{drive}} Q_{\text{cant}} \sin(-\phi) = A_{\text{free}} \sin(-\phi). \tag{15.4}$$

This shows that the amplitude decreases as the phase deviates from the resonance case $-90°$ of the free oscillation 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 (15.4) is shown in Fig. 15.3. It can be seen that a certain amplitude can be realized at two different phases, lower and higher than the phase of the free cantilever at resonance, i.e. $-90°$. This result, obtained under very general conditions, 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 (15.4). However, we can derive an expression $\phi(A/A_{\text{free}})$ from (2.25) and (2.28) by eliminating the dependence on ω, and (15.4) results.[1]

[1] The phase $\phi(A/A_{\text{free}})$ can be obtained numerically from (2.25) and (2.28). If this result is plotted in Fig. 15.3 it is indistinguishable on top of the curve obtained from (15.4). Alternatively (2.25) and (2.28) can be rearranged analytically leading to (15.4) in a very good approximation.

From Fig. 15.3 we see that an oscillation with a certain amplitude A can occur via two different phases. 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 (2.17) and include the static deflection introduced in Fig. 14.1, 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

$$m\ddot{z} = -\frac{m\omega_0}{Q_{\mathrm{cant}}}\dot{z} - k(z - (z_{\mathrm{drive}} + \Delta L)) + F_{\mathrm{ts}}(d + z). \tag{15.5}$$

For simplicity, we consider the driving frequency at the resonance frequency of the free cantilever, $\omega_{\mathrm{drive}} = \omega_0$, as it is often chosen in the tapping mode. Thus $z_{\mathrm{drive}} = A_{\mathrm{drive}} \cos \omega_0 t$. The resulting cantilever oscillation z and its time derivatives can be written as

$$z = A \cos(\omega_0 t + \phi) \ , \tag{15.6}$$

$$\dot{z} = -\omega_0 A \sin(\omega_0 t + \phi) \ , \tag{15.7}$$

$$\ddot{z} = -\omega_0^2 A \cos(\omega_0 t + \phi) = -\omega_0^2 z. \tag{15.8}$$

If we insert this into (15.5), the following equation results

$$-m\omega_0^2 z = \frac{m\omega_0^2 A}{Q_{\mathrm{cant}}} \sin(\omega_0 t + \phi) - k(z - \Delta L) + F_{\mathrm{ts}}(d + z) + kA_{\mathrm{drive}} \cos(\omega_0 t) \, . \tag{15.9}$$

Since $m\omega_0^2 = k$, the term on the left side of (15.9) cancels out the term $-kz$ on the right side. Now we multiply (15.9) by $A \cos(\omega_0 t + \phi)$ and integrate over one period. The integrals can be solved, or it can be seen form the symmetry that the first and the second term on the right side are zero after multiplication and integration. Thus the remaining equation reads as

$$A \int_0^T F_{\mathrm{ts}}(d + z) \cos(\omega_0 t + \phi) \, dt = -kAA_{\mathrm{drive}} \int_0^T \cos(\omega_0 t) \cos(\omega_0 t + \phi) \, dt. \tag{15.10}$$

The integral on the right side results as $1/2T \cos\phi$. Thus (15.10) can be written as

$$\frac{1}{T} \int_0^T F_{\mathrm{ts}}(d + z) A \cos(\omega_0 t + \phi) \, dt \equiv \langle F_{\mathrm{ts}} \cdot z \rangle = -\frac{1}{2} kAA_{\mathrm{drive}} \cos\phi. \tag{15.11}$$

If we finally use $A_{\text{free}} = Q_{\text{cant}} A_{\text{drive}}$, the cosine of the phase results as[2]

$$\cos\phi = \frac{-2Q_{\text{cant}}}{kAA_{\text{free}}} \langle F_{\text{ts}} \cdot z \rangle \,. \tag{15.12}$$

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. 15.1. Thus an attractive (negative) force will lead to a positive $\langle F_{\text{ts}} \cdot z \rangle$ and finally via (15.12) to a phase $\phi < -90°$. Correspondingly a repulsive force $F_{\text{ts}} > 0$ leads to a phase $\phi > -90°$. If $\langle F_{\text{ts}} \cdot z \rangle = 0$ the resonance phase of the free cantilever $\phi = -90°$ is restored.

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 corresponds to $\langle F_{\text{ts}} \cdot z \rangle$ being larger or smaller than zero, 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. 15.3. In this way, the measurement of the phase provides an unambiguous distinction between net attractive and net repulsive interactions.

15.2 Resonance Curve for an Anharmonic Force-Distance Dependence

The results in the previous section were obtained using very general considerations, either energy considerations, or averages over the equation of motion. Alternatively, the equation of motion for an anharmonic oscillator can be solved. This can be attempted either analytically [30], or by evaluating the solution of the equation of motion numerically for a particular model of the tip-sample force [31]. 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. 15.4, to odd shapes, for instance that shown in color in Fig. 15.4a.

For an anharmonic oscillator the resonance frequency of the oscillator changes with the amplitude, while for a harmonic oscillator the resonance frequency is independent of the oscillation amplitude. Thus 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

[2] If we approximate the tip-sample force by $F_{\text{ts}} = k'z$ (harmonic oscillator), $\langle F_{\text{ts}} \cdot z \rangle = -1/2\, k' A^2$ results (cf. (17.10)). Inserting this into (15.11) and remembering that according to (15.4) $A/A_{\text{free}} = -\sin\phi$, the following expression for the phase is obtained $\tan\phi = k/(k'Q_{\text{cant}})$, which corresponds to expression (14.14) obtained for the harmonic oscillator.

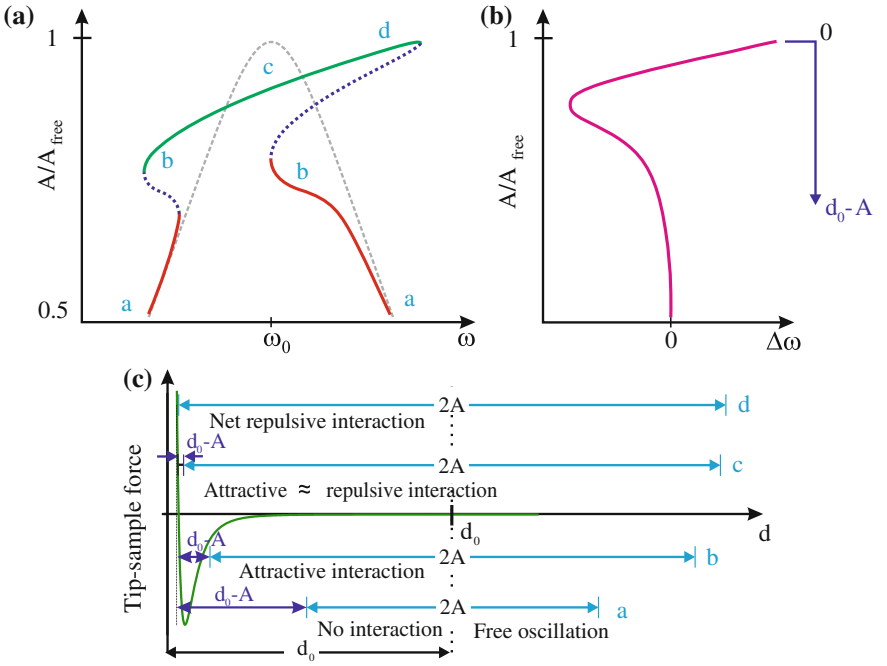

Fig. 15.4 **a** Resonance curve of an anharmonic oscillator (*solid line*) for a fixed average tip-sample distance $d = d_0$, compared to the free oscillation (*dashed gray curve*). For an anharmonic interaction the resonance curve becomes multivalued. The low-amplitude branch is shown in *red* and the high-amplitude branch in *green*. **b** "local" shift of the resonance frequency as function of the oscillation amplitude or alternatively as function of the distance from the sample surface to the lower turnaround point of the oscillation. **c** The oscillation ranges corresponding to the regions $a - d$ of the resonance curve (**a**) are indicated in a plot of the force-distance curve

of the resonance curve for an anharmonic oscillator as shown in Fig. 15.4a. 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 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.

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. 15.4c. Therefore, the resonance curve is very close to the resonance curve of the free cantilever (region a in Fig. 15.4a, 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. 15.4c. This results effectively in a local downshift of the resonance frequency explaining the "ear" seen to the left in region b in Fig. 15.4a.[3] For higher frequencies larger amplitudes occur and result 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. In region c in Fig. 15.4a the attractive contribution and the repulsive contribution compensate each other. At even higher frequencies, at the lower turnaround point the tip comes even closer to the surface, which effectively results in a shift of the resonance frequency towards higher frequencies, leading to the "ear" seen to the right in region d in Fig. 15.4a.

According to Fig. 15.4c, an increasing oscillation amplitude corresponds to an decreasing distance between the sample surface and the lower turnaround point $d_0 - A$. The amplitude dependent frequency shift of the resonance frequency of an anharmonic oscillator as a function of the amplitude $\Delta\omega(A)$ is qualitatively shown in Fig. 15.4b.

As seen from Fig. 15.4a, for certain ranges of frequencies the resonance curve of an anharmonic oscillator becomes multivalued. The solutions shown as blue dotted lines are unstable [30], while the low-amplitude branch (green in Fig. 15.4a) and the high-amplitude branch (red) correspond to two stable solutions of the equation of motion. 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 ω, in tapping mode atomic force microscopy the driving frequency is kept constant and we will discuss this case in the following.

15.3 Amplitude Instabilities for an Anharmonic Oscillator

In Fig. 15.5a 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. 15.2b. 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. 15.5a. The tip-sample approach is shown in red while the retraction is shown in green.

The jumps shown in Fig. 15.5a can be explained considering the resonance curves shown in Fig. 15.6a–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

[3] Correspondingly, the left "ear" also occurs on the high-frequency side of the resonance curve.

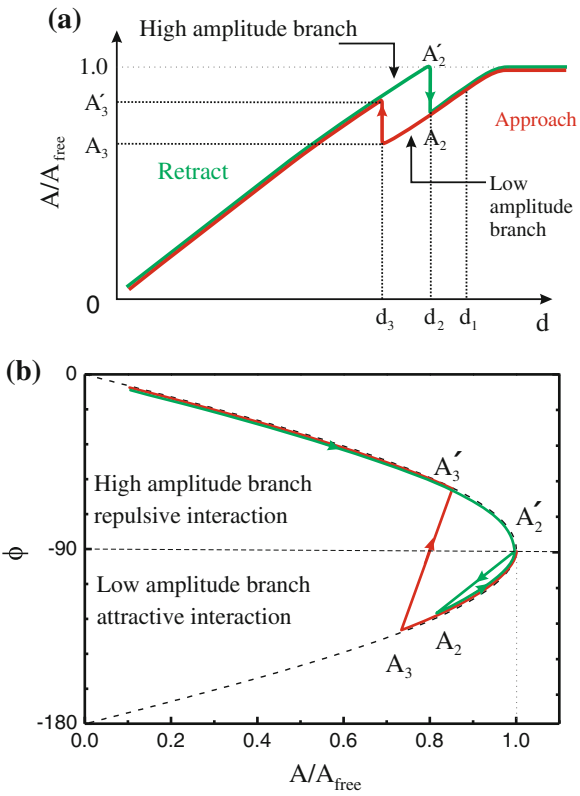

Fig. 15.5 **a** Amplitude distance curves with jumps between the low-amplitude branch and the high-amplitude branch shown for approach (*red*) and retraction (*green*). **b** Phase as function of the oscillation amplitude during approach (*red*) and retraction (*green*). Phase values below the −90° line correspond to an attractive interaction (low-amplitude branch), while phase values above the −90° line correspond to a net repulsive interaction (high-amplitude branch)

distance of d_1, the "ear" on the low-frequency side of the resonance curve visible in Fig. 15.6a 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 frequency is already reduced relative to the free amplitude (formation of the "ear" in Fig. 15.6a).

For smaller tip-sample distances d_3, an "ear" develops on the high-frequency side (Fig. 15.6b) for large amplitudes due to the repulsive tip-sample interaction. Due to this "ear" a low-amplitude branch and a high-amplitude branch develop. In Fig. 15.6b the situation is shown in which the low-amplitude branch of oscillation disappears at ω_0. The dotted line in Fig. 15.6b 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. 15.6b and 15.5a. 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 it approaches closer to the surface, i.e. smaller d (Fig. 15.5a).

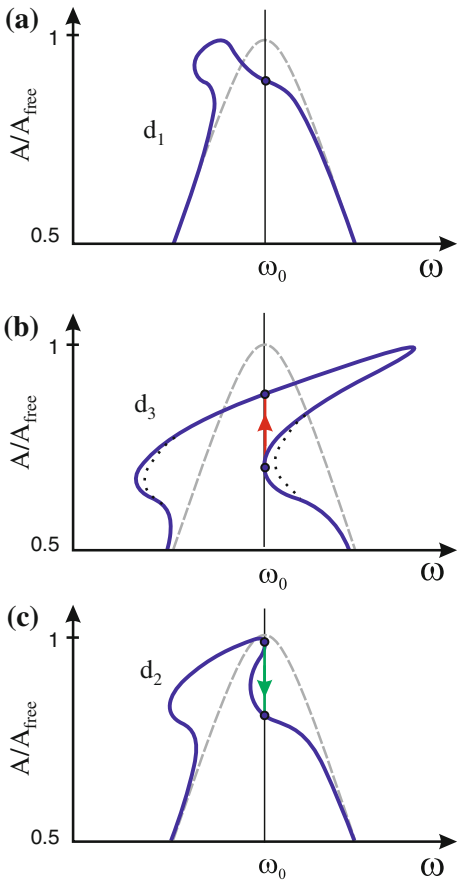

Fig. 15.6 Resonance curves for different average tip-sample distances d. The driving frequency is considered to be at the resonance of the free cantilever ω_0. **a** For large tip-sample distances (around d_1), at the lower turnaround point the tip only reaches the attractive regime, leading to an "ear" 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 distances 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 [31]

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. 15.6c and 15.5a). 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.

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 (15.12), $\phi < -90°$ corresponds to a net attractive interaction, while $\phi > -90°$ corresponds to a net repulsive interaction. In Fig. 15.5b, the double-valued dependence of the phase on the amplitude according to (15.4) 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 (15.12) (red line in Fig. 15.6b). 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 $-90°$ and the phase approaches zero for smaller tip-sample distances. During the retraction (increasing d), the green line is followed.[4] Thus according to (15.12) the high-amplitude branch with $\phi > -90°$ corresponds to a net repulsive interaction.

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 (repulsive interaction) can be selected for subsequent imaging. Depending on the material imaged, different interaction regimes may be desired. For a soft 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. When scanning from material A 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 on material A to the low-amplitude state on material B. The smaller oscillation amplitude leads to a reduction in the average tip-sample distance d by about $\sim$1 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 $\omega_{\text{drive}} = \omega_0$. 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 [31]. If in tapping mode the

[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.

[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 and thus change the force-distance behavior locally, resulting in a switch to another branch of the oscillation state.

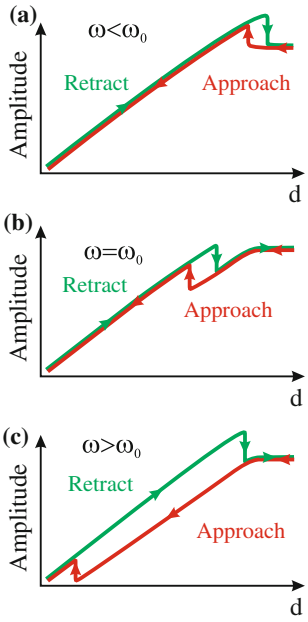

Fig. 15.7 Amplitude as a function of the average tip-sample distance d for driving frequencies **a** below ω_0, **b** at ω_0, and **c** above ω_0. The curves are shown for approach (*red*) and retraction (*green*). As an exercise, the dependences in (**a**)–(**c**) can be deduced from Fig. 15.6a–c This figure is adapted from [31]

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 as shown in Fig. 15.7a. 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. 15.7c. 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.

15.4 Energy Dissipation in Dynamic 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 (14.11). For the case of large amplitudes used in the tapping mode, the strength of

the dissipative 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. 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 and use the principle of energy conservation to include also the dissipative tip-sample interaction in the energy balance. In the steady-state, the power (average over one period) injected to the cantilever system by driving the cantilever according to (15.1) is equal to the power dissipated by the cantilever damping in the surrounding fluid plus the power dissipated due to the tip-sample interaction, as

$$\langle P_{\text{drive}} \rangle = \langle P_{\text{cant}} \rangle + \langle P_{\text{ts}} \rangle \,. \tag{15.13}$$

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 [28]. For simplicity, we avoid assigning a sign to the power and consider all (averaged) powers in (15.13) as positive.

The power pumped into the system by external driving of the cantilever was calculated in (15.2) as[6]

$$\langle P_{\text{drive}} \rangle = -\frac{1}{2} k A_{\text{drive}} A \omega \sin \phi. \tag{15.14}$$

The cantilever damping by the fluid is assumed to be proportional to $\dot{z}$, as $F_{\text{cant}}^{\text{damp}} = -\frac{m\omega_0}{Q_{\text{cant}}}\dot{z}$. Along the same lines as in (15.1), the power dissipated in the cantilever can be calculated as

$$\langle P_{\text{cant}} \rangle = \left\langle \frac{m\omega_0}{Q_{\text{cant}}} \dot{z}^2 \right\rangle = \frac{1}{T} \int_0^T \frac{m\omega_0}{Q_{\text{cant}}} A^2 \omega^2 \sin^2(\omega t + \phi) = \frac{k A^2 \omega^2}{2 Q_{\text{cant}} \omega_0}. \tag{15.15}$$

Due to (15.13), the power dissipated in the tip-sample interaction can be written as

$$\langle P_{\text{ts}} \rangle = \langle P_{\text{drive}} \rangle - \langle P_{\text{cant}} \rangle = \frac{k A^2 \omega}{2 Q_{\text{cant}}} \left(\frac{Q_{\text{cant}} A_{\text{drive}} \sin(-\phi)}{A} - \frac{\omega}{\omega_0} \right). \tag{15.16}$$

This result was obtained using the general law of energy (or power) conservation without any assumptions about the nature of the tip-sample interaction. The tip-sample

[6] Since $\phi < 0$, $\langle P_{\text{drive}} \rangle$ is positive.

interaction enters on the right hand side of (15.16) via the experimentally measured amplitude A. The frequencies relevant for the power of the oscillator are the actual oscillation frequency ω and the oscillation frequency of the free cantilever ω_0.

If the driving frequency ω is chosen at the resonance frequency of the free cantilever ω_0, (15.16) can be written as[7]

$$\langle P_{\mathrm{ts}} \rangle = \frac{k A^2 \omega_0}{2 Q_{\mathrm{cant}}} \left(\frac{A_{\mathrm{free}}}{A} \sin(-\phi) - 1 \right) , \tag{15.17}$$

with $A_{\mathrm{free}} = Q_{\mathrm{cant}} A_{\mathrm{drive}}$. Correspondingly, the dissipated energy per oscillation period T results as

$$\langle E_{\mathrm{ts}} \rangle = \frac{2\pi E_{\mathrm{osc}}}{Q_{\mathrm{cant}}} \left(\frac{A_{\mathrm{free}}}{A} \sin(-\phi) - 1 \right) , \tag{15.18}$$

with $E_{\mathrm{osc}} = 1/2\, k A^2$ being the energy contained in the cantilever oscillation. The last term in the bracket in (15.18) is proportional to the power dissipated by the cantilever damping, while the first term in (15.18) is proportional to the total dissipated power.

In the case that no dissipative interactions are present ($\langle E_{\mathrm{ts}} \rangle = 0$), a simple relation for the phase already obtained in (15.4) results as

$$\phi = -\arcsin\left(\frac{A}{A_{\mathrm{free}}} \right) . \tag{15.19}$$

We can rearrange (15.18) 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

$$\sin(-\phi) = \frac{A}{A_{\mathrm{free}}} \left(\frac{\langle E_{\mathrm{ts}} \rangle}{\langle E_{\mathrm{cant}} \rangle} + 1 \right) . \tag{15.20}$$

The second term in (15.20) is the contribution due to the elastic 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 (15.19)) if no dissipative tip-sample interaction is present. A phase change can therefore be exclusively attributed 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 elastic 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 (15.20), 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.

[7] While we used here the principle of energy conservation to derive (15.17), this equation can be obtained alternatively by multiplying (15.9) with $\omega_0 A \sin(\omega t + \phi)$ and integrating over one period.

For this case the $\phi(A/A_{\text{free}})$ dependence from (15.20) is displayed in Fig. 15.3 as a dashed curve for $\langle E_{\text{ts}} \rangle \,/\, \langle E_{\text{cant}} \rangle = 0.1$.

Finally, we give a quantitative example of the power dissipated into the tip-sample interaction. All variables in (15.17) are either known or can be measured. In a tapping mode experiment on a silicon wafer in air, a power dissipation of 0.3 pW was obtained independent of the oscillation amplitude [28].

15.5 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. The strongest force contribution in the tapping mode is the repulsive force at the lower turnaround point of one oscillation cycle. The behavior with respect to the contamination layer is an advantage of the tapping mode compared to the non-contact mode, where 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) 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 while it avoids their disadvantages. While the intermittent contact mode has several advantages when imaging a surface, 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 amplitude.

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 tip is measured by the detector and input to 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 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 $A/A_{\text{free}} \approx 0.9$.

As we have already seen, the amplitude has a monotonous dependence on the tip-sample distance (Fig. 15.2b). This leads advantageously to a clear unambiguous feedback signal. 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. 17.3.

15.6 Summary

- The intermittent contact mode (tapping mode) is a detection mode, which is different from the AM-slope detection considered in the previous chapter: (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. 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 a tip-sample interaction.
- An anharmonic tip-sample force leads to the coexistence of two vibrational modes with a low-amplitude and a high-amplitude, corresponding to a net attractive and 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 mode and $\phi > -90°$ for the high-amplitude 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.

Chapter 16
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 frequency shift in the dynamic AM mode, or constant amplitude in the tapping mode are measured. In Chap. 13 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 as a function of the tip-sample distance. Invoking force-distance curves 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, force volume or pulsed force mode. Besides access to the mechanical properties, this mode also allows high-resolution imaging, it is a tapping mode under force control.

16.1 Principles of Force-Distance Curve Mapping

In 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.

The force-distance curves considered in Chap. 13 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 acquired in less than a millisecond. In order to prevent

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_16

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. 16.1a. 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. 16.1b 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 to B, the attractive force increases. At point B snap-to-contact occurs. The repulsive force increases towards point C. The 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.

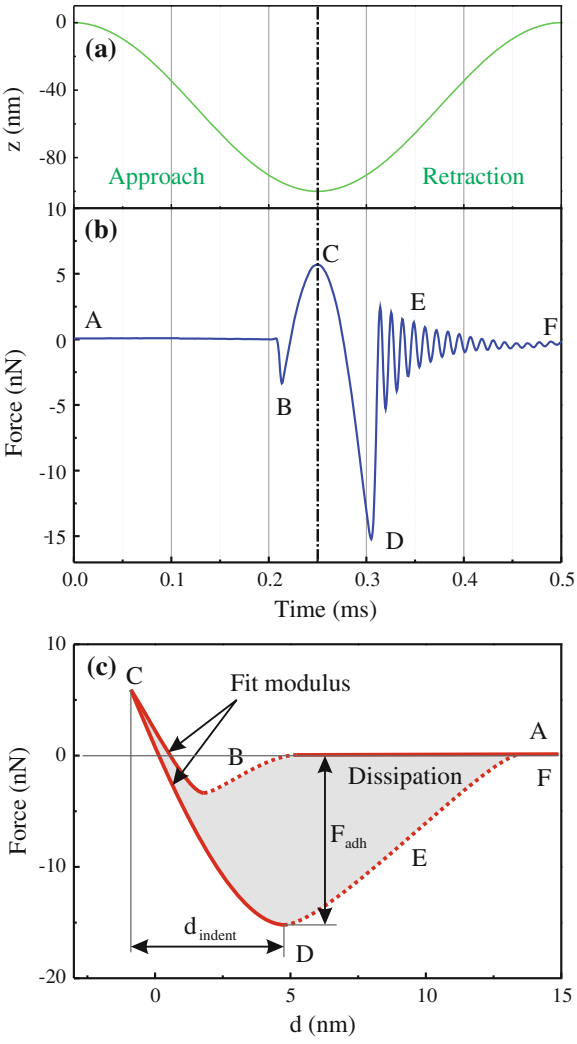

Fig. 16.1 **a** Sinusoidal change of the z-position of the tip or sample position during the acquisition of the force-distance curve. **b** Tip-sample force as a function of time for approach and retraction. **c** Tip-sample force as function of the tip-sample distance. From this curve, quantities like 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. Young's modulus can be determined by fitting a model for the mechanic contact to the approach force-distance curve

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. 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. 16.1c (cf. Fig. 13.5). 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. 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.

16.2 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. 16.1c. These quantities can be represented as images of (maximum) adhesion or indentation depth at the peak force used as the setpoint in the feedback.

The dissipation energy can be obtained as the area between the approach and the retraction curves, as

$$E_{\mathrm{diss}} = \int_{z_{\mathrm{min}}}^{z_{\mathrm{max}}} \left(F_{\mathrm{approach}} - F_{\mathrm{retract}}\right) dz, \tag{16.1}$$

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. 16.1c. 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 measured in the repulsive regime of the tip-sample contact can be fitted to an appropriate model of the tip-sample contact, for instance the Hertz model of the elastic contact, or other models also including inelastic contributions. 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 ratio of 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 distance above the sample 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.

16.3 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.

Chapter 17
Frequency Modulation (FM) Mode in Dynamic Atomic Force Microscopy—Non-contact Atomic Force Microscopy

In Chap. 15 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 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 scanning 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 obtained as a function of the distance.

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. 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 so large that the tip-sample force cannot be approximated as linear. 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. The time response in FM detection is not limited for high quality factors, as it is the case in AM detection. Therefore, the FM detection scheme can be used for cantilevers with high quality factors, i.e. in vacuum.

17.1 Principles of Dynamic Atomic Force Microscopy II

In Chap. 14, we derived the frequency shift in the limit of small oscillation amplitudes, i.e. the force was described as linear with the tip-sample distance in the range of the oscillation amplitude. In this limit, the frequency shift is proportional to the force gradient. However, for most cases of larger oscillation amplitudes or short-range

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_17

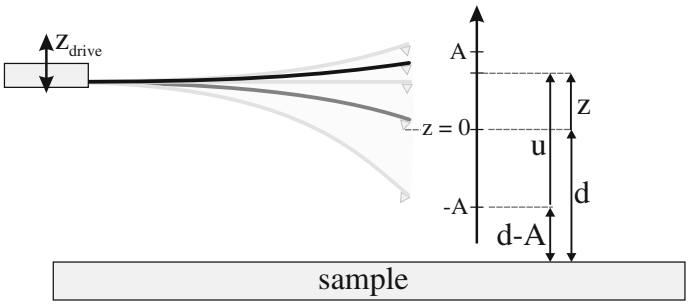

Fig. 17.1 Scheme of the cantilever vibration illustrating the corresponding coordinates

forces this limit does not hold at all. The interaction between tip and sample changes strongly on the scale of the vibrational amplitude of the cantilever.

In the following, we will consider a driven damped harmonic oscillator under the influence of a non-linear tip-sample force $F_{\mathrm{ts}}(d+z)$. The driving force is given by an external sinusoidal oscillation $z_{\mathrm{drive}} = A_{\mathrm{drive}} \cos(\omega t)$ of the cantilever base. This driving oscillation corresponds to a force $F_{\mathrm{drive}} = k z_{\mathrm{drive}}$. 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. $\omega = \omega_{\mathrm{drive}} = \omega_0'$. The equation of motion for the driven damped harmonic oscillator with an external tip-sample force $F_{\mathrm{ts}}(d+z)$ added is written according to (2.17) as

$$m\ddot{z} + \frac{m\omega_0}{Q_{\mathrm{cant}}}\dot{z} + k(z - z_{\mathrm{drive}} - \Delta L) = F_{\mathrm{ts}}(d+z). \tag{17.1}$$

The relevant coordinates are indicated in Fig. 17.1. The zero point for z ($z = 0$) is given by the condition that the tip-sample force is compensated by the static cantilever bending ΔL, cf. Fig. 14.1 and (14.3). In this case the tip-sample distance is d.

In spite of the fact that a non-linear tip-sample force is included into the equation of motion, we approximate the solution $z(t)$ by a harmonic oscillation $z(t) = A\cos(\omega t + \phi)$. 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 (17.1), however, we will calculate the shift of the resonance frequency. The relation between tip-sample force and frequency shift Δf is more complicated than the simple proportional relation between Δf and the force gradient obtained in the small amplitude limit (14.7). 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

[1] Under the influence of the tip-sample force the resonance frequency of the free cantilever, ω_0, shifts to ω_0'.

well as the expression for ω_{drive} into (17.1). Subsequently we multiply (17.1) by $z(t) = A\sin\omega t$ and integrate over one period resulting in the following expression

$$-\int_0^T m\omega^2 A^2 \sin^2\omega t\, dt + \int_0^T \frac{m\omega_0}{Q_{\text{cant}}} A^2\omega\cos\omega t\sin\omega t\, dt + \int_0^T kA^2\sin^2\omega t\, dt$$
$$-\int_0^T kA_{\text{drive}} A\cos\omega t\sin\omega t\, dt - \int_0^T k\Delta L A\sin\omega t\, dt$$
$$= \int_0^T F_{\text{ts}}(d + z(t))A\sin\omega t\, dt. \tag{17.2}$$

Since the integral of $\cos\omega t\sin\omega t$ over one period vanishes, the second and fourth terms on the left side in (17.2) 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 (17.2) can be written as

$$(k - m\omega^2)A^2\int_0^T \sin^2\omega t\, dt = \int_0^T F_{\text{ts}}(d + z(t))A\sin\omega t\, dt. \tag{17.3}$$

The integral $\int\sin^2\omega t dt$ within the limits from 0 to T can be calculated as $\frac{1}{2}T = \frac{\pi}{\omega}$, which results in

$$(k - m\omega^2)A^2\frac{\pi}{\omega} = \int_0^T F_{\text{ts}}(d + A\sin\omega t)A\sin\omega t dt. \tag{17.4}$$

The left hand side of (17.4) can be further evaluated as follows

$$\frac{A^2\pi}{\omega}\left(k - m\omega^2\right) = \frac{A^2 m\pi}{\omega}\left(\frac{k}{m} - \omega^2\right)$$
$$= \frac{A^2 m\pi}{\omega}\left(\omega_0^2 - \omega^2\right) = \frac{A^2 m\pi}{\omega}\left(\omega_0 + \omega\right)\left(\omega_0 - \omega\right). \tag{17.5}$$

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 (17.1) can be further written as

$$2\pi m A^2(\omega_0 - \omega) = -4\pi^2 m A^2(f - f_0) = -4\pi^2 m A^2\Delta f. \tag{17.6}$$

Now also taking the right-hand side of (17.4) into account the following expression for the frequency shift arises

$$\Delta f = -\frac{1}{4\pi^2 m A^2} \int_0^T F_{\mathrm{ts}}(d + A\sin\omega t) A \sin\omega t \; dt. \tag{17.7}$$

The time average of $F_{\mathrm{ts}}(t)$ times $z(t)$ over one period can be written as

$$\langle F_{\mathrm{ts}}(t) \cdot z(t) \rangle \equiv \frac{1}{T} \int_0^T F_{\mathrm{ts}}(d + A\sin\omega t) A \sin\omega t \; dt. \tag{17.8}$$

Using the above equation, (17.7) can be rewritten as the following expression for Δf (using $T = 1/f_0$ and $m = k/\omega_0^2$)

$$\Delta f = -\frac{f_0}{A^2 k} \langle F_{\mathrm{ts}}(t) \cdot z(t) \rangle. \tag{17.9}$$

The frequency shift is proportional to $\langle F \cdot z \rangle$, which is the time average of force times distance (tip-sample 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 (14.8). In contrast to the case of small amplitudes, the frequency shift depends as $1/A^2$ on the oscillation amplitude.

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

$$\langle F_{\mathrm{ts}} \cdot z \rangle = -\langle k' \cdot z^2 \rangle = \frac{1}{T} \int_0^T -k' A^2 \cos^2 \omega t dt = -\frac{1}{2} k' A^2, \tag{17.10}$$

which recovers the result of the frequency change found for the small amplitude limit $\Delta f = f_0 k'/(2k)$ (cf. 14.8). In analogy to this result for the small amplitude limit an effective tip-sample spring constant can generally be defined as

$$k' \equiv -\frac{2\langle F_{\mathrm{ts}} \cdot z \rangle}{A^2}, \tag{17.11}$$

in order to recover an equation of the same form as in the small amplitude limit $\Delta f = f_0 k'/(2k)$.

17.1.1 Expression for the Frequency Shift

When analyzing the time average in (17.10) 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 (17.10) 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 (17.10)

$$\langle F_{\mathrm{ts}}(d+z) \cdot z \rangle = \frac{1}{T} \int_0^T F_{\mathrm{ts}}(d+z(t)) \cdot z(t) dt, \tag{17.12}$$

with $z(t) = A \sin \omega t$. In order to convert the time average to a spatial average over the trajectory, we substitute in (17.12) the variable t by z as

$$\frac{dz}{dt} = A\omega \cos(\omega t) = A\omega \sqrt{1 - \sin^2(\omega t)} = \omega \sqrt{A^2 - z^2}. \tag{17.13}$$

Therefore, the average $\langle F \cdot z \rangle$ can be written as

$$\begin{aligned} \langle F_{\mathrm{ts}}(d+z) \cdot z \rangle &= \frac{1}{T} \int_0^T F_{\mathrm{ts}}(d+z(t)) \cdot z(t) dt \\ &= \frac{2}{\omega T} \int_{-A}^{+A} \frac{F_{\mathrm{ts}}(d+z) \cdot z}{\sqrt{A^2 - z^2}} dz \\ &= \frac{1}{\pi} \int_{-A}^{+A} \frac{F_{\mathrm{ts}}(d+z) \cdot z}{\sqrt{A^2 - z^2}} dz. \end{aligned} \tag{17.14}$$

Combining (17.9) and (17.14) the following expression for the frequency shift is obtained

$$\Delta f = -\frac{f_0}{\pi k A^2} \int_{-A}^{+A} F_{\mathrm{ts}}(d+z) \frac{z}{\sqrt{A^2 - z^2}} dz = -\frac{f_0}{\pi k A^2} \int_{-A}^{+A} F_{\mathrm{ts}}(d+z) g(z) dz. \tag{17.15}$$

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. 17.2a. 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 (17.15), the tip-sample force F_{ts}, must also be considered.

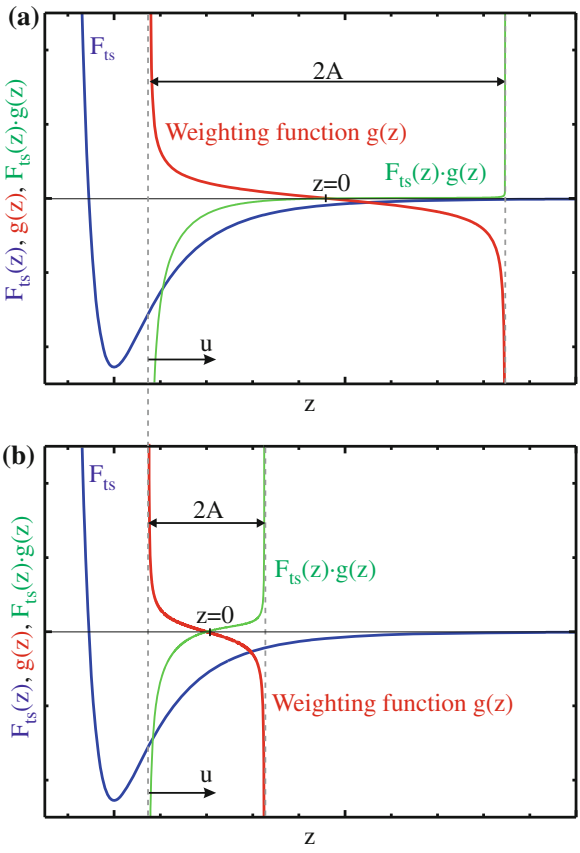

Fig. 17.2 The tip-sample force, the weighting function $g(z)$, and their product are displayed as a function of distance z for two different oscillation amplitudes A. In the large amplitude limit **a** the frequency shift signal is mainly picked up close to the lower turnaround point of the oscillation, 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**)

For the situation of a large amplitude shown in Fig. 17.2a 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 (17.15) is shown as a green line in Fig. 17.2a. In total, for large amplitudes the contributions to the frequency shift come only from regions close to the lower turnaround point.

The case of a smaller oscillation amplitude is shown in Fig. 17.2b. For better comparability, the lower turnaround point of the oscillation was placed in the same position as in Fig. 17.2a. In this case, the integrand of (17.15) 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. 17.2b.

Comparing the large amplitude case to the small amplitude case (Fig. 17.2a, b) we see that for the large amplitude case only the region close to the lower turnaround point contributes to the frequency shift, while the major part of the oscillation path does not result in a contribution to the frequency shift. In contrast, for small amplitudes

contributions to the frequency shift arise from all parts of the oscillation cycle. This means that for smaller oscillation amplitudes a stronger frequency shift signal is expected. In addition to this contribution from the integral in (17.15) also the prefactor $1/A^2$ enhances the frequency shift for small amplitudes. If we compare this amplitude dependence of the frequency shift, we note that in the previously treated small amplitude limit (14.8) 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. 18. Together, the important figure of merit, the signal-to-noise ratio, will be obtained.

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 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 (17.15) the total frequency shift can be split into contributions arising from the individual forces.

In this chapter, we have considered up to now conservative tip-sample interactions. In this case, the force is the same for a certain tip-sample distance independent of the direction of motion either for the approach towards the sample or for the retraction from the sample. For a dissipative tip sample interaction the forces at a certain point can be different for approach and retraction and this has to be considered. In this case, the tip-sample force in (17.15) can be replaced by $F_{\mathrm{ts}} = (F_{\mathrm{ts,approach}} + F_{\mathrm{ts,retraction}})/2$ [31, 32].

17.1.2 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. 17.1). 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. In some cases, 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. 17.1). If we substitute $z = u - A$ and express the tip-sample distance as $d + z = d - A + u$ the frequency shift (17.15) results in

$$\Delta f = -\frac{f_0}{\pi k A^2} \int_0^{2A} \frac{F_{\mathrm{ts}}(d - A + u)(u - A)}{\sqrt{A^2 - (u - A)^2}} du$$

$$= -\frac{f_0}{\pi k A^2} \int_0^{2A} \frac{F_{\mathrm{ts}}(d - A + u)(u - A)}{\sqrt{(2A - u)u}} du. \tag{17.16}$$

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 (17.15) or (17.16) has appreciable values only at tip positions very close to the lower turnaround point, as also indicated by the green line in Fig. 17.2a. 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

$$\Delta f = \frac{f_0}{\pi k A^2} \int_0^{\infty} \frac{F_{\mathrm{ts}}(d - A + u)A}{\sqrt{2Au}} du = \frac{f_0}{\sqrt{2}\pi k A^{3/2}} \int_0^{\infty} \frac{F_{\mathrm{ts}}(d - A + u)}{\sqrt{u}} du \ . \tag{17.17}$$

The dependences on resonance frequency and spring constant are the same as for the small amplitude limit (14.8). Furthermore, the frequency shift is proportional to $A^{-3/2}$.

The expression for the frequency shift in (17.17) contains two contributions. The frequency shift depends on the tip-sample force and also on the cantilever and experimental parameters. In order to separate the parameters out, a *normalized frequency shift* γ can be defined as

$$\gamma = \Delta f \frac{k A^{3/2}}{f_0} \ . \tag{17.18}$$

The normalized frequency shift has the following significance: Multiplying the experimentally measured frequency shift Δf by the factor $k A^{3/2}/f_0$, the expression (17.17) can be written as

$$\gamma = \frac{1}{\sqrt{2}\pi} \int_0^{\infty} \frac{F_{\mathrm{ts}}(d - A + u)}{\sqrt{u}} du \ . \tag{17.19}$$

The normalized frequency depends only on an integral over the tip-sample force, while the dependence on the experimental 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. 17.3a measurements on a graphite sample are shown. The frequency shift is plotted as a function of tip-sample distance. Different frequency shift curves

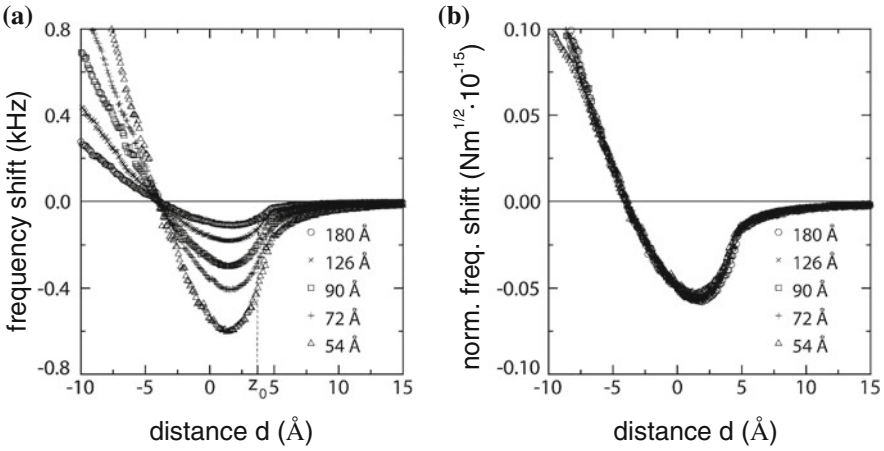

Fig. 17.3 **a** Experimentally measured frequency shift on a graphite sample as a function of the average tip-sample distance d for different values of the oscillation amplitude. The curves are shifted along the horizontal axis in order to make them comparable [33]. **b** If the normalized frequency 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 [33])

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. 17.3b 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 (17.19) as

$$\gamma = \frac{F_0}{\sqrt{2\pi}} \int_0^{\lambda} u^{-1/2} du = \frac{\sqrt{2}}{\pi} F_0 \sqrt{\lambda}. \tag{17.20}$$

To give some numbers: For $f_0 = 200\,\mathrm{kHz}$, $F_0 = 2\,\mathrm{nN}$, $A = 10\,\mathrm{nm}$, $k = 10\,\mathrm{N/m}$ and $\lambda = 0.1\,\mathrm{nm}$ a normalized frequency shift of $9\,\mathrm{fN}\sqrt{\mathrm{m}}$ results, corresponding to a frequency shift of $\Delta f = 180\,\mathrm{Hz}$. For an exponentially decaying force

$$F(z) = F_0 e^{-u/\lambda}, \tag{17.21}$$

the corresponding normalized frequency shift (17.19) can be calculated in the large amplitude limit as [34]

$$\gamma = \frac{1}{\sqrt{2\pi}} F_0 \sqrt{\lambda}, \tag{17.22}$$

which is (apart from a constant factor) the same result as obtained for a constant force F_0 with a range λ, shown in (17.20). 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 [34].

17.1.3 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 (17.15) this equation cannot easily be inverted analytically to a solution for $F_{\mathrm{ts}}(\Delta f)$. In the small amplitude limit the obtained equation

$$\Delta f(d) = -\frac{f_0}{2k} \frac{\partial F_{\mathrm{ts}}(d+z)}{\partial z}\bigg|_{z=0}, \tag{17.23}$$

can be inverted to

$$F_{\mathrm{ts}}(d) = \frac{2k}{f_0} \int_d^\infty \Delta f(z')dz'. \tag{17.24}$$

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, (17.15) can be inverted using approximations which allow the determination of the force with an accuracy of 5 % [32, 35].

17.2 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.

17.2.1 Self-excitation Mode

In the self-excitation mode the cantilever itself as a harmonic oscillator is the frequency-determining element in an oscillator circuit. A positive feedback is used

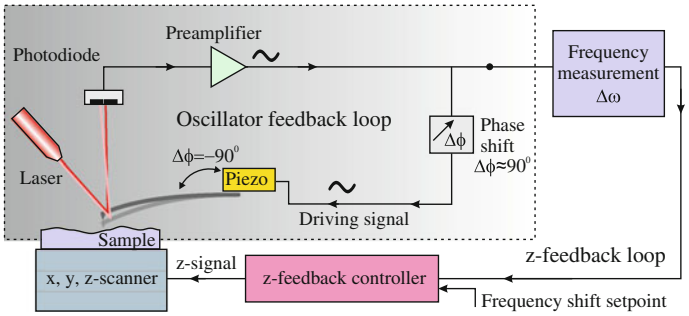

Fig. 17.4 Schematic of an FM detection setup operated in the self-excitation mode. In the circuit the measured cantilever oscillation signal is phase shifted and fed back to the actuator driving the cantilever. In addition to this (inner) oscillator feedback loop, the measurement of the shift of the resonance frequency $\Delta\omega$ is used in an outer z-feedback feedback loop in order to control the tip-sample distance

in order to self-excite the cantilever. A schematic of the implementation (Fig. 17.4) 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 $\Delta\omega$ is used in an outer z-feedback feedback loop in order to control the tip-sample distance.

In a mechanical harmonic oscillator oscillating at resonance there is a phase shift of $-90°$ between the displacement of the cantilever tip and the mechanical excitation, i.e. the cantilever oscillation is lagging the excitation. 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. 17.4). In order to excite the cantilever with the correct resonance phase, a phase shift of $+90°$ has to be applied to the oscillation signal before feeding it back as the driving signal. This phase shift "compensates" the $-90°$ phase shift between mechanical excitation and oscillation of the cantilever. For simplicity, we neglect all other phase shifts present in the loop, for instance in the preamplifier. 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.

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. 18). 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 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. 17.2.3.

Another question is: 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.[2] In the self-excitation mode, the cantilever is fed by its own oscillation. If the phase of the oscillator feedback loop is $-90°$, this means that the oscillator is automatically always fed at its resonance frequency. Due to this driving at resonance condition, the actual oscillation frequency will adapt to the new resonance frequency very fast.

This instantaneous adaption of the oscillation to the new resonance frequency can be demonstrated by including a term describing the self-oscillation loop in the equation of motion of the harmonic oscillator and subsequently solving this equation numerically. The self-excitation can be described in the equation of motion (2.17), replacing the driving term by the feedback term $\omega_0^2/Qz(t-t_0)$ [36]. The equation of motion for a harmonic oscillator with self-excitation then reads

$$\ddot{z} + \frac{\omega_0}{Q}\dot{z} + \omega_0^2 z = \frac{\omega_0^2}{Q} z(t-t_0). \tag{17.25}$$

The time shift $t - t_0$, with which the cantilever deflection signal is fed back as the driving signal $z(t - t_0)$, corresponds to a phase shift $\phi_0 = \omega t_0$, which is set to $-90°$. In order to demonstrate the tracking capability of the self-excitation mode, i.e. the fact that the actual cantilever oscillation frequency follows the change of the resonance frequency, the numerical solution of the equation of motion is analyzed. The response of the cantilever oscillation to an instantaneous change of the resonance frequency from ω_0 to ω_0' is simulated. Does the cantilever oscillation $z(t)$ follow the resonance frequency shift (tracking capability), and how rapidly is the new steady-state attained?

In Fig. 17.5 the deflection $z(t)$ obtained from the simulation is shown as a red line. The quick adaption of the oscillation to the new increased resonance frequency can be seen from the continuously increasing shift of the red curve relative to the reference curve (black line), corresponding to an oscillation without a change of the resonance frequency. In spite of the very large change of the resonance frequency of $\Delta\omega/\omega_0 = 5 \times 10^{-3}$, no transient occurs at $t = 0$. This is very different from

[2] For the case of AM detection, we have seen in Sect. 14.5 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.

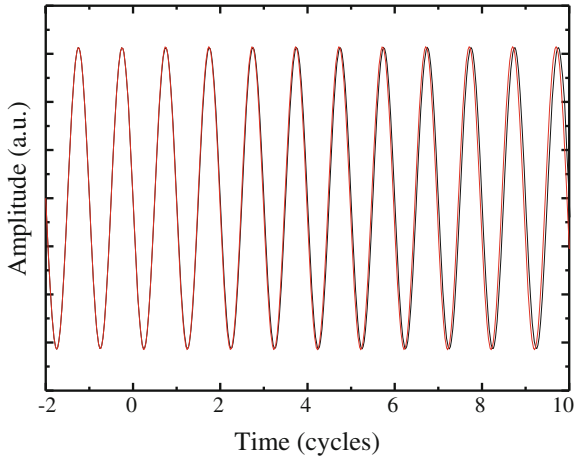

Fig. 17.5 Deflection $z(t)$ obtained from the simulation of a damped harmonic oscillator with self-excitation included in the equation of motion. At time $t = 0$ the resonance frequency of the harmonic oscillator changes from $f_0 = 1\,\text{MHz}$ by a large value of $\Delta f = 5\,\text{kHz}$. The simulated deflection (*red line*) is compared to the case without a change of the resonance frequency (*black line*). The fast adaption to the higher resonance frequency can be seen by a shift of the *red curve* with respect to the *black reference curve*. In spite of the large frequency shift assumed, the difference between the two *curves* is negligible at the time at which the resonance frequency changes ($t = 0$). This demonstrates the tracking capability of the self-oscillation mode with a very short time constant

AM detection, where a transient of about Q oscillation occurs before the oscillation has adapted to the new steady-state. The response of the cantilever oscillation to a change of the resonance frequency occurs instantaneously without a transient. After the change of the resonance frequency at $t = 0$ the amplitude remains constant. This is the case since the oscillation always remains in resonance. This is different from the AM case where the oscillation at ω_{drive} is off-resonance after a change of the resonance frequency. This leads to a reduced amplitude after a time constant of about Q oscillations in the AM mode, as seen in Fig. 14.8a.

The reason for the much shorter time constant in the self-excitation mode of the FM detection compared to the AM detection mode (cf. Sect. 14.5) can alternatively (to the analysis of the solution of the equation of motion) be rationalized by considering the change of the energy of the cantilever oscillation upon a change of the resonance frequency. 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 the following, we will compare the energy change during this transition for the AM and FM modes. The energy difference between the free oscillator and the state with

tip-sample interaction present are 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

$$\Delta E_{\mathrm{AM}} = E_{\mathrm{free}} - E_{\mathrm{ts}} = \frac{1}{2} m\omega_0^2 A^2 - \frac{1}{2} m\omega_0^2 (0.9A)^2 = 0.19 E_{\mathrm{free}}. \tag{17.26}$$

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

$$\begin{aligned} \Delta E_{\mathrm{FM}} &= E_{\mathrm{free}} - E_{\mathrm{ts}} = \frac{1}{2} m\omega_0^2 A^2 - \frac{1}{2} m\omega_0'^2 A^2 \\ &= \frac{1}{2} m\omega_0^2 A^2 \left(1 - \frac{\omega_0'^2}{\omega_0^2}\right) \approx E_{\mathrm{free}} \frac{2\Delta\omega}{\omega_0}. \end{aligned} \tag{17.27}$$

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 (17.27) the energy change between the free cantilever and the cantilever under tip-sample interaction is $2 \times 10^{-4} E_{\mathrm{free}}$ in the FM mode, which is thousand times smaller than in the AM mode according to (17.26).

According to the definition of the Q-factor in (2.41), a damped harmonic oscillator can gain/lose roughly $1/Q$th of its energy in per cycle $E_{\mathrm{diss}} = 2\pi E_{\mathrm{osc}}/Q$. Thus for a Q factor of 10,000 an energy of $6 \times 10^{-4} E_{\mathrm{free}}$ 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, increasing the intrinsic 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.

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.

---

[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 much smaller than the deviations in amplitude/frequency shift between the free cantilever and the situation with tip-sample interaction present.

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. 17.4. 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.

17.2.1.1 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 Fig. 14.9. 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 excitation amplitude. If energy is dissipated by the tip-sample interaction the oscillation amplitude would decrease. However, an increased excitation 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. 17.6.

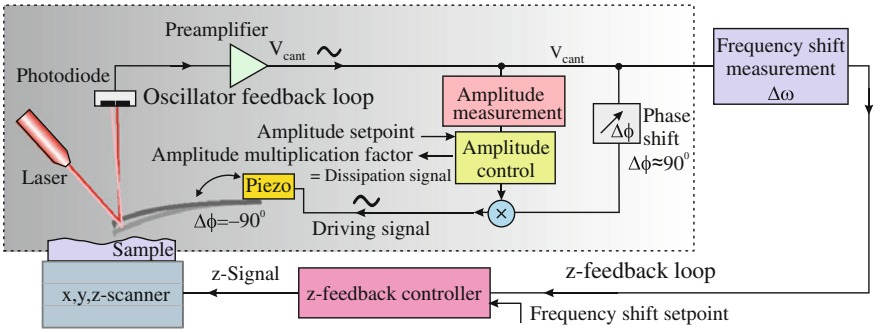

Fig. 17.6 Schematic of an FM detection setup operated with self-excitation including the amplitude control part. The cantilever oscillation amplitude is measured and maintained at a setpoint value by multiplying the driving signal by a proper multiplication factor. This factor relates to the energy dissipated by the tip-sample interaction

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. 17.6). 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 PI controller. In this way a constant cantilever oscillation amplitude is maintained by adjusting of 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 during a scan. The relation between the oscillation amplitude and the energy dissipated by the tip-sample interaction is given by (15.18).

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 (preamplifier) in the circuit have to be compensated. The phase shift in the box called the phase shift in Fig. 17.6 is adjusted (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).

To summarize, in the self-excitation mode the oscillation signal is fed back as the driving signal with a 90° phase shift. 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.

17.2.2 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

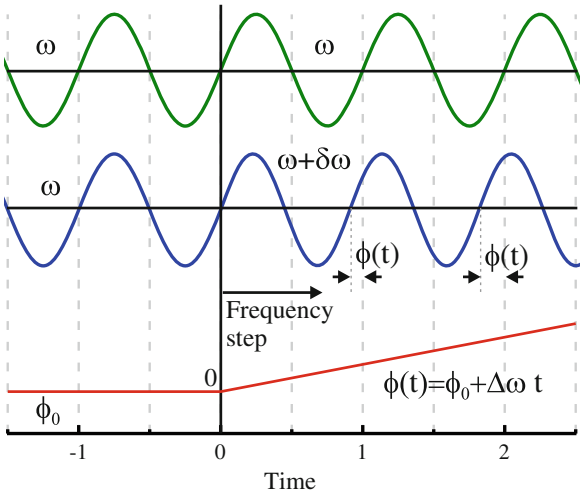

Fig. 17.7 The slightest frequency increase from ω to $\omega + \delta\omega$ leads to a linearly increasing phase $\phi(t)$. This phase (difference) can be detected using a phase detector. If the phase difference is maintained at zero, the two frequencies are exactly the same

at $t = 0$ form ω to $\omega + \delta\omega$, the oscillation can be expressed as $\cos[(\omega + \delta\omega)\, t + \phi_0]$. However, alternatively this expression can be rewritten as

$$\cos[(\omega + \delta\omega)\, t + \phi_0] = \cos[\omega t + (\delta\omega t + \phi_0)] = \cos(\omega t + \phi(t))\,, \qquad (17.28)$$

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. 17.7. The slightest frequency change corresponds to a linearly increasing phase signal. If the phase $\phi(t)$ is zero (or generally constant), the two frequencies are exactly the same.

In the following, the inner working of the frequency shift measurement (box in Fig. 17.6) 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 cantilever oscillation.

A PLL used in AFM is shown in Fig. 17.8 and consists of three main components: a phase detector, a Voltage-Controlled Oscillator (VCO), and a controller. First we introduce the phase detector and the VCO. Subsequently, their interaction in a phase-locked loop is described.

In the phase detector, 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

$$V_{\text{cant}} \cdot V_{\text{vco}} \propto \frac{1}{2}\left(\cos[(\omega_{\text{cant}} + \omega_{\text{vco}})t + \phi_0] + \cos[(\omega_{\text{vco}} - \omega_{\text{cant}})t + \phi_0]\right). \qquad (17.29)$$

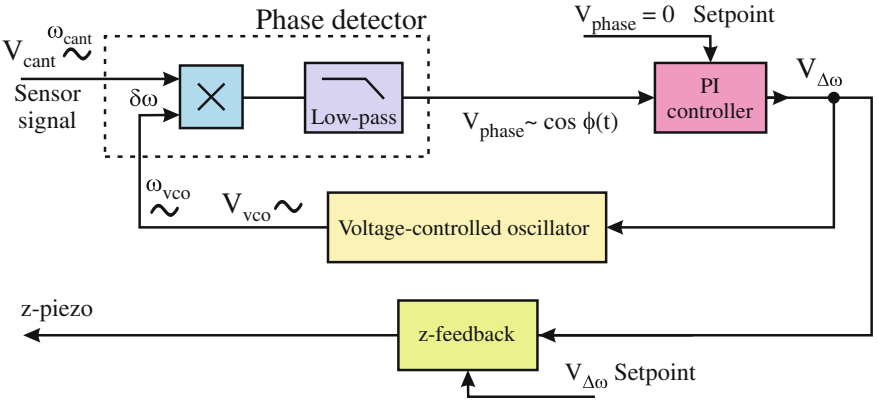

Fig. 17.8 The phase-locked loop consists of three main components: a phase detector, a Voltage-Controlled Oscillator (VCO) and a controller. These are combined to form a feedback loop in which the phase detector detects the phase difference between the cantilever oscillation signal V_{cant} and the VCO signal V_{vco}. The controller regulates the VCO frequency to a vanishing V_{phase}. This means that the VCO frequency adapts the 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 frequency shift $V_{\Delta\omega}$ is obtained at the output of the controller

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

$$V_{\text{phase}} \propto \cos\left[(\omega_{\text{vco}} - \omega_{\text{cant}})t + \phi_0\right] = \cos(\delta\omega t + \phi_0) = \cos(\phi(t)), \qquad (17.30)$$

with $\delta\omega = \omega_{\text{vco}} - \omega_{\text{cant}}$. The measured phase signal V_{phase} has the largest phase sensitivity for a phase close to 90°. Therefore, we consider $V_{\text{phase}} = 0$ 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 $V_{\text{phase}} \propto -\delta\omega t$. Including a proportionality factor K_{pd} which converts the phase into a voltage, the output voltage of the phase detector can be written as

$$V_{\text{phase}} = K_{\text{pd}} \cos\left(\delta\omega t + 90°\right) \approx -K_{\text{pd}}\delta\omega t. \qquad (17.31)$$

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

$$\omega_{\text{vco}} = \omega_{\text{work}} + K_{\text{vco}} V_{\Delta\omega}, \qquad (17.32)$$

with the proportionality factor K_{vco}, converting the input voltage $V_{\Delta\omega}$ to a frequency shift relative to the working frequency. The working frequency is the frequency of the free cantilever plus the frequency shift setpoint $\omega_{\text{work}} = \omega_{\text{free}} + \Delta\omega_{\text{set}}$.

Now we discuss the frequency tracking capability of the PLL. For the moment, we do not consider the PI controller shown in Fig. 17.8 and assume that the phase signal V_{phase} is directly fed into the input of the VCO, i.e. $V_{\text{phase}} = V_{\Delta\omega}$. Let us assume that initially the frequency of the VCO matches the oscillation frequency of the cantilever, $\omega_{\text{vco}} = \omega_{\text{cant}} = \omega_{\text{work}}$ and $\phi_0 = 90°$. At this working point $V_{\text{phase}} = 0$, which corresponds to the condition of maximum sensitivity for the phase, as shown above. In this case the input voltage at the VCO vanishes, i.e. $V_{\Delta\omega} = 0$.

Now we consider a change of the actual oscillation frequency of the cantilever, which results in a frequency difference $\delta\omega$ between the cantilever oscillation frequency and the VCO frequency. According to (17.31) this frequency difference leads to a phase difference signal measured by the phase detector $V_{\text{phase}} = K_{\text{pd}} \cos(\delta\omega t + \phi_0)$, which evolves approximately linearly with time. With this input, the output frequency of the VCO results according to (17.32) as

$$\omega_{\text{vco}} = \omega_{\text{work}} + K_{\text{pd}} K_{\text{vco}} \cos(\delta\omega t + \phi_0) \,. \tag{17.33}$$

Directly after the instantaneous frequency shift by $\Delta\omega$, the relations $\delta\omega = \Delta\omega$ and $\phi_0 = 90°$ hold. According to (17.33), the 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, i.e. $\delta\omega < \Delta\omega$. 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 (changed) frequency of the cantilever $\omega_{\text{work}} + \Delta\omega$. Due to this mechanism the VCO frequency is said to be locked to the cantilever frequency. In the steady-state $\omega_{\text{vco}} = \omega_{\text{cant}}$ and the frequency mismatch $\delta\omega = 0$ vanishes.[4]

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. 17.6).

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 (without z-feedback). The PLL converts this modulated high frequency signal to a

[4] While the PLL provides a frequency match $\omega_{\text{vco}} = \omega_{\text{cant}}$, a phase $\phi_0 \neq 0$ remains. The relation

$$\omega_{\text{cant}} = \omega_{\text{work}} + \Delta\omega \stackrel{!}{=} \omega_{\text{vco}} = \omega_{\text{work}} + K_{\text{pd}} K_{\text{vco}} \cos(\delta\omega t + \phi_0) \,, \tag{17.34}$$

results for the condition $\delta\omega = 0$ in

$$\Delta\omega = K_{\text{pd}} K_{\text{vco}} \cos\phi_0 . \tag{17.35}$$

Thus a static phase difference ϕ_0 different from $\phi_0 = 90°$ evolves in order to adapt the VCO frequency to the changed cantilever frequency.

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.

Without the use of the PI controller (not yet applied) the frequency match of the VCO frequency is achieved by a phase ϕ_0 different from 90°, as shown in (17.35). Thus the desired working point at $\phi_0 = 90°$ is left. In order to enforce a vanishing phase signal (i.e. to maintain the condition $\phi_0 = 90°$) a PI controller is used, which controls V_{phase} to zero by generating an appropriate controller output signal $V_{\Delta\omega}$, which is used as the input voltage for the voltage-controlled oscillator.

17.2.3 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 deviate from the ideal sinusoidal shape. Moreover, a cantilever 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 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. 17.9). The cantilever deflection signal (sensor signal) is fed to the input of the PLL (we neglect the amplitude control for the moment).

In the following, we analyze the time constants of the PLL tracking mode and obtain the result that this mode has a larger time constant than the self-excitation mode. We consider an instantaneous jump of the cantilever resonance frequency due to a tip-sample interaction from ω_0 to ω_0'. Initially after this jump the excitation frequency (PLL output) still remains at ω_0. This corresponds to the situation in the AM detection mode: excitation at a fixed frequency ω_0 and instantaneous change of the cantilever resonance frequency. For the case of AM detection, we found in Sect. 14.5 that the amplitude and now more importantly the phase changes with a time constant of $\tau_{\text{cant}} = 2Q/\omega_0$. The PLL detects this slowly changing phase and adapts the VCO frequency with the time constant τ_{cant} to the cantilever frequency.[5] Thus

[5] This is the case for a PLL with a fast time constant. If the PLL has a time constant longer than τ_{cant}, the PLL time constant will limit the overall time constant.

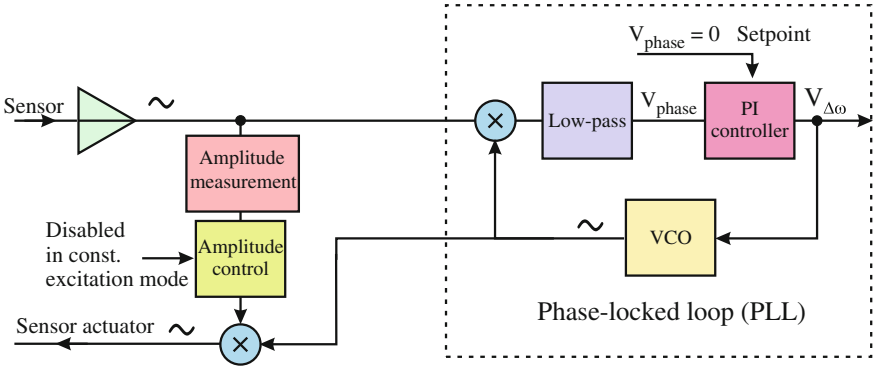

Fig. 17.9 Schematic of an FM AFM control in the PLL tracking mode. In this mode, the sensor is excited by a very clean sinusoidal driving signal taken from the voltage-controlled oscillator (VCO)

the PLL tracking mode is, compared to the self-excitation mode, a slow detection mode. As an example, for a Q-factor of 10^4 and $f_0 = 150\,\text{kHz}$ a time constant $\tau_{\text{cant}} = 130\,\text{ms}$ results.

Another disadvantage in using the excitation signal from the PLL is the following. If the PLL becomes unlocked, the cantilever will no longer be excited at its resonance and the z-feedback will not work properly anymore. In the self-excitation scheme the cantilever always oscillates at its resonance frequency independent of the PLL frequency detection.

The PI controller in the PLL loop (Fig. 17.9) is of specific importance if the VCO excites a harmonic oscillator (the cantilever) at resonance, as is the case in the PLL tracking mode.[6] Without the PI controller, according to (17.35), any deviation from the working frequency $\Delta\omega$ leads to a constant phase shift ϕ_0 different from 90°. This means that the cantilever is excited with a phase deviating from the proper resonance phase 90°. Specifically for cantilevers with high Q-factors, even a small phase shift leads to a driving out of resonance. The desired driving of the cantilever at resonance can be maintained by the use of a PI controller. Using the PI controller in the PLL loop, the phase signal ($V_{\text{phase}} = \cos(\delta\omega t + \phi_0)$) 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.

[6] In the PLL circuits used for example in communications, the PI controller is often not included.

17.3 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. 17.3a and schematically in Fig. 17.10a. Due to this, the tip-sample distance can only be controlled in a certain range of distances. As shown in the following, instabilities occur outside of this range.

In STM the measured signal (tunneling current) increases monotonously (exponentially) with decreasing tip-sample distance. This leads to stable feedback, i.e. the feedback controller "knows what to do". If the current becomes larger (e.g. due to

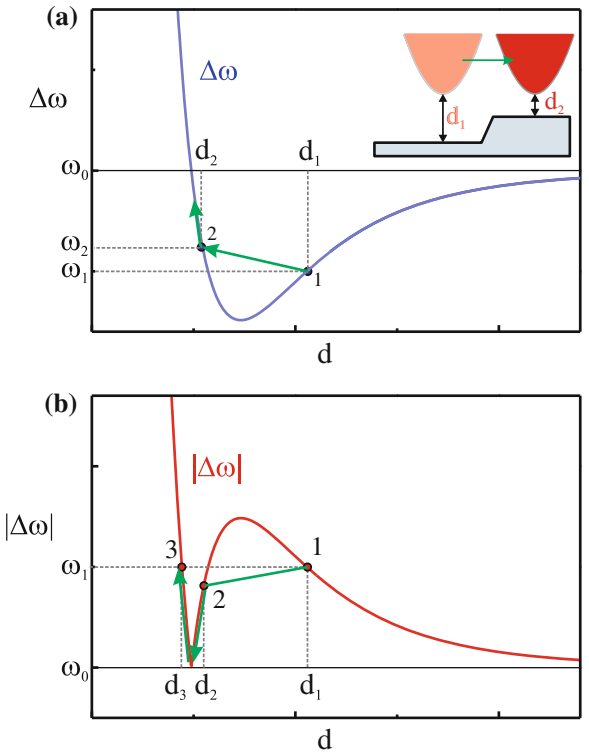

Fig. 17.10 Instabilities arise due to the non-monotonous dependence of the measured frequency shift as a function of the tip-sample distance. **a** An instability in the attractive regime at working point d_1 (induced for instance due to a fast scan over a steep step edge (inset)) results at a new working point at d_2 with opposite slope, leading to a wrong direction of the 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. 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

moving over a step edge), the tip has to be withdrawn from the sample in order to recover the desired tip-sample distance. 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 (point 1 in Fig. 17.10a). 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. 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 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 lateral change of the interaction potential (i.e. a branch of opposite slope is reached for a different lateral tip position on the sample).

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. 17.10b. 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 stable working point 3 is reached. Thus, using the absolute value of the frequency shift signal avoids catastrophic tip crashes and stabilizes the feedback (in the case of an instability) in the repulsive regime. However, the intended working point in the attractive regime is 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.

17.4 Comparison of Different AFM Modes

In the previous chapters, we have discussed several modes of AFM operation, which we will now compare. In Table 17.1 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

Table 17.1 Operating modes of AFM ordered in two "coordinates": static/dynamic mode and attractive/net-repulsive interactions

	Static AFM	Dynamic AFM
Net-repulsive interaction Contact	Contact mode: k ∼ 1 N/m	Tapping mode: k ∼ 20–100 N/m
Attractive interaction Non-contact	Non-contact mode: k ∼1 N/m	AM/FM non-contact mode: k ∼ 20–10^6 N/m

with non-contact AFM (attractive interaction). However, also the off-diagonal elements in Table 17.1 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. 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.

17.5 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

$$\Delta f = -\frac{f_0}{A^2 k}\langle F_{\mathrm{ts}}(t)\cdot z(t)\rangle = -\frac{f_0}{\pi k A^2}\int_{-A}^{+A} F_{\mathrm{ts}}(d+z)\frac{z}{\sqrt{A^2-z^2}}dz. \quad (17.36)$$

- 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

$$\gamma = \Delta f \frac{kA^{3/2}}{f_0} \,. \tag{17.37}$$

  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 the case of 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).

Chapter 18
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 which we want to measure. If atomic steps are to be measured, this is about 1 Å an atomic corrugation can have a much smaller signal of less than 0.1 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 with which the signal is detected.

In Sect. 12.3 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 $\sqrt{\langle \Delta z_{\mathrm{th}}^2 \rangle}$. In this chapter the effect of the thermal noise amplitude on the experimentally measured quantities in AFM such as the frequency shift is estimated.

18.1 Thermal Noise Density of a Harmonic Oscillator

The thermal displacement noise of the AFM cantilever can be estimated from the equipartition theorem, which states that each degree of freedom carries an average energy of $1/2\,k_B T$ 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 $E = 1/2\,kz^2 + 1/2\,mv^2$, 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 $k_B T$.

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_18

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

$$\left\langle \Delta z_{\text{th}}^2 \right\rangle = \frac{k_B T}{k}. \tag{18.1}$$

At room temperature and for a spring constant of $k = 10\,\text{N/m}$, an amplitude of $\sim 0.2\,\text{Å}$ 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 variable was introduced in Sect. 5.4. The noise variable 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)

$$N_{\text{th,osc}}(f) = N_{\text{th,exc}} G(f), \tag{18.2}$$

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, the transfer function of the harmonic oscillator is

$$\frac{A^2}{A_{\text{drive}}^2} \equiv G^2(f) = \frac{1}{\left(1 - \frac{f^2}{f_0^2}\right)^2 + \frac{1}{Q^2}\frac{f^2}{f_0^2}}. \tag{18.3}$$

The mean square thermal displacement can be calculated in analogy to (5.10). Another expression for the mean square displacement was obtained from the equipartition theorem as (18.1). Thus the following equation results

$$\left\langle \Delta z_{\text{th}}^2 \right\rangle = \int_0^\infty N_{\text{th,osc}}^2(f) df = N_{\text{th,exc}}^2 \int_0^\infty G^2(f) df = \frac{k_B T}{k}. \tag{18.4}$$

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 ($\int_0^\infty G^2(f) df = \pi Q f_0/2$), when the integration

limits are inserted. With this, the spectral noise density of a harmonic oscillator results as

$$N_{\mathrm{th,osc}}(f) = N_{\mathrm{th,exc}} G(f) = \sqrt{\frac{2k_B T}{\pi k Q f_0}}\, G(f). \tag{18.5}$$

Thus the spectral noise density of the harmonic oscillator consists of the strongly peaked transfer function of the harmonic oscillator $G(f)$ shown in Fig. 18.1a for two different Q-factors and a frequency independent white thermal excitation noise density given by (18.5) as

$$N_{\mathrm{th,exc}} = \sqrt{\frac{2k_B T}{\pi k Q f_0}}. \tag{18.6}$$

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. 18.1b. 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 (5.11) as

$$\left\langle \Delta z_{\mathrm{th}}^2(f_1, f_2) \right\rangle = \int_{f_1}^{f_2} N_{\mathrm{th,osc}}^2(f) df = \frac{2k_B T}{\pi k Q f_0} \int_{f_1}^{f_2} G^2(f) df. \tag{18.7}$$

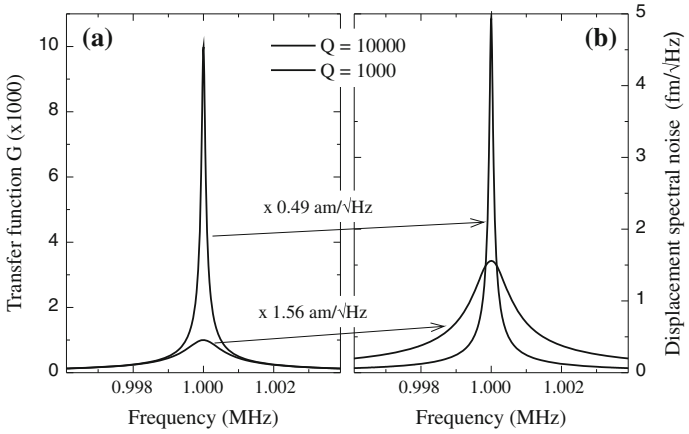

Fig. 18.1 **a** Transfer function of the harmonic oscillator $G(f)$. **b** Corresponding displacement spectral noise density at room temperature. The multiplication factor $N_{\mathrm{th,exc}}$ for going from (**a**) to (**b**) depends on the Q-factor

This equation will be used in the following in order to evaluate the mean square displacement in various circumstances.

18.2 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 (18.7), the mean square displacement in the static mode results with $B = f_2 - f_1$ as

$$\left\langle \Delta z^2_{\mathrm{th,stat}} \right\rangle = \frac{2k_B T B}{\pi k Q f_0}. \tag{18.8}$$

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 $\Delta F = k\sqrt{\langle \Delta z^2_{\mathrm{th}} \rangle}$. Therefore, the minimum detectable force (due to thermal noise) in static AFM (i.e. at low frequencies off-resonance) is

$$F^{\mathrm{static}}_{\mathrm{min,th}} = \sqrt{\frac{2k k_B T B}{\pi Q f_0}}. \tag{18.9}$$

18.3 Thermal Noise in the Dynamic AFM Mode with AM Detection

Here we consider a 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 (18.7), the mean square displacement in the dynamic mode results as

$$\left\langle \Delta z^2_{\mathrm{th,res}} \right\rangle = \frac{2k_B T Q(2B)}{\pi k f_0}, \tag{18.10}$$

with $2B$ being the two sided bandwidth, i.e. from $f_0 - B$ to $f_0 + B$. The thermal displacement noise (18.10) is Q times higher in the dynamic case than in the static case (18.8). 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. 14.3. 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 $\sqrt{\langle \Delta z_{\text{th}}^2 \rangle}$ is one half of that derived from (18.10).

As we have shown before in (14.8), in dynamic AFM for small amplitudes, the force gradient is related to the measured frequency shift by $\partial F/\partial z = \Delta f 2k/f_0$ (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

$$\frac{\partial F}{\partial z} = \frac{2k}{f_0}\Delta f = \frac{2k}{f_0}\frac{\Delta f}{\Delta A}\Delta A. \tag{18.11}$$

The inverse slope of the resonance curve at the working point can be written according to (2.33) as $\Delta f/\Delta A \approx f_0/(QA)$. If we identify the amplitude change ΔA with the thermal noise $\sqrt{\langle \Delta z_{\text{th}}^2 \rangle}$, the minimum detectable force gradient can be written as

$$\frac{\partial F}{\partial z} = \frac{2k}{f_0}\frac{f_0}{QA}\Delta A = \frac{2k}{QA}\sqrt{\frac{k_B T Q(2B)}{\pi k f_0}} = \sqrt{\frac{4k k_B T(2B)}{\pi Q f_0 A^2}}. \tag{18.12}$$

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. 14.5. 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 $\Delta A = \sqrt{\langle \Delta z_{\text{th}}^2 \rangle}$ translates to a noise in the topography signal z' via the slope of the amplitude distance relation dA/dz' as

$$\Delta z' = \Delta A \frac{dz'}{dA} = \frac{\sqrt{\langle \Delta z_{\text{th}}^2 \rangle}}{dA/dz'}. \tag{18.13}$$

Here the mean square displacement has to be taken from (18.10). For stiff materials the slope dA/dz' is about one, while it has a smaller value for soft materials and the noise in the topography signal becomes correspondingly larger.

18.4 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 C. 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 (18.7), we have to evaluate the transfer function at $f_0 + f_{\mathrm{mod}}$. When evaluating $G(f_0 + f_{\mathrm{mod}})$ 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 (18.3) and use (2.30), resulting in

$$G^2(f_0 + f_{\mathrm{mod}}) = \frac{1}{\left(1 - \frac{(f_0+f_{\mathrm{mod}})^2}{f_0^2}\right)^2 + \frac{1}{Q^2}\frac{(f_0+f_{\mathrm{mod}})^2}{f_0^2}} \approx \frac{1}{4\frac{f_{\mathrm{mod}}^2}{f_0^2} + \frac{1}{Q^2}}. \quad (18.14)$$

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 (18.14) can be neglected. In this case, the thermal displacement noise density can, according to (18.6), be written as

$$N_{\mathrm{th,osc}}^2(f_0 + f_{\mathrm{mod}}) \equiv N_{\mathrm{z,th}}^2(f_0 + f_{\mathrm{mod}}) = N_{\mathrm{th,exc}}^2 G^2(f_0 + f_{\mathrm{mod}}) = \frac{k_B T f_0}{2\pi k Q f_{\mathrm{mod}}^2}. \quad (18.15)$$

We change the notation here in order to distinguish between the thermal displacement noise density to $N_{\mathrm{z,th}}^2(f_0 + f_{\mathrm{mod}})$, and the thermal frequency noise density after demodulation $N_{f,\mathrm{th}}(f_0 + f_{\mathrm{mod}})$. In FM modulation, the displacement noise is transferred to a frequency noise and according to (C.11) in Appendix C we can write

$$N_{f,\mathrm{th}}(f_{\mathrm{mod}}) = \frac{\sqrt{2} f_{\mathrm{mod}}}{A} N_{\mathrm{z,th}}(f_0 + f_{\mathrm{mod}}). \quad (18.16)$$

For the thermal displacement noise according to (18.15), the frequency noise density results for the case $f_{\mathrm{mod}} > f_0/(2Q)$ as

$$N_{f,\mathrm{th}} = \sqrt{\frac{k_B T f_0}{\pi k Q A^2}} = const., \quad (18.17)$$

which does not depend on f_{mod}.

In the general case (independent of the limit $f_{\mathrm{mod}} > f_0/(2Q)$) the thermal displacement noise density can be written using (18.6) and (18.14) as

[1] We do not indicate explicitly that the carrier frequency is the shifted resonance frequency f_0'.

$$N^2_{z,\mathrm{th}}(f_{\mathrm{mod}}) = N^2_{\mathrm{th,exc}} G^2(f_0 + f_{\mathrm{mod}}) = \frac{2k_B T}{\pi k Q f_0} \frac{1}{4\frac{f^2_{\mathrm{mod}}}{f_0^2} + \frac{1}{Q^2}}. \tag{18.18}$$

The thermal frequency noise in FM detection can be calculated analogously to (5.11) and (18.16) by integration over f_{mod} up to the maximum $f_{\mathrm{mod,max}} = B$ as

$$\begin{aligned}\left\langle \Delta f^2_{\mathrm{th}} \right\rangle &= \int_0^B N^2_{f,\mathrm{th}}(f_{\mathrm{mod}}) df_{\mathrm{mod}} \\ &= \frac{4k_B T}{\pi A^2 k Q f_0} \int_0^B \frac{f^2_{\mathrm{mod}}}{4\frac{f^2_{\mathrm{mod}}}{f_0^2} + \frac{1}{Q^2}} df_{\mathrm{mod}} \\ &= \frac{k_B T}{\pi A^2 k Q} \left[f_0 B - \frac{1}{2Q} f_0^2 \arctan\left(\frac{2QB}{f_0}\right)\right]. \end{aligned} \tag{18.19}$$

The noise contributions from frequencies lower than ω_0 are already included by the factor $\sqrt{2}$ in (18.16). Thus in the FM case B is defined as $B = f_{\mathrm{mod,max}}$, i.e. as a single sided bandwidth.

If $B \gg f_0/(2Q)$, the second term in (18.19) can be neglected. In this limit the minimum detectable force gradient due to thermal noise can be written as

$$\frac{\partial F}{\partial z} = \frac{2k}{f_0}\sqrt{\left\langle \Delta f^2_{\mathrm{th}} \right\rangle} = \sqrt{\frac{4k k_B T B}{\pi Q f_0 A^2}}. \tag{18.20}$$

The mean square thermal displacement, which was calculated in (18.10) under the simplified assumption that $G^2 = Q^2$, can be calculated considering the integration over the transfer function (18.14) within twice the single-sided bandwidth as

$$\begin{aligned}\left\langle \Delta z^2_{\mathrm{th}} \right\rangle &= \int_{-B}^B N^2_{z,\mathrm{th}}(f_0 + f_{\mathrm{mod}}) df_{\mathrm{mod}} \\ &= \frac{2k_B T}{\pi k Q f_0} \int_{-B}^B \frac{1}{4\frac{f^2_{\mathrm{mod}}}{f_0^2} + \frac{1}{Q^2}} df_{\mathrm{mod}} \\ &= \frac{2k_B T}{\pi k} \arctan\left(\frac{Q2B}{f_0}\right). \end{aligned} \tag{18.21}$$

18.5 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. constant as a function of frequency within the considered detection bandwidth. Thus the mean square displacement due to the sensor displacement noise results according to (5.10) as

$$\left\langle \Delta z_{\mathrm{sens}}^2 \right\rangle = \int_0^B N_{\mathrm{z,sens}}^2 df = N_{\mathrm{z,sens}}^2 B. \tag{18.22}$$

Further, the minimum detectable force in the static mode results as

$$F_{\mathrm{min,sens}}^{\mathrm{static}} = k\sqrt{\left\langle \Delta z_{\mathrm{sens}}^2 \right\rangle} = kN_{\mathrm{z,sens}}\sqrt{B}. \tag{18.23}$$

In the dynamic mode the minimum detectable force gradient due to the sensor displacement noise results according to (18.12) as

$$\frac{\partial F}{\partial z} = \frac{2k}{QA}\sqrt{\left\langle \Delta z_{\mathrm{sens}}^2 \right\rangle} = \frac{\sqrt{2}k}{QA} N_{\mathrm{z,sens}}\sqrt{2B}, \tag{18.24}$$

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 (C.11) as

$$N_{f,\mathrm{sens}}(f_{\mathrm{mod}}) = \frac{\sqrt{2}\, f_{\mathrm{mod}}}{A} N_{\mathrm{z,sens}}. \tag{18.25}$$

The mean square frequency noise resulting from the sensor displacement noise is

$$\left\langle \Delta f_{\mathrm{sens}}^2 \right\rangle = \int_0^B N_{f,\mathrm{sens}}^2(f_{\mathrm{mod}}) df_{\mathrm{mod}} = \frac{2N_{\mathrm{z,sens}}^2}{A^2} \int_0^B f_{\mathrm{mod}}^2 df_{\mathrm{mod}}. \tag{18.26}$$

Thus the frequency noise due to the sensor displacement noise results as

$$\sqrt{\left\langle \Delta f_{\mathrm{sens}}^2 \right\rangle} = \sqrt{\frac{2N_{\mathrm{z,sens}}^2}{3A^2} B^3}. \tag{18.27}$$

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 (18.20), as

$$\frac{\partial F}{\partial z} = \frac{2k}{f_0}\sqrt{\langle \Delta f_{\text{sens}}^2 \rangle} = \sqrt{\frac{8}{3}}\frac{k N_{\text{z,sens}} B^{3/2}}{f_0 A}. \tag{18.28}$$

18.6 Total Noise in the FM Detection Mode

An example of an actually measured frequency noise density as a function of the modulation frequency is shown in Fig. 18.2. The experimentally measured noise density is characterized by a very small constant offset due to thermal noise (18.17) and a linear increase of the noise density with the modulation frequency according to (18.25). These two independent noise contributions add up to a total noise density as $N_{f,\text{tot}}^2 = N_{f,\text{th}}^2 + N_{f,\text{sens}}^2$. Due to the bandwidth of the frequency demodulator electronics, which has a bandwidth limit of 1 kHz, the measured noise density levels off and decreases beyond this frequency.

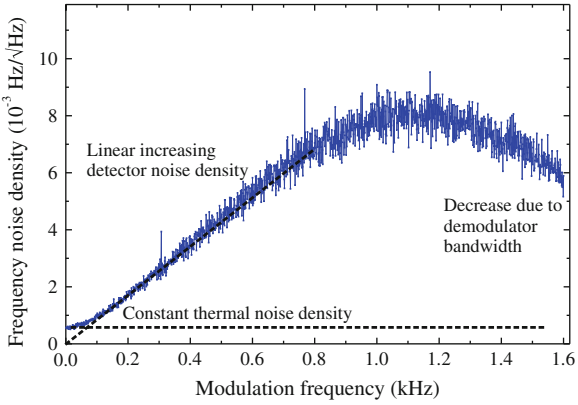

Fig. 18.2 Experimentally measured frequency noise density $N_{f,\text{sens}}(f_{\text{mod}})$ of an FM atomic force microscopy setup

Table 18.1 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
$N_{\text{z,sens}}$ (fm/$\sqrt{\text{Hz}}$)	100	50	2

Table 18.2 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	Eq. no.
Static	$\sqrt{\langle \Delta z^2 \rangle}$ (fm)	94	3.9	0.013	(18.8)
	$F_{\mathrm{min,th}}$ (pN)	0.94	6.9	14	(18.9)
	$F_{\mathrm{min,sens}}$ (pN)	32	2,800	68,000	(18.23)
AM	$\sqrt{\langle \Delta z_{\mathrm{th}}^2 \rangle}$ (pm)	40	17	0.27	(18.10)
	$(\partial F/\partial z)_{\mathrm{th}}$ (N/m)	0.0005	0.1	0.3	(18.12)
	$\sqrt{\langle \Delta z_{\mathrm{sens}}^2 \rangle}$ (pm)	3.2	1.6	0.063	(18.22)
	$(\partial F/\partial z)_{\mathrm{sens}}$ (N/m)	0.0001	0.019	0.091	(18.24)
FM	$\sqrt{\langle \Delta z_{\mathrm{th}}^2 \rangle}$ (pm)	19	1.5	0.061	(18.21)
	$(\partial F/\partial z)_{\mathrm{th}}$ (N/m)	0.0003	0.5	0.2	(18.20)
	$\sqrt{\langle \Delta f_{\mathrm{th}}^2 \rangle}$ (Hz)	1.7	0.9	0.09	(18.19)
	$\sqrt{\langle \Delta f_{\mathrm{sens}}^2 \rangle}$ (Hz)	0.6	13	0.52	(18.27)
	$(\partial F/\partial z)_{\mathrm{sens}}$ (N/m)	0.0001	1.5	1.1	(18.28)

The total noise is composed of the thermal noise and the detector noise (and other sources of noise such as the oscillator noise [37], which we neglect here for simplicity). If these are assumed to be independent, the total noise results as

$$\Delta f_{\mathrm{total}} = \sqrt{\Delta f_{\mathrm{th}}^2 + \Delta f_{\mathrm{sens}}^2}. \tag{18.29}$$

The minimum detectable force gradients combine correspondingly. In Table 18.1, 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 [38]. In Table 18.2 numerical values for the noise estimated in this chapter are compared for the three different kinds of sensors.

18.7 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 (5.26) we have seen that the fundamental limit for the detection of the tunneling current using a transimpedance amplifier is the Johnson noise in the feedback resistor, which was written as

$$\Delta I = \sqrt{4k_B T B/R}. \tag{18.30}$$

For a 100 MΩ resistor and a bandwidth of 3 kHz, a (RMS) noise current of $\Delta I = 0.3\,\text{pA}$ 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

$$\Delta z = \frac{\Delta I}{|dI/dz|}. \tag{18.31}$$

Assuming a tunneling current of $I_0 = 0.1\,\text{nA}$ at the working point and $\kappa = 0.1\,\text{Å}^{-1}$, the slope of the $I(z)$ curve results as $dI/dz = -2\kappa I_0$. 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 (18.30) 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 us to work with a larger bandwidth in STM compared to FM detection in AFM.

18.8 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 $\Delta f/f_0$ 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 for which the signal-to-noise ratio is largest [33]. 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

$$F(u) = F_0 e^{-u/\lambda}, \tag{18.32}$$

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 (17.20) and (17.24)) be expressed as

$$\frac{\Delta f}{f_0} = \frac{1}{\sqrt{2\pi}} \frac{F\sqrt{\lambda}}{kA^{3/2}}. \tag{18.33}$$

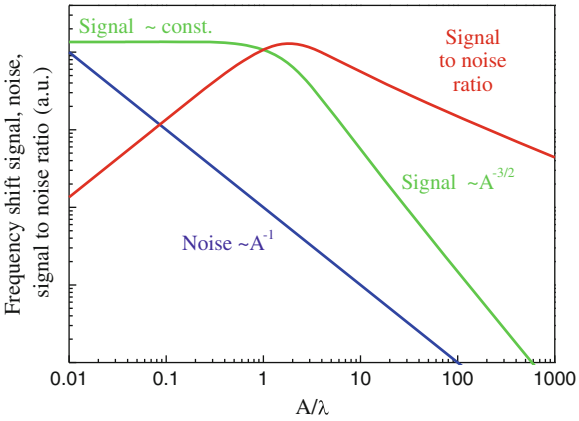

Fig. 18.3 The frequency shift signal, the corresponding noise and the signal-to-noise ratio in FM detection are shown as a function of the cantilever (sensor) oscillation amplitude A, which is normalized to the tip-sample interaction length λ. (adapted from [33])

This means that for large amplitudes the frequency shift signal depends on the amplitude proportional to $A^{-3/2}$, as shown in Fig. 18.3.

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 (14.8). This can be evaluated further using the force law in (18.32) as

$$\frac{\Delta f}{f_0} = \frac{k_{\rm ts}}{2k} = \frac{-F'}{2k} = \frac{F}{2k\lambda}. \tag{18.34}$$

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. 18.3 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. 18.3.

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 (18.19) and (18.27). In Fig. 18.3 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 $\lambda^{\rm short} = 0.1\,{\rm nm}$ and an interaction with a range of $\lambda^{\rm long} = 5\,{\rm nm}$. If we consider

the limiting case $A > \lambda^{\text{long}}$, using (18.33) 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 (18.34), 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.

18.9 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.

Chapter 19
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). It was discovered that 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.

19.1 Tuning Fork Quartz Sensor

One example of a quartz sensor is the quartz tuning fork, frequently used in wristwatches, as shown in Fig. 19.1a on the lower right. In Fig. 19.1b an encapsulated tuning fork quartz oscillator is shown (left), as well as one without housing (right). 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, due to its use in watches. The bending mode of such a tuning fork is like that known from a macroscopic tuning fork as indicated by the arrows in Fig. 19.1b. 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, one prong is fixed to a holder with high mass as shown in Fig. 19.1c. This configuration is called qPlus configuration [33].

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 bending oscillation of the prong of the tuning fork is performed electrically using

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_19

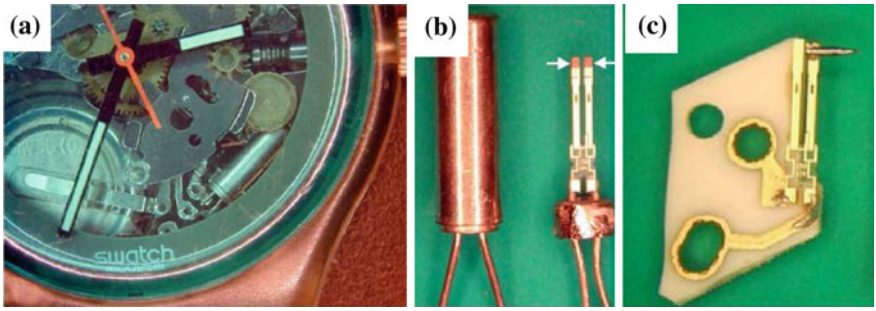

Fig. 19.1 **a** A tuning fork quartz oscillator as used in wristwatches. **b** Tuning fork oscillator with and without housing. **c** The tuning fork oscillator can be used as a force sensor in AFM. One prong is glued to a support and a tip is attached to the other prong (reproduced with permission from [39])

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.

19.2 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. 19.2. 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. 19.2, which allows for an electrical excitation without any additional driving piezo by applying the AC driving signal 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. 19.3a. 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. 19.3b. 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. 19.4. 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.

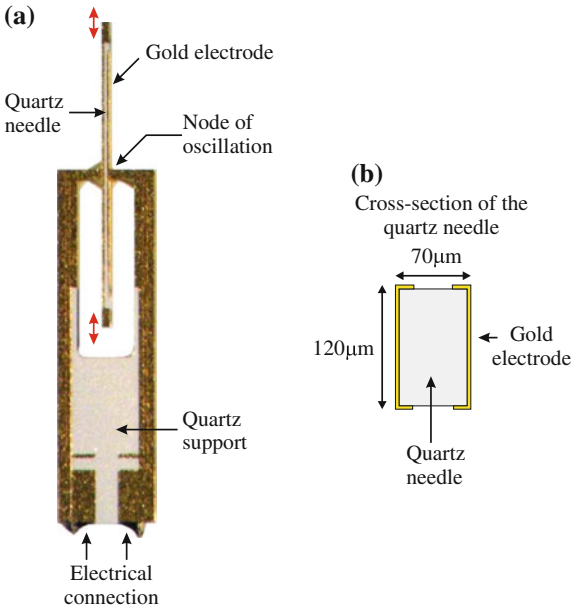

Fig. 19.2 Photo of a needle sensor (**a**) and schematic cross section through the needle (**b**). The needle sensor is an extensional type quartz oscillator which can be used as force sensor in AFM

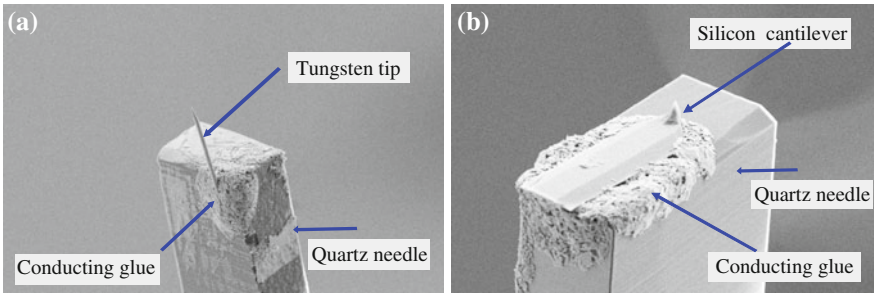

Fig. 19.3 Tips glued to the top of a needle sensor. **a** Electrochemically etched tungsten tip. **b** End part of a silicon cantilever

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 measured at the sample and used as a feedback signal for control of the tip-sample distance. We call this mode of operation 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.

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).

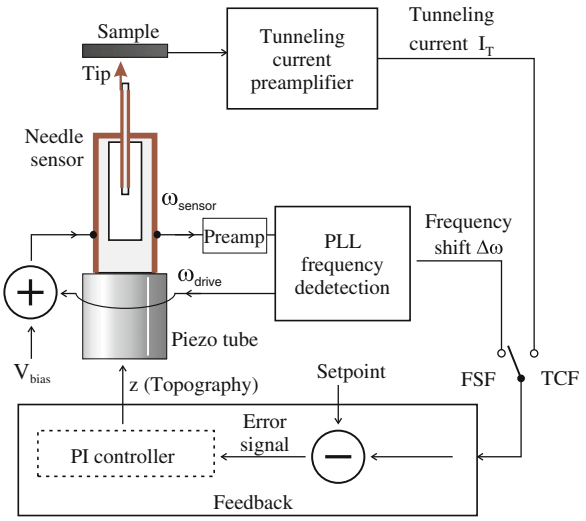

Fig. 19.4 Schematic circuit for driving the needle sensor as a force sensor in AFM. Alternatively the frequency shift signal (FSF) or the tunneling current (TCF) can be used as feedback signals [40]

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 19.1 typical properties of three types of AFM force sensors are compared: silicon cantilever, tuning fork and needle sensor. 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

Table 19.1 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 k–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 k–20 k	5 k–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 $10\,\text{nN} < kA_{min}$

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.

19.3 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

$$A_{\text{sensor}} = S_{\text{sensor}} V_{\text{sensor}}. \tag{19.1}$$

The cantilever oscillation amplitudes are only known from the experimental quantity V_{sensor} in volt. In the calibration procedure, the sensitivity factor is determined which converts V_{sensor} to an oscillation amplitude A_{sensor} in nm. The aim of the amplitude calibration procedure is to find the calibration factor S_{sensor} (in nm per volt). 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. 12.5. 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. 19.5 we compare two cases of different oscillation amplitudes A_{sensor} and A'_{sensor}. When the tips are brought close to the surface and a certain frequency shift setpoint Δf is requested, this will result in different values for the average tip-sample position of the cantilever d, as shown in Fig. 19.5. Since the main contribution to the frequency shift signal comes from the lower turnaround point of the oscillation (as shown in Chap. 17), 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

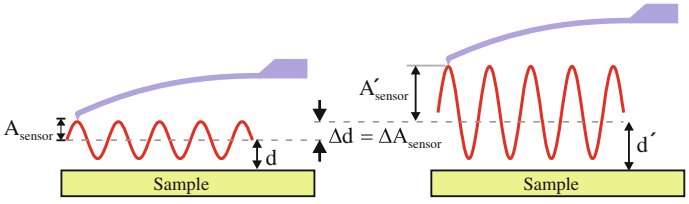

Fig. 19.5 Principle of the determination of the oscillation amplitude used for quartz sensors. The distance between the lower turnaround point of the tip oscillation and the sample is approximately the same for different oscillation amplitudes. Thus the change of the sensor amplitude ΔA_{sensor} is equal to the retraction of the equilibrium position of the tip Δd. A cantilever sensor is shown schematically instead of a quartz sensor

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

$$S_{\text{sensor}} = \frac{\Delta A_{\text{sensor}}}{\Delta V_{\text{sensor}}} = \frac{\Delta d}{\Delta V_{\text{sensor}}}. \tag{19.2}$$

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 (17.18)) is measured as a function of the tip-sample distance d [41]. The measured frequency shift curves have the usual (Lennard-Jones-type) shape, as shown in Fig. 19.6. 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. 17.3). 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. 19.5. The mutual shift (for a voltage increase ΔV_{sensor} of 0.1 V) amounts to about $\Delta d = 0.5$ nm as indicated in Fig. 19.6. 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 $S_{\text{sensor}} = 5$ nm/V can be obtained from the relation

$$\Delta A_{\text{sensor}} = \Delta d = S_{\text{sensor}} \Delta V_{\text{sensor}}. \tag{19.3}$$

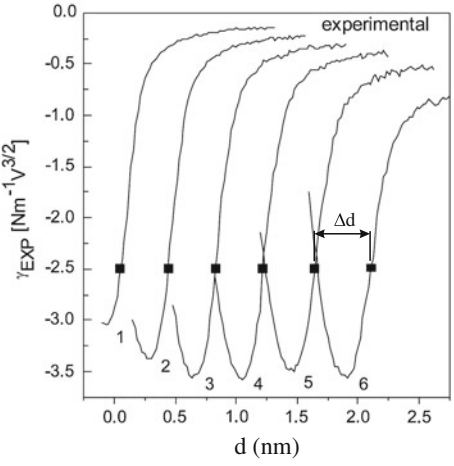

Fig. 19.6 Normalized frequency shift as a function of the average tip-sample position d, for different cantilever oscillation amplitudes. The sensor amplitude V_{sensor} increases from *curve 1* to *curve 6* from 0.2 to 0.7 V, respectively. Each increase of the amplitude by 0.1 V leads to a shift of the curves by $\Delta d = 0.5$ nm, showing the proportionality $\Delta d \propto \Delta V_{\text{sensor}}$ (reproduced with permission from [41])

19.4 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.

Part III
Scanning Tunneling Microscopy and Spectroscopy

Chapter 20
Scanning Tunneling Microscopy

The problem of the tunneling junction (electrode-gap-electrode) can be treated in different approximations. First we consider a simple wave function matching at a one-dimensional square barrier. Then the Bardeen model of tunneling will be considered. Subsequently, we discuss the Bardeen model of tunneling in several approximations. First we focus on the energy dependence, while restricting the analysis to a one-dimensional barrier. This will lead to an energy- and bias-dependent transmission factor and relates the tunneling current to the energy dependence of the density of states of the sample. In the complementary Tersoff-Hamann approximation, the tunneling voltage is considered to be small so that only electrons close to the Fermi level contribute to the tunneling current. In this approximation, the tunneling current is proportional to the density of states of the sample at the position of the tip. Finally, it is shown that voltage-dependent imaging makes it possible to distinguish chemical elements with atomic precision.

20.1 One-Dimensional Potential Barrier Model

As a one-dimensional model for tunneling we consider a square potential barrier $V(z)$ of height V_0 above the bottom of the potential ($V = 0$) in the region between $z = 0$ and $z = d$, as shown in Fig. 20.1. E is the energy of the electron tunneling through the barrier. The time-dependent Schrödinger equation reads as

$$i\hbar\frac{\partial}{\partial t}\Psi(\boldsymbol{r},t) = \left(-\frac{\hbar^2}{2m}\Delta + V(\boldsymbol{r},t)\right)\Psi(\boldsymbol{r},t). \tag{20.1}$$

Since the potential is not time-dependent the time dependence of the solution can be split into a separate factor $\phi(t) = \exp(-\frac{i}{\hbar}Et)$ as known from quantum mechanics, i.e. $\Psi(\boldsymbol{r},t) = \exp(-\frac{i}{\hbar}Et)\psi(\boldsymbol{r})$. This means that there are solutions with a definite fixed (time-independent) energy E. The spatial dependence of the solution of the

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_20

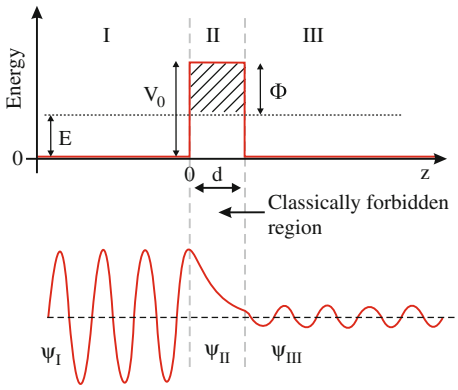

Fig. 20.1 One-dimensional metal-vacuum-metal tunneling junction. If the barrier is thin, the wave function can penetrate to the other side of the barrier

Schrödinger equation, $\psi(\boldsymbol{r})$, can be obtained from the time-independent Schrödinger equation which can be written in the one-dimensional case as

$$\frac{\hbar^2}{2m}\frac{\partial^2}{\partial z^2}\psi(z) = [V(z) - E]\,\psi(z). \tag{20.2}$$

We will first find solutions for regions I, II, and III in Fig. 20.1 separately. If we insert the expression for a right traveling plane wave, $\psi = e^{ikz}$, into the time-independent Schrödinger equation, (20.2), this results in

$$-\frac{\hbar^2}{2m}k^2 = V(z) - E. \tag{20.3}$$

We obtain for k

$$k = \sqrt{\frac{2m}{\hbar^2}\,[E - V(z)]}. \tag{20.4}$$

In regions I and III, outside the barrier, $V = 0$, and the solution has the form of an oscillating wave[1] for a free electron

$$\psi_{\text{free}} = e^{ikz} \text{ with } k = \sqrt{\frac{2m}{\hbar^2}E}. \tag{20.5}$$

[1] The wave function as drawn in Fig. 20.1 has its maximum value at the position $z = 0$. One could think that when moving the barrier, e.g. to the left, the amplitude of the wave function changes and could even vanish if the barrier is moved by one quarter of the wave length. This misconception results from the fact that the wave function is a complex function which is difficult to draw in two dimensions. The cosine function which is drawn in Fig. 20.1 corresponds to the the real part of the complex wave function $\psi(z) = e^{ikz}$. The probability for the incoming wave is $|\psi(z)|^2 = 1$ is constant, i.e. independent of z.

In region II, inside the barrier $V = V_0$ and $E - V_0 < 0$. Therefore, k in (20.4) is imaginary. If we define κ by $k = i\kappa$, the real variable κ results as

$$\kappa = \sqrt{\frac{2m}{\hbar^2}(V_0 - E)}. \tag{20.6}$$

Thus the assumption $\psi = e^{ikz}$ for the solution of the Schrödinger equation leads to a wave function $\psi = e^{-\kappa z}$ if the energy of the particle E is less than the height of the potential barrier V_0. For region II, inside the barrier the solution ψ_{barrier} is given as

$$\psi_{\text{barrier}} = e^{-\kappa z}. \tag{20.7}$$

This is not an oscillating solution as found for the regions in which the potential vanishes, but an exponentially decaying (real) wave function which is generally found for regions in which the potential is larger than the particle energy. In classical physics, a particle would not enter such a potential region. In quantum mechanics, a particle can enter or even pass through such a classically forbidden region. This process is called "tunneling".

Up to now we have only considered the right-traveling wave $\psi = e^{ikz}$ for region I and III. Another independent solution is a left-traveling wave, $\psi = e^{-ikz}$. The general solution is a linear combination of the two. In region II, the general solution is a linear combination of (20.7) and $\psi_{\text{barrier}} = e^{+\kappa z}$.

In the following, we consider a barrier which is higher than the particle energy ($V_0 > E$) and obtain a solution of the time-independent Schrödinger equation, which is valid not only in one of the three regions, but for all z. Before we start, we should mention that any solution obtained above can be multiplied by a constant (complex number with amplitude and phase), as $c \cdot e^{i\alpha}$, and still remains a solution. Therefore, our solution of an incoming wave traveling from the left is combined from the solutions in the three regions as

$$\psi(z) = \begin{cases} Ae^{ikz} + Be^{-ikz} & z < 0 \ \text{(region I)} \\ Ce^{-\kappa z} + De^{\kappa z} & 0 \le z \le d \ \text{(region II)} \\ Fe^{ikz} & z > d \ \text{(region III)} \end{cases} \tag{20.8}$$

with k and κ as defined in (20.5) and (20.6), respectively, and B, C, D, and F being complex numbers. We assume the amplitude A of the incoming wave to be unity. In region I ($z < 0$) the incoming wave is accompanied by a certain amount B of a reflected wave. In region II, we have already interpreted the term proportional to C as the wave function tunneling into the barrier. The term proportional to D can be interpreted as a "reflection" from the downward potential jump at position d. In contrast to classical mechanics, in quantum mechanics the wave function can also be "reflected" from a downward potential step. In region III, the transmitted wave (with positive k vector) is proportional to F. Since we assume no incoming particle from the right, no wave with negative k vector is assumed in region III.

The coefficients B, C, D, and F can be calculated from certain continuity conditions which hold at the borders of the regions as will be discussed now. First we mention that the wave function is finite everywhere. This is the case due to the probability interpretation of the wave function. Since the probability of a particle being at a certain place $|\psi(z)|^2$ is finite, also the wave function itself has to be finite. If we now consider the time-independent Schrödinger equation (20.2) with a finite potential (as in our case of a barrier), we can conclude that the right-hand side of (20.2) is always finite. Therefore, $\frac{\partial^2}{\partial z^2}\psi(z)$ can be integrated and $\frac{\partial}{\partial z}\psi(z)$, as well. Thus $\psi(z)$ and $\frac{\partial}{\partial z}\psi(z)$ are continuous, as the integral over a finite function is always continuous.

Since the wave function and its derivative have to be continuous, we apply these continuity conditions at the two boundaries between the regions. At the position $z = 0$ the two equations

$$1 + B = C + D \quad \text{and} \quad ik\,(1 - B) = \kappa\,(D - C)\,, \tag{20.9}$$

for the wave function and its derivative, respectively. At the position $z = d$ the two equations are

$$Ce^{-\kappa d} + De^{\kappa d} = Fe^{ikd}, \quad \text{and} \quad \kappa\left(De^{\kappa d} - Ce^{-\kappa d}\right) = ikFe^{ikd}, \tag{20.10}$$

for the wave function and its derivative, respectively. These four (complex) equations, fix the four (complex) coefficients B, C, D, and F. The solutions can be found, for instance, using a computer algebra system. In the following, we are interested in the factor F, which is the amplitude of the wave function past the barrier. The absolute square of the coefficient F results as

$$T = |F|^2 = \frac{4k^2\kappa^2}{(k^2 + \kappa^2)^2 \sinh^2(\kappa d) + 4k^2\kappa^2}. \tag{20.11}$$

This factor T is the probability of finding an electron at the end of the potential barrier, $|\psi(d)|^2$. We will show in the next section that T corresponds to the flux (particle flux or electric current) through the potential barrier. For this reason T is also called the transmission factor.

In the limit $\kappa d \gg 1$ (20.11) can be simplified. In this case the $\sinh^2$ term can be approximated by

$$\sinh^2 \kappa d \approx \frac{1}{4} e^{2\kappa d}. \tag{20.12}$$

As another approximation, we also neglect the last term in the denominator (we will verify the validity of this approximations later) and the transmission factor becomes

$$T = \frac{16k^2\kappa^2}{(k^2 + \kappa^2)^2} e^{-2\kappa d} = \frac{16E(V_0 - E)}{V_0^2} \exp\left[-2d\sqrt{\frac{2m}{\hbar^2}(V_0 - E)}\right], \tag{20.13}$$

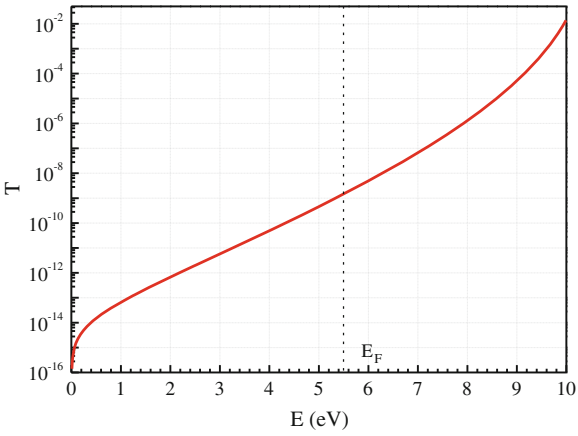

Fig. 20.2 Transmission factor as a function of the electron energy E for a square potential of barrier height $V_0 = 10\,\text{eV}$ and thickness $d = 0.5\,\text{nm}$ for the case that E is smaller than the barrier V_0. The exact solution (20.11) shown is indistinguishable from the approximation (20.13)

using (20.5) and (20.6). If we neglect the energy dependence of the preexponential factor, we see that the transmission factor and therefore also the measured tunneling current in STM experiments decreases exponentially with the barrier width d, which is the tip-sample distance. Secondly, the tunneling current decreases exponentially with the square root of $V_0 - E$. In the preexponential factor, E and V_0 enter separately, therefore also the depth of the potential, i.e. the zero level for the potential in Fig. 20.1, becomes important. Here we use $V_0 = 10\,\text{eV}$ and $d = 0.5\,\text{nm}$. In Fig. 20.2, the exact solution for the transmission factor (20.11) is plotted as a function of the energy E. The curve for the approximation is indistinguishable from the exact solution, which shows that the approximations made in order to obtain (20.13) are justified. The transmission factor still remains much smaller than unity up to $E = V_0$.

In a metal, E can be identified with the Fermi energy E_F, i.e. the energy of the highest occupied state, and V_0 corresponds to the vacuum energy E_{vac} relative to the bottom of the potential. The difference between vacuum energy and Fermi energy is called the work function $\Phi = E_{\text{vac}} - E_F = V_0 - E_F$. Thus E_F can be written as $E_F = V_0 - \Phi$, and for a usual value for the work function of $\Phi = 4.5\,\text{eV}$ the Fermi energy results as $E_F = 5.5\,\text{eV}$ from the bottom of the potential. As seen in Fig. 20.2, the transmission factor at E_F has very small values of about 10^{-9}.

In the following, we consider the transmission factor for energies larger than the barrier height. For $E > V_0$ the constant κ becomes imaginary according to (20.6) and we define $\kappa = ik'$. If we insert this into (20.11) and use the identity $\sinh(ik'd) = i\sin(k'd)$, the following expression for the transmission factor results

$$T = |F|^2 = \frac{4k^2k'^2}{(k^2 - k'^2)^2 \sin^2(k'd) + 4k^2k'^2}. \tag{20.14}$$

For the case $E > V_0$ the transmission factor has an oscillatory character approaching a transmission factor of unity for particular energies, as shown in Fig. 20.3.

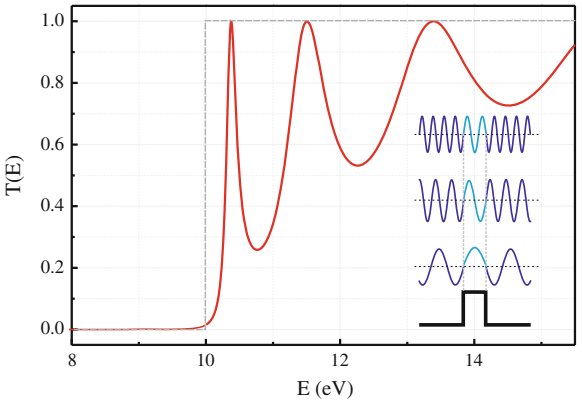

Fig. 20.3 Transmission factor as a function of the particle energy E for a square potential of barrier height $V_0 = 10\,\text{eV}$ and thickness $d = 0.5\,\text{nm}$. In classical mechanics, the transmission is zero if the particle energy is lower than the barrier and one if the energy is larger than the barrier (*dashed line*). For energies larger than the barrier the quantum mechanical solution of the problem results in an oscillatory behavior of the transmission factor according to (20.14), as shown by the *red line*. The transmission factor reaches unity, if an integer of the half-wavelength of the wave function in the barrier fits into the barrier width d, as shown in the *inset*

This occurs if the sine term in (20.14) vanishes, i.e. if the condition $k'd = n\pi$ is fulfilled. Since the wavelength of the wave function is related to the wave number by $k' = 2\pi/\lambda$, the following condition between the wavelength λ and the thickness of the barrier d results

$$n\frac{\lambda}{2} = d. \tag{20.15}$$

Thus the condition for a transmission factor of one is that an integer of half the wavelength of the wave function inside the barrier fits into the barrier width d, as shown in the inset of Fig. 20.3. These barrier resonances will be considered in Sect. 21.11.

While the one-dimensional wave function matching approach proved quite useful, it also involves several problems. First, the wave functions are not normalized. Since the wave functions extend to infinity the integral over $\psi\psi^*$ is infinite. Second, in this approach no voltage difference between the electrodes is considered. Also only one electron state at energy E is considered and the electronic structure of the sample (and the tip) does not enter in this approach.

20.2 Flux of Matter and Charge in Quantum Mechanics

In the previous section, we used the yet unproven ad hoc assumption that the absolute square of the wave function behind the barrier (relative to the incoming wave) is proportional to the flux density of the particles (or the electric current). In the following,

we will show that this results from a more general description of the flux of matter in quantum mechanics.

In electricity or fluid dynamics, a continuity equation relates a density ρ to a flux density, or current density $\boldsymbol{j}$ as

$$\frac{\partial \rho(\boldsymbol{r}, t)}{\partial t} + \operatorname{div} \boldsymbol{j}(\boldsymbol{r}, t) = 0. \tag{20.16}$$

In quantum mechanics, the probability density is defined as

$$\rho(\mathbf{r}, t) = \Psi(\mathbf{r}, t)\Psi^*(\mathbf{r}, t). \tag{20.17}$$

In searching for a flux density $\boldsymbol{j}(\boldsymbol{r}, t)$ which fulfills a continuum equation for the probability density given above, we take the first derivative of (20.17) with respect to time

$$\frac{\partial \left[\Psi(\mathbf{r}, t)\Psi^*(\mathbf{r}, t)\right]}{\partial t} = \frac{\partial \Psi}{\partial t}\Psi^* + \Psi \frac{\partial \Psi^*}{\partial t}. \tag{20.18}$$

Inserting the time-dependent Schrödinger equation (20.1) (slightly reordered)

$$\frac{\partial \Psi}{\partial t} = \frac{i\hbar}{2m}\Delta\Psi - \frac{i}{\hbar}V\Psi \tag{20.19}$$

into the right part of (20.18) results in

$$\frac{\partial(\Psi\Psi^*)}{\partial t} = \frac{i\hbar}{2m}\left(\Delta\Psi\Psi^* - \Psi\Delta\Psi^*\right). \tag{20.20}$$

The terms proportional to V cancel out. The bracket in (20.20) can be written as a divergence of another term. This can be easily seen by calculating the following divergence

$$\operatorname{div}\left(\Psi^*\nabla\Psi - \Psi\nabla\Psi^*\right) = \nabla\Psi^*\nabla\Psi + \Psi^*\Delta\Psi - \nabla\Psi\nabla\Psi^* - \Psi\Delta\Psi^*. \tag{20.21}$$

As the first and the third term cancel, the continuity equation (20.16) holds with the flux density (or probability current) $\boldsymbol{j}$, as

$$\boldsymbol{j}(\boldsymbol{r}, t) = \frac{-i\hbar}{2m}\left(\Psi^*\nabla\Psi - \Psi\nabla\Psi^*\right). \tag{20.22}$$

The continuity equation for the probability means that the probability of finding a particle is conserved in a local sense. Generally, the probability would also be conserved if a particle disappeared at one place and appeared at a distant place at the same time. Due to the validity of the continuity equation any change of the probability with time is related to an inward flowing probability current and the probability is conserved *locally*.

Now we calculate the probability current for the specific case of the wave function for a square barrier given by (20.8). We perform the calculation for the easiest case (i.e. region III). However, the result will be the same for the other regions. The wave function in region III is a one-dimensional plane wave given by $\Psi = Fe^{ikz}$. In calculating the probability current according to (20.22), in the one-dimensional case the Nabla operator is replaced by the partial derivative $\frac{\partial}{\partial z}$. With this the (scalar) probability current can be written as

$$j = \frac{-i\hbar}{2m}\left(F^* e^{-ikz}\frac{\partial F e^{ikz}}{\partial z} - F e^{ikz}\frac{\partial F^* e^{-ikz}}{\partial z}\right) = |F|^2\,\frac{\hbar k}{m}. \tag{20.23}$$

In the semi-classical transport theory of solid state physics [42], the particle velocity is written as

$$v = \frac{p}{m} = \frac{\hbar k}{m}. \tag{20.24}$$

Thus the probability current density is given by the absolute square of the amplitude of the wave function times the particle velocity as $j = |F|^2\, v$. One point to note is that in spite of the fact that the wave function is stationary, i.e. does not change with time, a flux density (probability current) occurs and the particle is "in motion".

There still remains the problem that the wave function in (20.8) is not normalized, because it extends to minus and plus infinity. In principle, one could multiply the probability current density of one particle by the number of particles and obtain a particle flux density, and by multiplication with the particle charge also an electric current density. Finally, if divided by the area, an electric current results.

20.3 The WKB Approximation for Tunneling

Only for a few types of potential barriers can the Schrödinger equation be solved analytically. For a more general type of one-dimensional potential barriers, the semi-classical Wentzel-Kramers-Brillouin (WKB) approximation is often applied. While this approximation does not describe some typical hallmarks of quantum behavior, such as the oscillatory behavior of the transmission factor, the exponential decay of the transmission factor in a potential barrier is reproduced.

This method can be considered as an extension of the solution of the square barrier to more general shapes of a one-dimensional barrier. When discussing the square barrier model, we found that the wave vector k is modified inside the barrier. For a free particle with energy E which moves in a constant potential V, the wave function of the time-independent Schrödinger equation can be written as

$$\psi(x) = Ce^{\pm ikz}, \tag{20.25}$$

with C being a normalization constant and $k = \frac{1}{\hbar}\sqrt{2m(E-V)}$. We notice that for different values of the potential V the phase ikz of the wave function is modified.

We now consider a spatially varying one-dimensional potential to be composed of small segments of constant potential. Infinitesimal rectangular barriers of barrier height $V(z)$ and thickness dz are considered. The wave function will be still approximated by a plane wave. The total phase shift can be calculated by integration over the infinitesimal phase shifts between z_0 and z_1. We obtain for the wave function at position z_1

$$\psi(z_1) = \psi(z_0) \exp\left[\pm i \int_{z_0}^{z_1} k(z) dz\right]. \tag{20.26}$$

For the case of a potential barrier $E - V(z) < 0$ and omitting the preexponential factor in (20.13), the WKB approximation the transmission factor results as

$$T = \left|\frac{\psi(z_1)}{\psi(z_0)}\right|^2 = \exp\left[-\frac{2\sqrt{2m}}{\hbar} \int_{z_0}^{z_1} \sqrt{(V(z) - E)} dz\right]. \tag{20.27}$$

For a constant value of the potential, the exponential dependence of the transmission factor which was found by the wave matching method is recovered. For a more general form of the potential, the integral along the barrier in (20.27) can be evaluated numerically for any barrier shape.

As an example of the WKB approximation, we consider a trapezoidal barrier as shown in Fig. 20.4. The potential as function of z is written as $V(z) = E + \Phi_{\text{tip}} - z/d(\Phi_{\text{tip}} - \Phi_{\text{sample}})$. The integral in (20.27) is written as

$$\int_0^d \sqrt{\Phi_{\text{tip}} - z/d(\Phi_{\text{tip}} - \Phi_{\text{sample}})} dz = \frac{2d}{3(\Phi_{\text{tip}} - \Phi_{\text{sample}})} \left[\Phi_{\text{tip}}^{3/2} - \Phi_{\text{sample}}^{3/2}\right]. \tag{20.28}$$

If this expression for the integral is inserted in (20.27) an analytic expression for the transmission through a trapezoidal barrier results in the WKB approximation.

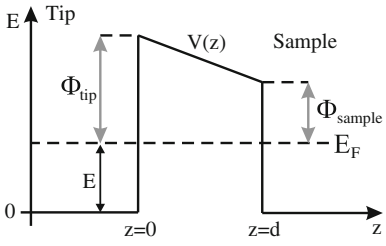

Fig. 20.4 Trapezoidal barrier arising due to different work functions Φ considered for tip and sample. An electron with energy E above the bottom of the potential is considered

20.4 Density of States

Since in the following sections the concept of the density of states will be used, we will introduce it here. The density of states $\rho(E)$ is a distribution function describing the number of electronic states of a system in the range between E and $E + dE$ and is defined for instance by the relation

$$dN(E, E + dE) = \rho(E)dE. \tag{20.29}$$

The number of states in a finite energy range between E_1 and E_2 is then given by

$$N(E_1, E_2) = \int_{E_1}^{E_2} \rho(E)dE. \tag{20.30}$$

For a system with n discrete states (where also all degenerate states need an individual number) the density of states can be formally written as

$$\rho(E) = \sum_n \delta(E - E_n), \tag{20.31}$$

where δ represents the Dirac delta function, which can be defined as a limit of the normalized Gauss distribution

$$\delta(x) = \lim_{a \to 0} \frac{1}{a\sqrt{2\pi}} e^{-\frac{x^2}{2a^2}}. \tag{20.32}$$

From this it can be inferred that the unit of the Dirac delta function $\delta(x)$ is $1/\text{unit(x)}$.

Like the Gauss distribution, the Dirac delta function is normalized

$$\int_{-\infty}^{\infty} \delta(E)dE = 1. \tag{20.33}$$

The following expression also holds for a finite range of integration

$$\int_{E_1}^{E_2} \delta(E - E_i)dE = 1, \tag{20.34}$$

if $E_i \in [E_1, E_2]$. Coming back to the definition of the density of states in (20.31) we calculate the number of states between E_1 and E_2 as in (20.30)

$$N(E_1, E_2) = \int_{E_1}^{E_2} \rho(E) dE = \int_{E_1}^{E_2} \sum_n \delta(E - E_n) dE = \sum_n 1, \tag{20.35}$$

where the sum extends over all n for which $E_n \in [E_1, E_2]$. This is the expected result and confirms that the definition of the DOS in (20.31) is reasonable. Up to now we have only focused on the energy dependence of the density of states. However, it is useful that the number of states is also referred to a certain volume as well. The number of states (like the one in (20.35)) is referred either to the whole system, or to a unit volume for periodic systems.

Another useful normalization is to consider the spatial distribution of the density of states. We do this by defining a new distribution function, the local density of states (LDOS), by weighting with the probability $|\psi_n(\boldsymbol{r})|^2$ of a particle to be at position $\boldsymbol{r}$ as

$$\text{LDOS} = \rho(E, \boldsymbol{r}) = \sum_n |\psi_n(\boldsymbol{r})|^2 \delta(E - E_n). \tag{20.36}$$

20.5 Bardeen Model for Tunneling

In a simple interpretation, the STM image is the surface topography of the sample. However, on the atomic scale it is not clear what would be meant by the word "topography". One reasonable definition would be that a topographic image is an image of constant surface charge density. However, as we will see in this section the STM tip follows the local density of states at the Fermi level, whereas electrons at all energies contribute to the charge density. In the following, we describe an interpretation of STM images, applicable even in the case of atomic resolution.

Bardeen developed a model for tunneling in solids long before the invention of the STM. He considered tunneling in metal-insulator-metal tunneling junctions. In the following, we transfer his model to the case of the STM. His approach was to consider the tip plus barrier and the sample plus barrier as two separate systems.[2] The electronic states of the individual subsystems may be obtained by solving the time-independent Schrödinger equations of the two subsystems. In the simplified one-dimensional case, the solutions are just oscillatory wave functions of different energies with an exponentially decaying tail inside the barrier. In the full three-dimensional case, these are the complete wave functions (and corresponding energy eigenvalues) of tip and sample taking the atomic arrangement into account.

Subsequently, the transition (scattering) from the initial (tip) states to the final (sample) states is considered within the time-dependent perturbation theory. An electron which is initially in a tip state can scatter (be transferred) into a sample state. The tip and sample states are indicated in Fig. 20.5 with a bias voltage V applied between

[2] Strictly speaking, he considered a metal electrode plus oxide barrier and another metal electrode, since STM had not been invented yet.

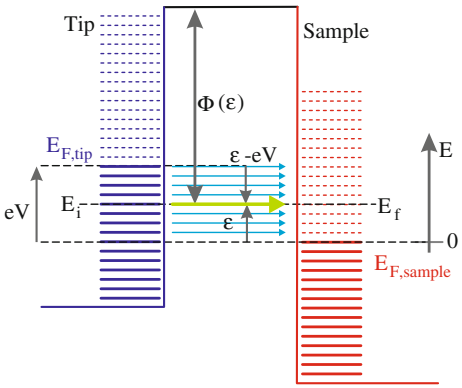

Fig. 20.5 Energy diagram of tip and sample states for the case of positive sample bias voltage V. Tunneling with energy conservation can only occur within the bias window (*blue arrows*). Above the bias window the initial states are empty and below the final states are occupied. All energies are referred relative to the sample Fermi level

tip and sample. When a positive voltage is applied to an electrode, the energy of the states in this electrode is decreased. In the reverse case, a negative voltage at an electrode leads to an upward shift of the energy levels (for electrons it is energetically unfavorable to hop onto a negatively charged electrode). Thus Fig. 20.5 corresponds to the case in which a positive voltage V is applied to the sample or equivalently a negative voltage is applied to the tip. Furthermore, we use the zero temperature limit in which all levels are filled up to the (tip or sample) Fermi level and are empty above.[3] This means that tunneling (or scattering from one electrode to another) can only occur in the bias window between $E_{F,\text{sample}}$ and $E_{F,\text{tip}}$.

In Bardeen's approach, the transition rate from one electrode to the other is calculated using the time-dependent perturbation theory assuming weak coupling between the two electrodes. Specifically, a variant of Fermi's golden rule is applied in order to calculate the transition rate. Since Fermi's golden rule is often not a part of the introductory courses in quantum mechanics, Bardeen's variant for tunneling is derived in Appendix B with emphasis on the case of scanning tunneling microscopy.

Applied to the case of tunneling, Fermi's golden rule shows that scattering from a particular (initial) tip state i at $E_{\text{tip},i}$ to a particular (final) sample state f at $E_{\text{sample},f}$ results according to (B.19) at a transition rate (number of electrons per time) of

$$w_{\text{tip},i\to\text{sample},f} = \frac{2\pi}{\hbar}\left|M_{fi}\right|^2 \delta(E_{\text{sample},f} - E_{\text{tip},i}), \tag{20.37}$$

with the matrix element M_{fi} calculated according to (B.24) as

$$M_{fi} = \frac{\hbar^2}{2m}\int\limits_{S_{\text{tip/sample}}} \left[\psi_{\text{tip},i}(\boldsymbol{r})\nabla\psi^*_{\text{sample},f}(\boldsymbol{r}) - \psi^*_{\text{sample},f}(\boldsymbol{r})\nabla\psi_{\text{tip},i}(\boldsymbol{r})\right]\cdot d\boldsymbol{S}. \tag{20.38}$$

[3] This arises because electrons are fermions and only one electron can occupy an electronic state due to the Pauli principle.

The Dirac delta function in (20.37) shows that the energy of the final state, must be the same as the energy of the initial state, as expected from energy conservation. The calculation of the matrix element involves an integration over an (arbitrary) tip-sample separation surface $S_{\text{tip/sample}}$.

In order to obtain the total rate for the transfer from any initial state to any final state a sum over all pairs of initial and final states has to be calculated[4]

$$w_{\text{tip}\to\text{sample}} = \frac{2\pi}{\hbar} \sum_{i,f} \left|M_{fi}\right|^2 \delta(E_f - E_i), \tag{20.39}$$

where in the low-temperature limit the sum has to be performed only within the bias window. In order to obtain the current, we multiply the transition rate $w_{\text{tip}\to\text{sample}}$ by the electron charge (we skip the negative sign here and consider only the absolute value of the current) and take a factor of two due to spin degeneracy into account. This results in

$$I = \frac{4\pi e}{\hbar} \sum_{i,f} \left|M_{fi}\right|^2 \delta(E_f - E_i), \tag{20.40}$$

The Bardeen equation derived above for the tunneling current is quite general and will be evaluated further in certain approximations. Generally, two kinds of approximations are used. In the Tersoff-Hamann approximation, the tunneling voltage is considered to be so small that the energy dependence of the matrix element as well as the energy dependence of the densities of states can be neglected. The emphasis is put on considering the spatial dependence of the surface wave functions realistically. Furthermore, the tip wave function is approximated by a spherical s-wave. In this case, the Tersoff-Hamann model results in a very simple relation between the surface wave functions and the tunneling current and can be used to simulate STM images for specific models of surface structures. We will discuss this approximation in detail in a later section of this chapter.

In the energy-dependent approximation of the Bardeen equation, complementary assumptions are made. The emphasis is put on the energy dependence of the density of states of tip and sample. The treatment of the surface wave functions is replaced by the approximate treatment of the tunneling barrier in a one-dimensional model with a rectangular shape of the tunneling barrier. This approximation is used in scanning tunneling spectroscopy in order to obtain information on the density of states of the sample from the energy-dependent measurement of the tunneling current and its derivative.

[4] For the sake of easier notation, we abbreviate $E_{\text{sample},f}$ by E_f and correspondingly for the tip $E_{\text{tip},i}$ by E_i.

20.5.1 Energy-Dependent Approximation of the Bardeen Model

Since the double sum over the tip and sample states in (20.40) is an abstract entity, this will be replaced in the energy-dependent approximation of the Bardeen model by the introduction of the (energy-dependent) density of states of tip and sample.

Since each wave function corresponds to a particular energy, the dependence of the matrix element on the wave functions can be replaced by a dependence on the energy as

$$M_{fi}(\psi^*_{\text{sample},f}(\boldsymbol{r}), \psi_{\text{tip},i}(\boldsymbol{r})) = M(E_f, E_i) = M(E_f), \tag{20.41}$$

which is evident for non-degenerate states (each wave function corresponds to a certain energy). For degenerate states, the energy dependent matrix element is the sum of the matrix elements of the degenerate states, because several states contribute to the transition. The last equality in (20.41) arises if the matrix element appears together with the delta function, e.g. in (20.40).

In order to introduce the densities of states, we use the the following identity for the Dirac delta function

$$\int_{-\infty}^{\infty} f(\varepsilon)\delta(\varepsilon - E_f)d\varepsilon = f(E_f), \tag{20.42}$$

and insert $f(\varepsilon) = |M(\varepsilon)|^2\,\delta(\varepsilon - E_i)$, which results in

$$\left|M(E_f)\right|^2 \delta(E_f - E_i) = \int_{-\infty}^{\infty} |M(\varepsilon)|^2\,\delta(\varepsilon - E_i)\delta(\varepsilon - E_f)d\varepsilon. \tag{20.43}$$

Inserting this into (20.39) and extending the integration only over the bias window (low-temperature approximation) results in

$$\begin{aligned} w_{\text{tip}\to\text{sample}} &= \frac{2\pi}{\hbar}\sum_{i,f}\int_{E_{F,\text{sample}}}^{E_{F,\text{tip}}} |M(\varepsilon)|^2\,\delta(\varepsilon - E_i)\delta(\varepsilon - E_f)d\varepsilon \\ &= \frac{2\pi}{\hbar}\int_{E_{F,\text{sample}}}^{E_{F,\text{tip}}} |M(\varepsilon)|^2 \sum_i \delta(\varepsilon - E_i)\sum_f \delta(\varepsilon - E_f)d\varepsilon. \end{aligned} \tag{20.44}$$

The sums over i and f can be replaced according to (20.31) by the density of states of tip and sample $\rho_{\text{tip/sample}}(\varepsilon)$. This results in the following expression for the transition rate as

$$w_{\text{tip}\to\text{sample}} = \frac{2\pi}{\hbar} \int_{E_{F,\text{sample}}}^{E_{F,\text{tip}}} |M(\varepsilon)|^2 \, \rho_{\text{tip}}(\varepsilon)\rho_{\text{sample}}(\varepsilon) d\varepsilon. \tag{20.45}$$

In order to obtain the current, we multiply the transition rate $w_{\text{tip}\to\text{sample}}$ by the electron charge and again take the factor of two due to spin degeneracy into account. This results in

$$I = 2e \; w_{\text{tip}\to\text{sample}} = \frac{4\pi e}{\hbar} \int_{E_{F,\text{sample}}}^{E_{F,\text{tip}}} \rho_{\text{tip}}(\varepsilon)\rho_{\text{sample}}(\varepsilon) \, |M(\varepsilon)|^2 \, d\varepsilon. \tag{20.46}$$

It is seen from the above equation that the two electrodes enter symmetrically. The tunneling current is a convolution of the states of the sample *and* the tip. This means that in order to obtain information about the surface the (electronic) structure of the tip must be known. However, in most practical experiments the density of states of the tip is unknown and we will show later under which approximations the tip properties can be taken out of the problem.

Now we choose the sample Fermi energy as a reference point for the energies: $E_{F,\text{sample}} = 0$. Thus the tip Fermi energy results as $E_{F,\text{tip}} = eV$ as also shown in Fig. 20.5. Inserting this choice of reference into the integration limits (20.46) can be written as

$$I = \frac{4\pi e}{\hbar} \int_{0}^{eV} \rho_{\text{tip}}(\varepsilon)\rho_{\text{sample}}(\varepsilon) \, |M(\varepsilon)|^2 \, d\varepsilon. \tag{20.47}$$

In this equation, the energy variables of the densities of states of tip and sample are both referred to a common energy reference namely $E_{F,\text{sample}}$, which was set to zero. What is inconvenient about this notation is that the density of the *tip* states $\rho_{\text{tip}}(\varepsilon)$ is referred to the Fermi energy of the *sample*. It is more convenient if the energy variable in the density of the tip states is relative to the "its own" (i.e. tip) Fermi energy (Fig. 20.5). This corresponds to a change in reference from $\rho_{\text{tip}}(\varepsilon)$ (relative to the sample Fermi level) to $\rho_{\text{tip}}(\varepsilon - eV)$ (relative to the tip, i.e. "its own" Fermi level). The energy term $\varepsilon - eV$ describes the distance between tip Fermi energy and the green arrow in Fig. 20.5. This results in the expression for the tunneling current

$$I = \frac{4\pi e}{\hbar} \int_{0}^{eV} \rho_{\text{tip}}(\varepsilon - eV)\rho_{\text{sample}}(\varepsilon) \, |M(\varepsilon)|^2 \, d\varepsilon, \tag{20.48}$$

and the matrix element

$$M(\varepsilon) = \frac{\hbar^2}{2m} \int\limits_{S_{\text{tip/sample}}} \left[\psi_{\text{tip}}(\boldsymbol{r}, \varepsilon)\nabla\psi^*_{\text{sample}}(\boldsymbol{r}, \varepsilon) - \psi^*_{\text{sample}}(\boldsymbol{r}, \varepsilon)\nabla\psi_{\text{tip}}(\boldsymbol{r}, \varepsilon)\right] \cdot d\boldsymbol{S}. \tag{20.49}$$

In this approximation of the Bardeen equation, the double sum over the initial and final states has been replaced by an energy integral over the densities of states of tip and sample.

In the following, we will work out this approximation further by approximating the Bardeen matrix element for the case of a one-dimensional rectangular barrier.

20.5.1.1 Bardeen Matrix Elements for a One-Dimensional Barrier

The matrix elements in the Bardeen equation for the tunneling current (20.48) can be written according to (20.49). These matrix elements can be calculated if the tip and sample wave functions are known on the separation surface $S_{\text{tip/sample}}$. According to (20.37), we only consider elastic tunneling, thus we consider only matrix elements for which $E_i = E_f = E$ and therefore $M_{fi} = M(E)$. In the following, we calculate the matrix elements explicitly for a simple model of a one-dimensional rectangular barrier. In the one-dimensional case (20.49) reduces to

$$M(E) = \frac{\hbar^2}{2m} \int\limits_{z=z_s} \left[\psi_{\text{tip}}(z, E)\frac{\partial\psi^*_{\text{sample}}(z, E)}{\partial z} - \psi^*_{\text{sample}}(z, E)\frac{\partial\psi_{\text{tip}}(z, E)}{\partial z}\right] dS. \tag{20.50}$$

The integration is performed at a separation surface $S_{\text{tip/sample}}$ located at constant height above the sample surface z_S. The tip and sample wave functions are shown schematically in Fig. 20.6. The wave functions $\psi_{\text{tip},i}$ and $\psi_{\text{sample},f}$ are the solutions of the individual tip and sample systems, as considered in Appendix B. Inside a one-dimensional rectangular barrier the wave functions decay exponentially as discussed previously and can be written as

$$\psi_{\text{tip}}(z) = \psi_{\text{tip}}(0)e^{-\kappa z}, \tag{20.51}$$

and

$$\psi_{\text{sample}}(z) = \psi_{\text{sample}}(d)e^{\kappa(z-d)}, \tag{20.52}$$

where the energy dependence enters via the decay constant $\kappa = \sqrt{2m(V_0 - E)}/\hbar$, corresponding to a state with energy E. Since we only consider elastic tunneling, the energy in the initial and final states is the same. The zero point on the z-axis is chosen at the left end of the barrier, as indicated in Fig. 20.6. At the position of the right

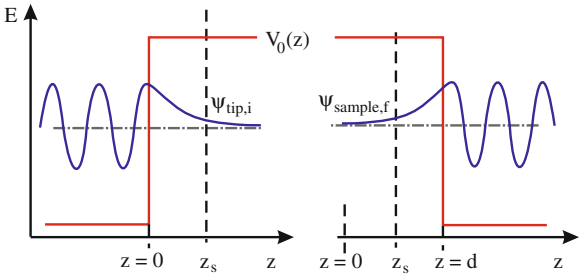

Fig. 20.6 Evaluation of the matrix element for a one-dimensional rectangular barrier. The wave functions of both separate tip and sample systems decay exponentially into the barrier. The matrix element is evaluated by integration at a position z_S inside the barrier

end of the barrier ($z = d$) the exponent in (20.52) is zero and for smaller z-values (i.e. positions inside the barrier) the exponent becomes negative, corresponding to an exponentially decaying final state wave function inside the barrier as shown in Fig. 20.6.

The matrix element can be evaluated by inserting the above one-dimensional wave functions at the position of the separation surface z_s. Inserting (20.51) and (20.52) into (20.50), we obtain

$$M(E) = \frac{\hbar^2}{2m} \int\limits_{z=z_s} 2\kappa \psi_{\text{tip},i}\,(0)\,\psi_{\text{sample},f}\,(d)\,e^{-\kappa z_s} e^{\kappa(z_s-d)}\,dS$$
$$= \frac{\hbar^2}{m} \kappa \psi_{\text{tip},i}\,(0)\,\psi_{\text{sample},f}\,(d)\,A e^{-\kappa d}. \tag{20.53}$$

The tunneling matrix elements are independent of the position z_s of the separation surface (as it should be), since the wave function expressions are taken at a specific constant z-position (namely zero and d), and κ does not depend on z. Moreover, all of the expressions in the integral do not depend on x or y, since we consider a one-dimensional model. Therefore, the integration results just in the area A of the electrodes (area of the tunneling contact). The energy dependence of $M(E)$ enters through the decay constant κ. While κ also appears in the preexponential factor, the energy dependence is dominated by the contribution of κ in the exponent. The term which enters into the equation for the tunneling current is $|M(E)|^2$ is called the transmission factor $T(\Phi, d) = |M(E)|^2$ in the one-dimensional approximation, with $\Phi = V_0 - E$. If we neglect the energy dependence of the pre-exponential factor, the energy dependence of the transmission factor can be expressed by the exponential factor as

$$T(\Phi, d) \propto \exp(-2\kappa d) = \exp\left(-2d\sqrt{\frac{2m}{\hbar^2}\Phi}\right). \tag{20.54}$$

In the one-dimensional approximation, the Bardeen model results in the same energy dependence of the transmission factor as that already obtained by the wave function matching model according to (20.13).

In summary, if in the evaluation of the matrix element in (20.48) a one-dimensional barrier is used for simplicity, the matrix element $|M(E)|^2$ can be approximated by a T transmission factor according to (20.54).

20.5.1.2 Barrier Height and Transmission Factor as Function of Voltage and Energy

In the previous section, we have evaluated the transmission factor for a given constant barrier height Φ of a rectangular barrier. In the following we replace this simple barrier by an effective barrier including effects such as different tip and sample work functions, the tunneling voltage, and the particular energy ε of a tunneling electron in the bias window.

The easiest case is to consider different work functions for tip and sample. As shown in Fig. 20.7a, this leads to a a trapezoidal barrier replacing the rectangular barrier (The other cases to be considered will also lead to a trapezoidal barrier).

As a general approach, we approximate the height of a trapezoidal barrier by a rectangular barrier with the average height of the trapezoidal barrier, i.e. at the middle of the barrier (dotted line in Fig. 20.7a). For the case of the different tip and sample work functions, the average barrier height is given by an average work function as $\overline{\Phi} = (\Phi_{\text{tip}} + \Phi_{\text{sample}})/2$.

This approximation is not very well justified since the transmission factor increases exponentially (not linearly) with decreasing barrier height and thus this approximation underestimates the tunneling current. Nevertheless, this approximation is generally used because of its simplicity and because the barrier thickness in STM experiments is quite small (0.5–1 nm).

As a next effect we include the dependence of the barrier height on the tunneling voltage V. Therefore, we assume for the moment the same work function for tip and sample. For states at the sample Fermi energy, the average tunneling barrier increases

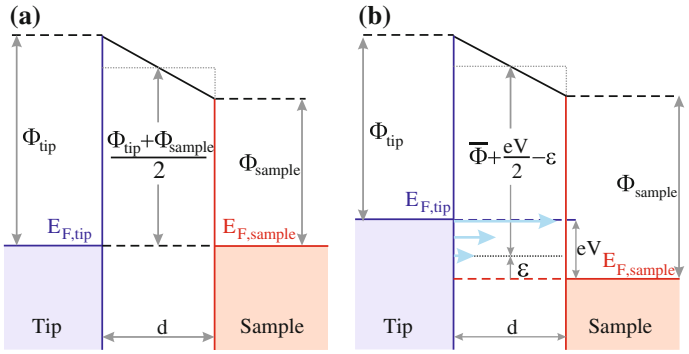

Fig. 20.7 **a** Average tunneling barrier if the work functions of tip and sample are different. **b** Dependence of the barrier height on the tunneling voltage (with tip and sample work function being the same)

to $\overline{\Phi} + \frac{eV}{2}$. In the general case for states at energy ε relative to the sample Fermi level the tunneling barrier height is given by $\overline{\Phi} + \frac{eV}{2} - \varepsilon$, as shown in Fig. 20.7b. For the two limiting cases $\varepsilon = 0$ and $\varepsilon = eV$ the barrier heights result as $\overline{\Phi} + \frac{eV}{2}$ and $\overline{\Phi} - \frac{eV}{2}$, respectively. This corresponds to an increase/reduction of the barrier height for electrons at the sample/tip Fermi level, respectively. If electrons in the middle of the bias window are considered $\varepsilon = \frac{eV}{2}$, the barrier results as $\overline{\Phi}$, i.e. independent of V.

Taking all the effects together, the average or effective barrier for an electron tunneling at energy ε relative to the sample Fermi level is

$$\Phi_{\text{eff}} = \frac{\Phi_{\text{tip}} + \Phi_{\text{sample}}}{2} + \frac{eV}{2} - \varepsilon. \tag{20.55}$$

If we now replace Φ in the equation for the transmission factor (20.54) by the effective barrier height given in (20.55), the following equation results for the transmission factor

$$T(\varepsilon, V, d) \propto \exp\left[-2d\sqrt{\frac{2m}{\hbar^2}\left(\frac{\Phi_{\text{tip}} + \Phi_{\text{sample}}}{2} + \frac{eV}{2} - \varepsilon\right)}\right]. \tag{20.56}$$

The transmission factor decreases exponentially for lower electron energies (smaller ε) because the effective barrier appears higher to these electrons as indicated by the horizontal arrows with different lengths shown in Fig. 20.7.

Since in the one-dimensional barrier approximation the matrix element $|M(\varepsilon)|^2$ is replaced by the transmission factor $T(\varepsilon, V, d)$, (20.48) can be written as

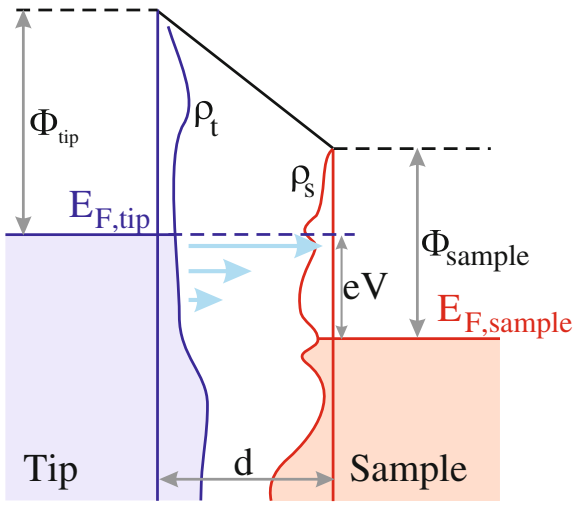

Fig. 20.8 Energy level diagram of the tunneling junction. The applied bias shifts the Fermi level by eV. Density of states are represented by ρ_{tip} and ρ_{sample} (the filled states are colored)

$$I = \frac{4\pi e}{\hbar} \int_0^{eV} \rho_{\text{tip}}(\varepsilon - eV)\rho_{\text{sample}}(\varepsilon)T(\varepsilon, V, d)d\varepsilon, \tag{20.57}$$

This is the Bardeen equation for a one-dimensional barrier in the limit of zero temperature (taking the Fermi functions as step functions). Since this is an important equation we will discuss it in the following using the graphic in Fig. 20.8. The appearance of the combined density of states $\rho_{\text{tip}} \cdot \rho_{\text{sample}}$ in (20.57) can be seen in the graphic representation of the energy diagram of the tunneling junction. Figure 20.8 includes the density of states of tip and sample, ρ_{tip}, and ρ_{sample}, respectively. When a positive bias voltage is applied, a tunneling current flows from the occupied states of the tip to the unoccupied states of the sample. The unoccupied states in the sample are shifted downwards by eV relative to the Fermi level of the tip. It is reasonable to assume that the contribution to the tunneling current at a certain energy is proportional to the density of occupied states in the tip as well as to the density of the unoccupied states in the sample. Therefore, this contribution to the tunneling current should be proportional to the product of both, as

$$dI \propto \rho_{\text{sample}}\ \rho_{\text{tip}} d\varepsilon. \tag{20.58}$$

This product has to be multiplied by the transmission factor $T(\varepsilon, V, d)$, which corresponds to the horizontal arrows in Fig. 20.8, and integrated over the range of energies in which occupied states in the tip and empty states in the sample exist. The horizontal arrows indicate that the tunneling is energy-conserving (elastic). The transmission factor decreases exponentially for lower energies, indicated by the shorter blue arrows. A larger effective barrier for the tunneling electrons applies to these low-lying states.

As we have seen, the transmission factor is not a constant but depends on the applied bias voltage and the energy of the tunneling electrons. However, for small tunneling voltages $eV \ll \Phi$ the energy-dependent term of the transmission factor ε can be replaced by an average energy $\bar{\varepsilon} = (E_{F,\text{tip}} + E_{F,\text{sample}})/2 = eV/2$. In this approximation, the transmission factor $T(d)$ is independent of the energy ε and the voltage V. Therefore, the tunneling current can be written as

$$I = \frac{4\pi e}{\hbar} T(d) \int_0^{eV} \rho_{\text{tip}}(\varepsilon - eV)\rho_{\text{sample}}(\varepsilon)d\varepsilon. \tag{20.59}$$

In this case, the tunneling current is proportional to the combined density of the states of tip and sample integrated up to the bias voltage. For a constant tip density of states, this can be put in front of the integral. If the sample density of states is (approximately) constant the integration leads to a proportionality between bias voltage V and the tunneling current ($I \propto V$). If the voltage is very small the result of the the tunneling

current is proportional to the sample density of states at the energy E_F. This also results from the Tersoff-Hamann approximation which we will discuss later.

20.5.1.3 Inclusion of the Fermi Functions at Finite Temperatures

Up to now we have considered the low-temperature case, i.e. all levels below the Fermi level have been treated as completely filled and all levels above the Fermi level have been considered as completely empty. In this approximation tunneling occurs only in the bias window. Considering Fermi functions, the Fermi functions were approximated as step functions. Now we take into account the actual Fermi distribution for the filling of the levels at energy E, which is not a step function, but a broadened (smeared out) step function. The Fermi-Dirac distribution $f(E - E_F)$ is defined as

$$f(E - E_F) = \frac{1}{1 + \exp\left[(E - E_F)/k_B T\right]}, \tag{20.60}$$

and gives the occupation number of filled states at energy $E - E_F$ and temperature T.[5] The occupation number of the empty states is correspondingly $1 - f(E - E_F)$.The contribution to the tunneling current due to the occupation of levels can be expressed as: the occupation number in the electrode from which the tunneling starts *times* the occupation number of the empty levels in the electrode to which the electron tunnels. The part $I^+(E)$ arises for electrons tunneling at a certain energy E which tunnel from the occupied states of the tip to the unoccupied states of the sample. There is also a part $I^-(E)$ for electrons tunneling from sample to tip (Fig. 20.9). $I^+(E)$ can be written as

$$I^+(E) = I_{\text{tip,filled}\rightarrow\text{sample,empty}} \propto f(E - E_{F,\text{tip}}) \cdot [1 - f(E - E_{F,\text{sample}})]. \tag{20.61}$$

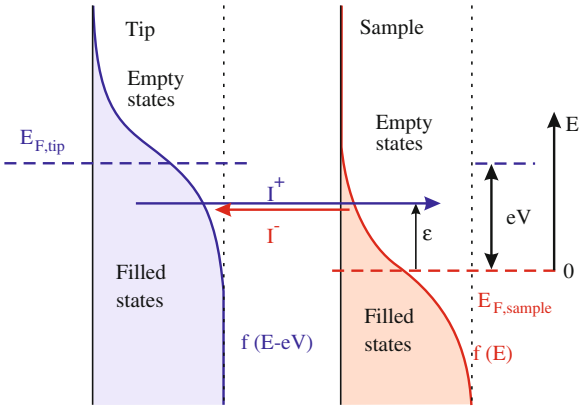

Fig. 20.9 Occupation numbers for tip and sample states. The total tunneling current is composed of a part I^+ for electrons tunneling from tip to sample and a part I^- for electrons tunneling from sample to tip

[5] Here we explicitly use $E - E_F$ as the argument, since we will refer the Fermi function to different Fermi levels, i.e. tip and sample Fermi level.

The reverse current contribution from the filled states of the sample to the empty states of the tip is written as

$$I^{-}(E) = I_{\text{sample,filled}\rightarrow\text{tip,empty}} \propto f(E - E_{F,\text{sample}}) \cdot [1 - f(E - E_{F,\text{tip}})] \quad (20.62)$$

so that the total contribution due to the occupation numbers, $I^{\text{total}}(E)$, is obtained as

$$I^{\text{total}}(E) = I^{+} - I^{-} \propto f(E - E_{F,\text{tip}}) - f(E - E_{F,\text{sample}}). \quad (20.63)$$

Since $E_{F,\text{sample}} = 0$ and $E_{F,\text{tip}} = eV$, the total contribution can be written as

$$I^{\text{total}}(E) = I^{+} - I^{-} \propto f(E - eV) - f(E). \quad (20.64)$$

The tunneling current is then obtained by including this factor in (20.57), renaming E to the integration variable ε, and extending the integration from minus infinity to infinity, as

$$I = \frac{4\pi e}{\hbar} \int_{-\infty}^{\infty} \{f(\varepsilon - eV) - f(\varepsilon)\} \rho_{\text{tip}}(\varepsilon - eV) \rho_{\text{sample}}(\varepsilon) T(\varepsilon, V, d) d\varepsilon. \quad (20.65)$$

The term in the curly brackets is also sometimes called the window function, because in the limit of low-temperatures, where the Fermi function becomes a step function, this term has the value one in the range between zero and eV, and zero everywhere else.[6]

20.5.2 Tersoff-Hamann Approximation of the Bardeen Model

The Bardeen tunneling model was developed before the invention of the STM in order to describe planar MIM (metal insulator metal) tunneling junctions. Shortly after the invention of the STM, Tersoff and Hamann adapted the Bardeen model for the case of the STM. Their theory of STM applies in the limit of very small tunneling voltages and was the first approximation including realistic surface wave functions (not the simple one-dimensional barrier).

As mentioned previously, the main challenge in the Bardeen theory is the calculation of the tunneling matrix elements. Tersoff and Hamann neglected the energy dependence of the matrix element and evaluated the matrix element in the limit of small voltages, i.e. at the Fermi level. They chose a plane above the surface and performed the integration in (20.49). In order to perform the calculation of the tunneling matrix element, explicit wave functions for the two electrodes (surface and tip) have

[6] Here we considered that tip and sample have the same temperature. If this is not the case, Fermi functions with different temperatures have to be considered for tip and sample.

to be inserted into (20.49). To describe the wave function of the surface they used a plane wave Fourier expansion. The structure of the surface is given by the Fourier components in the expansion.

To calculate the integral also the wave function of the tip has to be known. Unfortunately, the structure of the tip is usually not known. Tersoff and Hamann therefore assumed the simplest possible approximation for the tip. If the ideal STM tip consisted of a mathematical point source all tip properties would be taken out of the problem. They showed that, if the position of the tip point source is at $\boldsymbol{r}_t$, the current at small voltages reduces to

$$I \propto \sum_n |\psi_n(\boldsymbol{r}_t)|^2 \delta(E_n - E_F). \tag{20.66}$$

According to (20.36) this expression is the definition for the local density of states (LDOS). The Dirac delta function ensures the energy conservation (elastic tunneling). ψ_n are the surface wave functions and the term $|\psi_n(\boldsymbol{r}_t)|^2$ describes the probability of finding a surface state electron at the position of the tip. The tip probes the surface wave functions at the position $\boldsymbol{r}_t$.

Thus the expression for the tunneling current in (20.66) can be identified with the local density of states (LDOS) of the sample states at energy E_F and at the position of the point like tip $\boldsymbol{r}_t$ as

$$I \propto \sum_n |\psi_n(\boldsymbol{r}_t)|^2 \, \delta(E_F - E_n) \equiv \rho_{\text{sample}}(E_F, \boldsymbol{r}_t). \tag{20.67}$$

The ideal STM with the tip considered as a point source simply measures $\rho_{\text{sample}}(E_F, \boldsymbol{r}_t)$. Within this approximation, STM has a quite simple interpretation as measuring a property of the surface, without reference to the complex tip-sample system. In an STM measurement, the LDOS of the surface alone is measured (i.e. without an influence due to the tip), but at the position at which the tip is located.

Tersoff and Hamann have shown that the above equation remains valid, regardless of the size of the tip as long as the tip wave function can be approximated by an s-wave, i.e. a spherical wave function. The tip position $\boldsymbol{r}_t$ must then be interpreted as the effective center of curvature of the tip (Fig. 20.10), i.e. the origin of the s-wave, which best approximates the tip wave functions. This means that the STM measures the LDOS of the surface at the Fermi level at a distance z_0 above the surface (not d_{min}).

In order to interpret STM images quantitatively an STM image is calculated for a proposed structure and compared with the measured images. The Tersoff-Hamann approximation is one of the most widely used methods to interpret images based on ab initio calculations.

We can compare the large voltage approximation to the Tersoff-Hamann approximation by considering the limit of very small voltages in (20.57). In this limit, the current is proportional to the voltage and to $\rho_{\text{sample}}(\varepsilon = 0) T(\varepsilon = 0, V = 0, d)$. If we identify $d = z_0$ the following correspondence results

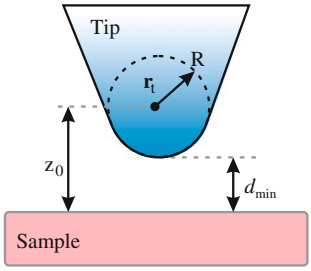

Fig. 20.10 In the Tersoff-Hamann approximation of the STM the tip is modeled as a locally spherical potential with a radius of curvature R centered at $\boldsymbol{r}_t$

$$\rho_{\text{sample}}^{\text{Ters.-Ham.}}(E_F, z_0) = LDOS(E_F, z_0) = \rho_{\text{sample}}^{\text{large volt. appr.}}(\varepsilon = 0) \cdot T(\varepsilon = 0, V = 0, d = z_0). \tag{20.68}$$

This means that the density of states of the sample at the very surface position times the transmission factor for a barrier of thickness d corresponds to the density of states at the position of the center of the tip a distance d outside the surface, also called the local density of sample states at the tip position.

20.6 Constant Current Mode and Constant Height Mode

By far the most common mode of STM operation is the constant current mode. In this mode the tunneling current is kept constant by adjusting an appropriate height of the tip above the sample surface via feedback. This mode is visualized in Fig. 20.11. The sample local density of states (LDOS) at E_F (absolute square of the wave function at the Fermi energy) is shown to have an oscillatory behavior along a coordinate x parallel to the surface at the surface positions, i.e. $z = 0$. We assume here that this modulation in the local density of states arises due to an atomic surface structure with

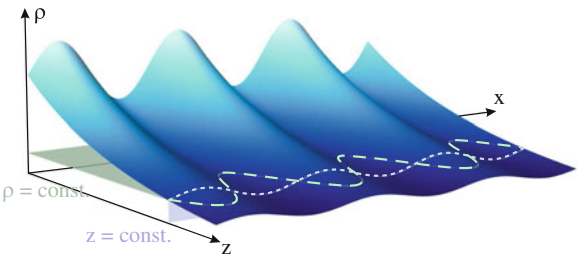

Fig. 20.11 Local sample density of states at E_F with an oscillatory modulation assumed to arise due to an atomic structure at the surface $z = 0$. This LDOS modulation is periodic along the surface (x-direction) and decays exponentially with increasing distance from the surface (z-direction). The contour of constant density of states is visualized in the constant current mode and shown as a *dashed line*. The constant height mode is visualized as the variation of the density of states at a constant distance from the surface ($z = \text{const.}$)

a high density of states at positions where an atom is present and a lower density of states at positions in between the atoms. As seen in the previous sections, the wave functions (and with them the density of states) decay exponentially with the distance from the surface, as they enter the barrier region outside the solid $z > 0$. A z dependent contour of a constant density of states (LDOS) is shown in Fig. 20.11 as a dashed line, extending further from the surface at x-positions at which an atom is present, and more closer to the surface at x-positions in between the atoms. In this way, the contour of constant density of states images the topography of the atomic structure. Below we will discuss that there are many exceptions in which this simplified picture is not valid.

Another mode in STM operation is the constant height mode. In this mode, the feedback is switched off and the tip is scanned at constant height over the surface, while recording the tunneling current. This mode is visualized in Fig. 20.11 by different values for the density of states at a constant distance from the surface ($z = const.$). Since in the Tersoff-Hamann approximation the tunneling current is proportional to the (local) sample density of states at the Fermi energy, this quantity is measured in the constant height mode. While this mode of operation has conceptually a simple interpretation, there are in practice several obstacles to the implementation of this mode. To maintain a constant height during a scan over an atomically flat terrace is very difficult due to effects of thermal drift and piezo creep (cf. Sect. 3.6). For these reasons the constant height mode is mostly applied in low-temperature STM experiments, where also the thermal drift and piezo creep are negligible. Furthermore, a scan over an atomic step edge will change the tip-sample distance by several Å. This will lead to a change in the tunneling current of several orders of magnitude. Two atomic steps come already close to a typical tip-sample distance and can lead to an undesirable tip-sample contact. Moreover, also the tilt between sample and scanner which is always present (cf. Chap. 7) makes the practical implementation of the constant height mode difficult. For these reasons, the constant current mode is usually confined to very small scans on an atomically flat terrace. A practical advantage of the constant height mode is that it is a fast mode of STM data acquisition only limited by the bandwidth of the current amplifier and not by the (lower) bandwidth of a feedback loop.

The two modes can also be combined, for instance to allow for fast scanning. In this case, the average height of the tip above the surface is followed (relatively slowly) using the feedback in the constant current mode. Since the feedback follows the surface topography only slowly, variations in the tunneling current remain on shorter time scales which cannot be compensated by the slow feedback. These tunneling current variations on a shorter time scale (error signal) correspond to a constant height mode and usually contain information on the atomic structure, or more generally information at small time and length scales (smaller than the constant current feedback can follow).

We would not like to conclude this section without mentioning that several effects can alter the simplified picture of interpreting contours of constant tunneling current as contours of the "topography of the surface (atoms)". For instance, atoms of different chemical elements, located with their nuclei at the same height above the surface,

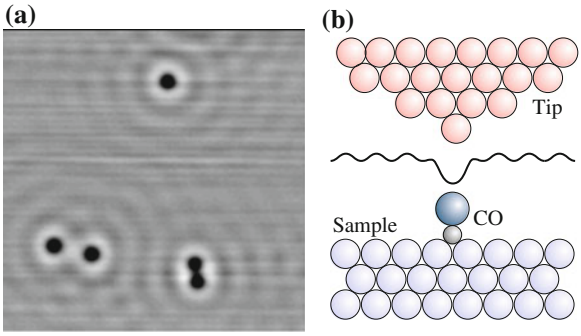

Fig. 20.12 **a** STM images of CO molecules on a Cu(111) surface. In spite of the fact that the CO molecules are adsorbed on top of the Cu atoms at the surface, they are imaged as depressions (*dark contrast*). This is an example of an electronic effect of chemical nature. **b** Schematic *side view* of the metal sample with adsorbed CO molecule and tip. The *black line* shows the contour followed by the tip in the constant current mode. Due to the low local density of states (LDOS) above the CO, the molecule is imaged as a depression

will give rise to different values of the density of states due to their different chemical nature. This will give rise to a different apparent (topographic) height measured in STM.

An extreme example in which a complete contrast inversion is found will be discussed in the following. In this example, a carbon monoxide molecule sticking out of the surface (as known from other experimental techniques than STM) is imaged in the constant current mode as a depression, as shown in Fig. 20.12. This results due to a reduced density of states compared to the bare metal surface. Due to this chemical effect and other electronic effects the simplified interpretation of the constant current contour measured in STM as "the topography" of the surface is often not applicable.

20.7 Voltage-Dependent Imaging

Semiconductors show a strong variation of the LDOS with energy; as an example we consider here GaAs. Due to the different chemical nature of Ga and As, electronic charge is transferred from Ga to As, giving the covalent bond a somewhat ionic character. This leads to occupied states at the As atoms (somewhat below the valence band edge) and to unoccupied states located at the Ga atoms (somewhat above the conduction band edge). For positive voltages, electrons tunnel from filled states of the tip into the empty conduction band states of the sample, located at the Ga atoms (Fig. 20.13a). With negative sample voltages, electrons tunnel out of the occupied valence band states located at the As atoms (Fig. 20.13b). Thus, using voltage-dependent imaging at different voltage polarities, empty states and filled states can be imaged at a semiconductor surface.

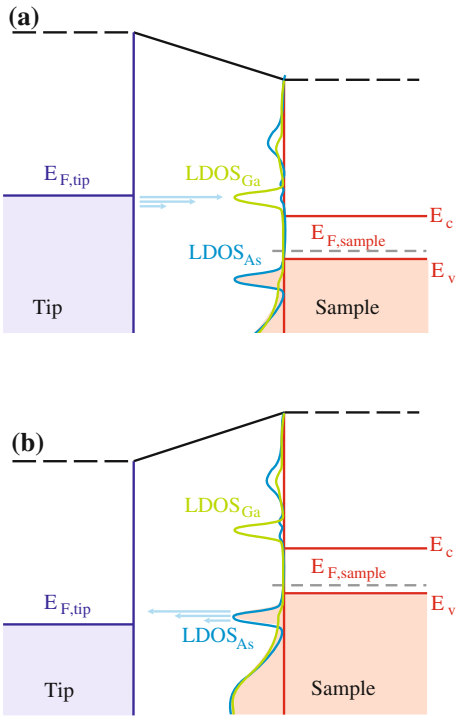

Fig. 20.13 Energy levels involved in tunneling at a semiconductor surface, here gallium arsenide (GaAs) is used as an example. **a** At positive sample bias, current flows from occupied tip states to unoccupied Ga sample states. **b** At negative sample bias, current flows from occupied As sample states to unoccupied tip states

As an example, we consider here the GaAs(110) surface which consists of zigzag chains of atoms alternating between Ga and As, as shown in Fig. 20.14c. Voltage-dependent imaging is experimentally realized by scanning the same area of a sample with two different voltages. Each scan line is recorded twice, first scanning with one voltage and then using another voltage. This interlacing technique results in two images recorded at two different tunneling voltages. Two such images recorded at sample voltages of $+1.9$ and $-1.9\,\mathrm{V}$ are shown in Fig. 20.14a, b, respectively. In order to highlight the differences between the two images a black rectangle is drawn in both images at the same position and in Fig. 20.14c as well. According to the previous discussion, Ga atoms (states) should be imaged in Fig. 20.14a and As atoms in Fig. 20.14b. Indeed the atomic protrusions in both images are arranged as shown in the top view of the model of the GaAs surface which is shown in Fig. 20.14c. The color image Fig. 20.14d shows an overlay of both images with Ga atoms in green and As atoms in red. This example shows that voltage-dependent imaging can (in fortunate cases) lead to element specific imaging in STM by imaging empty and occupied states at different bias voltage polarities. However, it has to be

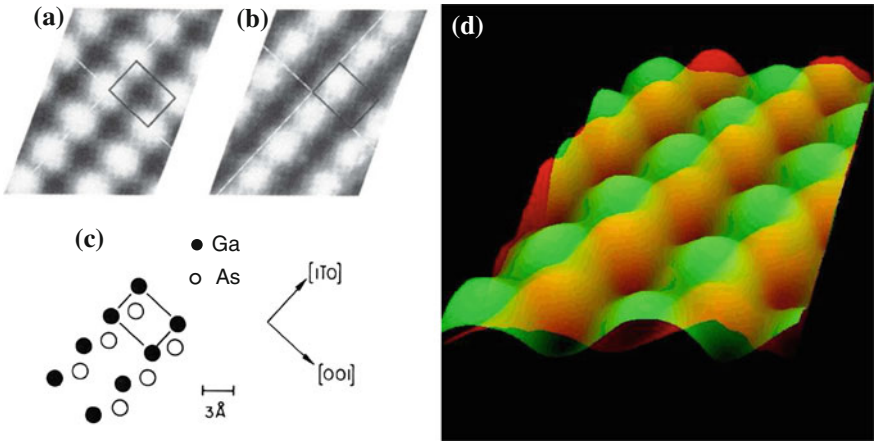

Fig. 20.14 STM images of a GaAs(110) surface acquired at sample voltages +1.9 V (**a**) and −1.9 V (**b**) (reproduced with permission from [43]). The empty states image in (**a**) shows the Ga atoms, while the filled states image in (**b**) shows the As atoms. **c** Schematic *top view* of the GaAs(110) surface structure. The Ga and As atoms are shown as *solid* and *open circles*, respectively. **d** Overlay of the filled and empty states images showing the Ga atoms in *green* and As atoms in *red* (reproduced with permission from R.M. Feenstra)

mentioned that this imaging with chemical sensitivity is generally more an exception than the rule. For instance, silicon and germanium cannot be distinguished by voltage-dependent imaging because their chemical nature is too similar.

20.8 Summary

- The problem of tunneling through a barrier can be treated in different approximations. In the simplest one-dimensional model, an incoming wave decays exponentially inside a barrier. The conditions of wave function matching of the incoming wave, the exponentially decaying wave and the transmitted wave at both ends of the barrier allow the transmitted amplitude to be calculated. In the limit that the particle energy is much smaller than the vacuum level the transmission coefficient depends exponentially on the barrier thickness and the square root of the barrier height, as

$$T \propto \exp\left(-const.\, d\sqrt{\Phi}\right). \tag{20.69}$$

- In the Bardeen model, the tunneling current can be written in the low-temperature limit as

$$I = \frac{4\pi e}{\hbar} \sum_{i,f} \left|M_{fi}\right|^2 \delta(E_f - E_i), \tag{20.70}$$

with the following expression for the matrix element

$$M_{fi} = \frac{\hbar^2}{2m} \int\limits_{S_{\text{tip/sample}}} \left[\psi_{\text{tip},i}(\boldsymbol{r}) \nabla \psi^*_{\text{sample},f}(\boldsymbol{r}) - \psi^*_{\text{sample},f}(\boldsymbol{r}) \nabla \psi_{\text{tip},i}(\boldsymbol{r}) \right] \cdot d\boldsymbol{S}. \tag{20.71}$$

with the integration extending over an arbitrary separation surface inside the barrier. The Bardeen equation is quite general and can take in principle the three-dimensional character of the problem into account. Two approximations are considered in the following.

- In the energy-dependent approximation of the Bardeen model, the double sum over initial and final states is replaced by integrating the densities of states of tip and sample over the energy. This approximation is most suitable to consider the energy dependence of the tunneling current. In this approximation, the current can be written depending on the combined density of states $\rho_{\text{tip}} \cdot \rho_{\text{sample}}$ as

$$I = \frac{4\pi e}{\hbar} \int\limits_0^{eV} \rho_{\text{tip}}(\varepsilon - eV) \rho_{\text{sample}}(\varepsilon) \, |M(\varepsilon)|^2 \, d\varepsilon, \tag{20.72}$$

with the energy variable of the tip and sample density of states referred relative to the respective Fermi levels. The matrix element is considered as a function of the energy. In this energy-dependent approximation, the matrix element is further evaluated using a one-dimensional barrier model, resulting in a transmission factor replacing the matrix element as

$$|M(\varepsilon)|^2 = T(\varepsilon, V, d) \propto \exp\left(-2d \sqrt{\frac{2m}{\hbar^2} \left(\frac{\Phi_{\text{tip}} + \Phi_{\text{sample}}}{2} + \frac{eV}{2} - \varepsilon \right)} \right). \tag{20.73}$$

- If at finite temperatures the Fermi functions are no longer considered as step functions, also a small "reverse current" from the filled sample states to the empty tip states occurs for positive sample bias voltages. This leads to

$$I = \frac{4\pi e}{\hbar} \int\limits_{-\infty}^{\infty} \{ f(\varepsilon - eV) - f(\varepsilon) \} \, \rho_{\text{tip}}(\varepsilon - eV) \rho_{\text{sample}}(\varepsilon) T(\varepsilon, V, d) d\varepsilon. \tag{20.74}$$

- The Tersoff-Hamann approximation is an evaluation of the Bardeen model in the limit of small tunneling voltages and the limit of a spherical tip. In this limit, the tunneling current is proportional to the density of states of the surface at the Fermi energy at the position of the center of the tip, i.e. the local density of states (LDOS).

- In the constant current mode, the tunneling current is kept constant by adjusting an appropriate height of the tip above the sample surface via feedback. In the simplest approximation in this mode, the tip follows the topography of the surface. In the constant height mode, the tunneling current is recorded during scanning without feedback. This mode can only be used on atomically flat surfaces.
- Voltage-dependent imaging of semiconductor surfaces allows the empty states of the sample to be imaged at positive sample bias voltages and the filled states for the reverse polarity. In the case of GaAs, voltage-dependent imaging allows to perform chemically sensitive imaging of As and Ga atoms separately.

Chapter 21
Scanning Tunneling Spectroscopy (STS)

One of the fascinating potentials of scanning tunneling microscopy is its ability to obtain energy-resolved spectroscopic data with atomic resolution. As we will see in this chapter, the STM allows us to measure directly the spatial and energy dependence of the local density of states. In the last chapter, we saw that apart from the influence of the thickness and the height of the barrier the tunneling current also depends on the applied bias voltage. In this chapter, we will focus further on the voltage dependence of the tunneling current. For very low bias voltages, when the matrix element or transmission factor as well as the densities of states can be considered independent of the bias voltage, the current is proportional to the applied voltage (Tersoff-Hamann approximation). For higher bias voltages and specifically for semiconductor samples, the bias dependence of the density of states and the transmission coefficient cannot be omitted.

21.1 Scanning Tunneling Spectroscopy—Overview

In the previous chapter the following expression was obtained for the tunneling current (20.48)

$$I = \frac{4\pi e}{\hbar} \int_0^{eV} \rho_{\mathrm{tip}}(\varepsilon - eV)\rho_{\mathrm{sample}}(\varepsilon)T(\varepsilon, V, d)d\varepsilon, \tag{21.1}$$

with ρ_{tip} and ρ_{sample} being the density of states of tip and sample, with their energy variables referred relative to their respective Fermi level. The term $T(\varepsilon, V, d)$ corresponds to the transmission factor for tunneling from a tip state to a sample state, whose energy dependence is referred relative to the sample Fermi level. In scanning tunneling spectroscopy (STS), the aim is to measure the density of states of the sample. This is accomplished by measuring the current-to-voltage characteristic of the tunneling junction. From Fig. 20.8, we can see that a small increase of the voltage

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_21

dV shifts all sample states down and an additional current dI contributes to the total current I. Graphically, this corresponds to a new blue arrow appearing in Fig. 20.8. In a first approximation this additional current dI per voltage increase $dV = d\varepsilon/e$ is given by the integrand in (21.1) taken at the upper limit of the integral $\varepsilon = eV$ as

$$\frac{dI}{dV} \approx \frac{4\pi e^2}{\hbar} \rho_{\text{tip}}(0)\, \rho_{\text{sample}}(eV)\, T(eV, V, d). \tag{21.2}$$

This is only a first approximation realizing the principle of scanning tunneling spectroscopy. We neglected the fact that the current contributions of all lower-lying levels are modified because all sample states are shifted down by edV and a new (smaller) transmission factor applies for all the sample states. The effective tunneling barrier becomes higher if the sample states are shifted "downwards". This effect leads to an additional contribution to the current when the voltage is changed, i.e. to the differential conductance dI/dV, as we will show in detail later.

In the simplest approximation, the density of states of the tip and the transmission factor are considered to be voltage independent. Hence the differential conductance is proportional to the energy dependent density of states of the sample

$$\frac{dI}{dV} \propto \rho_{\text{sample}}(eV). \tag{21.3}$$

In this approximation, the differential conductance dI/dV measures the sample density of states at the energy eV relative to the Fermi energy of the sample.

21.2 Experimental Realization of Spectroscopy with STM

There are several variants of spectroscopic measurements using the STM. We consider first the measurement of dI/dV at constant tip-sample distance. In practice, an STS spectrum (differential conductance) is acquired using the following method. The tip is positioned over a certain desired lateral position of the surface at a certain voltage and a certain current flows (usually called stabilization voltage and stabilization current). For a specific density of states these conditions define a certain tip-sample distance. Then the feedback loop is disabled (leaving the tip-sample distance constant during spectroscopy) and the dI/dV signal is recorded over the desired range of voltages.

The primary measured signal is the current I as a function of the voltage V: $I = f(V)$. In the following, we show how a modulation technique (lock-in technique) is used in order to obtain the derivative signal dI/dV as a function of the voltage. Since generically STM is not an AC technique, in STS an AC signal is generated by adding a small modulation voltage $V_M \cos \omega t$ to the applied bias voltage V. The measured modulated current will be

$$I(t) = f(V + V_M \cos \omega t). \tag{21.4}$$

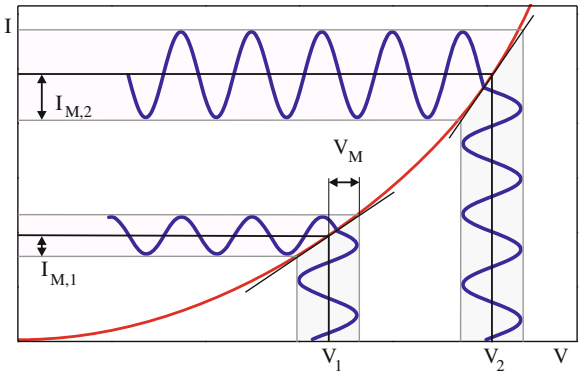

Fig. 21.1 Graphic representation of the measurement of the first derivative of the I-V curve. The voltage is modulated around a value V_1 or V_2. The measured amplitude of the resulting modulated (tunneling) current is proportional to the slope of the I-V curve (dI/dV) at V_1 or V_2, respectively

This signal can be detected by a lock-in amplifier and its amplitude is for small modulation voltages proportional to dI/dV as can be inferred from Fig. 21.1, which shows a hypothetical I-V curve. For the voltage V_1 the derivative (dI/dV) is small and for V_2 the derivative is larger. If now a modulation voltage is applied around a center voltage, as indicated by V_M, this will lead to a corresponding modulation of the measured tunneling current. Figure 21.1 shows that the AC amplitude of the tunneling current I_M is proportional to the slope (derivative) of the I-V curve. As can be seen from Fig. 21.1, the measured slope is averaged over the amplitude of the modulation voltage around the center voltage. Therefore, the energy resolution of the measured dI/dV signal in STS is proportional to the amplitude of the modulation voltage.

In the following, it will be shown more formally that the first derivative of the current signal dI/dV is proportional to the AC amplitude of the signal at the modulation frequency. Moreover, in some spectroscopy techniques such as inelastic spectroscopy the second derivative of the $I(V)$ signal is required. As we will see later, this signal is proportional to vibrational excitations. Therefore, a general statement about the nth derivative of the I-V curve will be derived, which states that the nth derivative of the I-V curve at voltage V is proportional to the AC amplitude of the signal at n-times the modulation frequency. The I-V curve is represented by a function f as $I = f(V)$. If the voltage at position V is modulated with a small harmonic modulation voltage of amplitude V_M, a modulated current $I = f(V + V_M \cos \omega t)$ results. The Taylor expansion of I around the voltage V as a polynomial function of the modulation voltage up to the fourth order can be written as

$$I = \sum_{k=0}^{\infty} \frac{V_M^k}{k!} \frac{d^k f(V)}{dV^k} \cos^k \omega t$$

$$= f(V) + V_M \frac{df(V)}{dV} \cos \omega t + \frac{V_M^2}{2} \frac{d^2 f(V)}{dV^2} \cos^2 \omega t$$

$$+ \frac{V_M^3}{6} \frac{d^3 f(V)}{dV^3} \cos^3 \omega t + \frac{V_M^4}{24} \frac{d^4 f(V)}{dV^4} \cos^4 \omega t + \ldots$$

This Taylor expansion comprises the nth derivative of the I-V curve at V (highlighted in green) and the nth powers of the modulation voltages (highlighted in blue). The nth powers of the $\cos \omega t$ terms can also be expressed as a sum of cos-terms with up to n-times higher frequency (due to a mathematical identity) as

$$\begin{aligned}
\cos^2 \omega t &= \frac{1}{2} + \frac{1}{2} \cos 2\omega t \\
\cos^3 \omega t &= \frac{3}{4} \cos \omega t + \frac{1}{4} \cos 3\omega t \\
\cos^4 \omega t &= \frac{3}{8} + \frac{1}{2} \cos 2\omega t + \frac{1}{8} \cos 4\omega t
\end{aligned}$$

If we now replace the nth powers of the $\cos \omega t$-terms by the expressions with the n-times ωt contributions in the Taylor expansion and group the terms with the same multiples of the modulation frequency, the following expression results for the Taylor expansion

$$\begin{aligned}
I = \sum_{k=0}^{\infty} \frac{V_M^k}{k!} \frac{d^k f(V)}{dV^k} \cos^k \omega t = \\
1 \quad \left[f(V) + \frac{d^2 f(V)}{dV^2} \frac{V_M^2}{4} + \cdots \right] + \\
+V_M \cos \omega t \quad \left[\frac{df(V)}{dV} + \frac{d^3 f(V)}{dV^3} \frac{V_M^3}{8} + \cdots \right] + \\
+\tfrac{1}{4} V_M^3 \cos 2\omega t \quad \left[\frac{d^2 f(V)}{dV^2} + \frac{d^4 f(V)}{dV^4} \frac{V_M^4}{12} + \cdots \right] + \\
+\tfrac{1}{24} V_M^3 \cos 3\omega t \quad \left[\frac{d^3 f(V)}{dV^3} + \frac{d^5 f(V)}{dV^5} \frac{V_M^5}{20} + \cdots \right] + \cdots
\end{aligned} \tag{21.5}$$

If the higher order terms are neglected (not highlighted in (21.5)), the Taylor expansion reduces to a sum of harmonic functions with frequencies of $n\omega t$ (blue in (21.5) times the nth derivative of the I-V curve. Thus the amplitude of the signal at n-times the modulation frequency ω is proportional to the nth derivative of the I-V curve at V. This means in order to measure the nth derivative of the amplitude, the signal component at n-times the modulation frequency has to be measured. This can be done in practice by using the lock-in technique and frequency multiplication of the reference signal n times before multiplication with the measurement signal, as will be shown in the following.

The experimental setup used to measure the nth derivative of the I-V curve is shown in Fig. 21.2. The modulation voltage $V_M \cos \omega t$ is generated by an oscillator and added to the voltage V, which is slowly swept during the measurement of the derivative of the I-V curve. The total bias voltage $V + V_M \cos \omega t$ is then applied to the sample. The oscillator reference frequency is doubled if the second derivative is measured. A phase shift can be applied, which can compensate a possible phase shift in the experimental setup. Then this reference signal is multiplied by the voltage proportional to the tunneling current measured by the preamplifier. The low-pass-

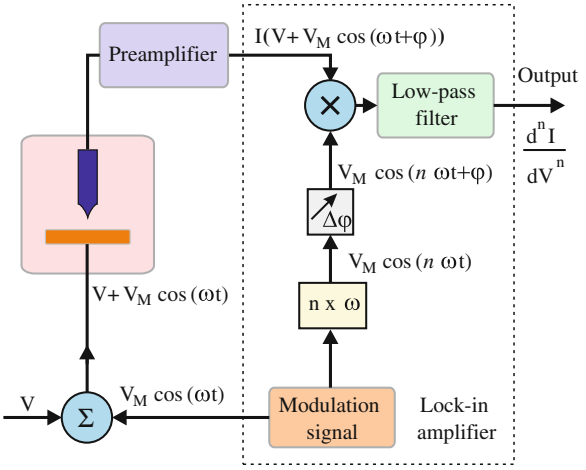

Fig. 21.2 Experimental setup for the measurement of the *n*th derivative of the I-V curve using the lock-in technique

filtered output corresponds to the *n*th derivative of the I-V curve. By a slow scan of the voltage V, the derivative can be measured as a function of the voltage. Long averaging times of the lock-in detection result in a good signal-to-noise ratio. On the other hand, the time to take a spectrum is usually limited by drift which changes the tip-sample distance and thus also the measured current.

Measuring the derivative using the lock-in technique requires some effort. It could be considered possible to obtain the same result by simply measuring the I-V curve and then afterwards obtaining the *n*th derivative by numerical differentiation. However, the numerical derivative of an I-V curve is much more noisy than the derivative obtained using the lock-in technique. Also smoothing of the numerically differentiated signal will not improve the signal recovery to the level achieved with the lock-in technique. Without the lock-in technique all signals are amplified, in the lock-in technique only the signal component at the modulation frequency and with the same phase as the modulation frequency is selected.

21.3 Normalized Differential Conductance

Before we analyze the differential conductance in detail, we introduce a way to normalize the dI/dV spectra. It is often useful to normalize the dI/dV spectra, because the transmission factor induces a background which increases exponentially with the voltage V. The desired signal rides on this large background signal. The normalization procedure has the great advantage that solely the measured data are required and no fit to a model, no calculations or simulations are required. This normalization is an easy and convenient way to plot the dI/dV data. The differential conductance dI/dV is only proportional to the sample density of states if the transmission factor T is considered as constant. However, the transmission factor strongly depends on

voltage. We have seen in Chap. 20 in (20.56) that the transmission factor can be written as

$$T(\varepsilon, V, d) \propto \exp\left(-2d\frac{\sqrt{2m}}{\hbar}\sqrt{\bar{\Phi} + \frac{eV}{2} - \varepsilon}\right). \tag{21.6}$$

Now we consider the effective barrier height, i.e. the term under the square root, as function of the voltage V. The tunneling barrier takes values from $\Phi_{\text{eff}} = \bar{\Phi} + \frac{eV}{2}$ for states at the lower end of the bias window ($\varepsilon = 0$), and reducing down to $\Phi_{\text{eff}} = \bar{\Phi} - \frac{eV}{2}$, for states at the upper end of the bias window ($\varepsilon = eV$), compare also Fig. 20.7. The states with the smallest tunneling barrier, contribute most to the tunneling current. For those states at the upper end of the bias window the tunneling barrier decreases with increasing tunneling voltage as $\Phi_{\text{eff}} = \bar{\Phi} - \frac{eV}{2}$. Due to the exponential dependence in (21.6) the transmission factor increases strongly with the tunneling voltage. This exponential increase of the "background" current with increasing bias voltage is a major problem in scanning tunneling spectroscopy. This background tends to mask density of states (DOS) features in the dI/dV signal. In Fig. 21.3a the dI/dV signal on a Si(111) 2×1 surface is shown. It can be clearly seen that the conductance rises sharply with the applied voltage as expected from the exponentially increasing transmission coefficient. The traces at higher voltages could only be measured by increasing the tip-sample separation (in order to decrease the transmission coefficient). Any small features in the dI/dV curve arising from the density of states are hidden in the exponentially increasing transmission coefficient.

As shown by Feenstra [44, 45], this voltage dependence (and also the distance dependence) can be removed by the normalization of the differential conductance dI/dV by the total conductance I/V. The obtained dimensionless quantity, $(dI/dV)/(I/V)$ provides a convenient plot of the data. The normalized conductance for the data shown in Fig. 21.3a is displayed in Fig. 21.3b. It does not diverge for larger voltages (as the original (dI/dV)) and shows clear peaks at four voltages, which can be assigned to a large density of states at those energies. Moreover, all the dI/dV spectra in Fig. 21.3a, taken at different tip-sample distances, collapse in a single curve showing the practical use of the normalization to suppress effects of the varying tip-sample separation. We note that sometimes the differential conductance is also written as $(dI/dV)/(I/V) = d(\ln I)/d(\ln V)$, which arises because $d(\ln x)/dx = 1/x$. An obvious problem with this normalization arises for semiconductors with a surface band gap, where the current and the differential conductance may go to zero if no surface states are present in the band gap. In this case the normalization procedure is modified by broadening of I/V which results in $I/V > 0$ for all voltages [45].

Apart from providing a convenient plot of the data, the normalized conductance is often identified with the density of states of the sample. In the following, we present the reasoning for giving this assignment. Assuming a constant DOS of the tip according to (21.1) and (21.2), the normalized conductance can be written as

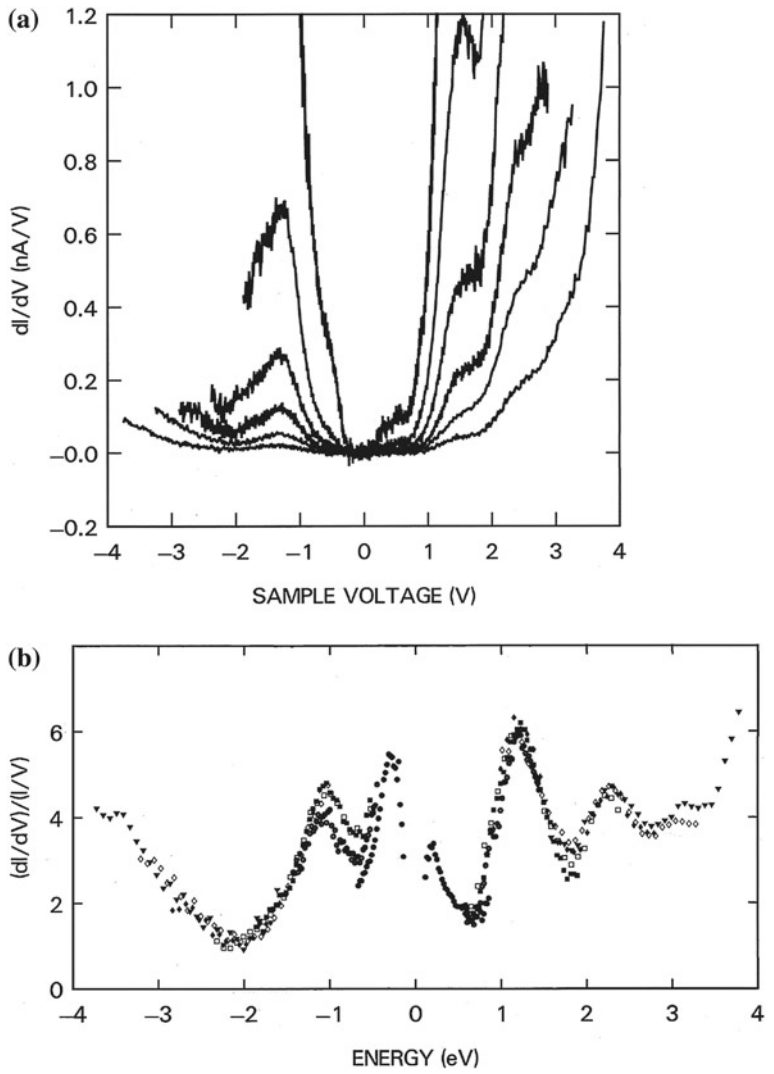

Fig. 21.3 **a** Derivative of current versus voltage data on Si(111) 2×1. The different traces are for different tip-sample distances. DOS features are barely visible on top of the huge background signal. **b** Normalized dI/dV signal: $(dI/dV)/(I/V)$ of the data shown in (**a**). Data for different tip-sample separations are shown as different symbols. In the plot of the normalized conductance, the DOS peaks, which are barely visible in (**a**), are clearly visible (reproduced with permission from [44])

$$\frac{dI/dV}{I/V} \approx \frac{\rho_{\text{sample}}(eV)\, T(\varepsilon = eV, V)}{\frac{1}{eV}\int_0^{eV} \rho_{\text{sample}}(\varepsilon) T(\varepsilon, V) d\varepsilon}$$
$$= \frac{\rho_{\text{sample}}(eV)}{\frac{1}{eV}\int_0^{eV} \rho_{\text{sample}}(\varepsilon)\, \frac{T(\varepsilon, V)}{T(\varepsilon = eV, V)} d\varepsilon}. \tag{21.7}$$

As the *ratio* of transmission factors occurs in the denominator and their dependence on the voltage and tip-sample separation is similar, the denominator can be considered as slowly varying with voltage. Thus the normalized differential conductance can be considered roughly proportional to the density of states at eV. However, it must be stressed that several crude approximations entered in order to derive this estimate. The normalized conductance is still more a convenient way to plot the data without further analysis than a quantitative estimate of the sample density of states.

In the data considered in Fig. 21.3a, several measurements were performed at different tip-sample distances in order to extend the dynamic range of the measurements. An alternative approach is to vary the tip-sample distance continuously during one measurement, i.e. to approach the tip towards the sample at smaller bias voltage in order to obtain a larger signal of the differential conductance (to compensate the decreasing transmission factor). Such an acquisition method gives a wide dynamic range of tunneling current and the differential conductivity.

Another important issue to be mentioned is the role of the tip density of states. The spectra acquired with STS always also contain information about the electronic states of the tip. dI/dV can be related to the sample density of states only for tips with a constant (flat) DOS ($\rho_{\text{tip}}(\varepsilon) \approx const.$). In the actual experiment, however, the tip DOS is not always constant and results should be reproduced with several tips.

21.4 Relation Between Differential Conductance and the Density of States

In the following, we will analyze the differential conductance dI/dV in more detail and explore whether information on the sample density of states can be extracted from the scanning tunneling spectroscopy spectra. The expression for the tunneling current calculated within the large voltage approximation

$$I = \frac{4\pi e}{\hbar} \int_0^{eV} \rho_{\text{tip}}(\varepsilon - eV) \rho_{\text{sample}}(\varepsilon) T(\varepsilon, V) d\varepsilon, \tag{21.8}$$

involves integration over all the electron states between the tip and sample Fermi levels. In scanning tunneling spectroscopy dI/dV is measured. Therefore, we now calculate the derivative of the current. There is a rule (Leibniz integral rule) for the

differentiation of an integral with respect to a parameter x, if also the integration boundaries depend on this parameter as $b = b(x)$

$$\frac{d}{dx}\int_0^{b(x)} f(t,x)dt = \frac{db}{dx} f[b(x),x] + \int_0^{b(x)} \frac{\partial}{\partial x} f(t,x)dt. \tag{21.9}$$

With the assignments $x = V$, $b(x) = eV$, $t = \varepsilon$, and $f(t,x) = f(\varepsilon, V) = \rho_{\text{tip}}(\varepsilon - eV)\rho_{\text{sample}}(\varepsilon)T(\varepsilon, V)$, the above rule can be used to take the derivative of (21.8) at the voltage V

$$\begin{aligned}
\frac{dI}{dV}\frac{\hbar}{4\pi e} &= \frac{d}{dV}\left[\int_0^{eV} \rho_{\text{tip}}(\varepsilon - eV)\rho_{\text{sample}}(\varepsilon)T(\varepsilon, V)d\varepsilon\right] \\
&= e\rho_{\text{tip}}(0)\rho_{\text{sample}}(eV)T(eV, V) \\
&\quad + \int_0^{eV} \frac{\partial}{\partial V}\left[\rho_{\text{tip}}(\varepsilon - eV)\rho_{\text{sample}}(\varepsilon)T(\varepsilon, V)\right]d\varepsilon \qquad (21.10) \\
&= e\rho_{\text{tip}}(0)\rho_{\text{sample}}(eV)T(eV, V) \\
&\quad + \int_0^{eV} \frac{\partial\rho_{\text{tip}}(\varepsilon - eV)}{\partial V}\rho_{\text{sample}}(\varepsilon)T(\varepsilon, V)d\varepsilon \\
&\quad + \int_0^{eV} \rho_{\text{tip}}(\varepsilon - eV)\rho_{\text{sample}}(\varepsilon)\frac{\partial T(\varepsilon, V)}{\partial V}d\varepsilon. \qquad (21.11)
\end{aligned}$$

In the following, we assume for simplicity (and because the tip density of states is usually unknown) that ρ_{tip} is constant. Therefore, the second term in (21.11) vanishes. If also the transmission factor were be constant, the third term vanishes and the initial approximation (21.2) is recovered. This approximation is often used because it leads to the very simple result that the measured dI/dV signal is proportional to the sample density of states. However, since we know that the transmission factor is exponentially dependent on the tunneling voltage this approximation is not justified. Therefore, we will not neglect the third term of (21.11) in the following. Before we proceed with a detailed analysis of (21.11) we would like to explain the physical significance of the two remaining terms in this equation.

In Fig. 21.4 tunneling between tip and sample at an applied bias voltage eV is shown in blue. We consider the tip density of states as constant for simplicity. The situation for a slightly larger bias voltage is shown in red. The sample DOS is shifted downwards by edV. Due to this, a new sample state enters the bias window at $e(V + dV)$ above the sample Fermi level (the largest peak in the sample DOS in Fig. 21.4).

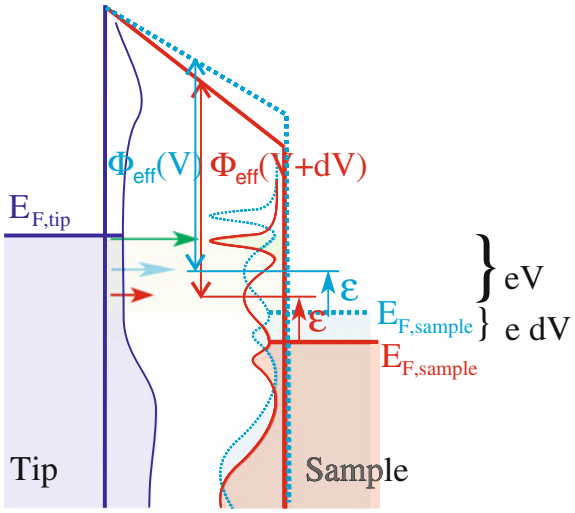

Fig. 21.4 Tunneling barrier shown in *blue* for an applied bias voltage of eV. The corresponding situation for a tunneling voltage increased by edV is shown in *red*. There are two contributions to the tunneling current upon a small increase of the tunneling voltage. First, new sample states shift into the *top* of the bias window. This results in an additional contribution to the current proportional to $\rho_{\text{sample}}(eV+edV)\,T(eV+edV, V+dV)$ and corresponds to the first term in (21.11). The second contribution to the current comes from all sample states between zero and eV, which are shifted down due to the increase of the tunneling voltage. A modified (decreased) transmission factor applies to all these states, because the effective barrier $\Phi_{\text{eff}}(V+dV)$ is increased by the increased tunneling voltage. The smaller transmission factor is indicated by the shorter *horizontal red arrow* compared to the *blue arrow* symbolizing the transmission factor. This contribution corresponds to the last term in (21.11)

The corresponding additional contribution to the tunneling current can be written as

$$dI_1 \propto \rho_{\text{sample}}[e(V+dV)]\,T[e(V+dV), V+dV]edV, \tag{21.12}$$

corresponding to the green horizontal arrow in Fig. 21.4. This contribution corresponds to the first term in (21.11).

However, there is another contribution to the tunneling current arising upon an increase of the tunneling voltage. This arises from contributions to the tunneling current from all sample states in the bias window (i.e. between $E_{F,\text{sample}}$ and $E_{F,\text{sample}}+eV$). Due to the bias voltage increase all these states shift down by edV, as indicated in Fig. 21.4 (red). While the contribution to the tunneling current still arises from the same sample states within the bias window (same DOS at ε) the associated transmission factor is now modified. The new transmission factor $T(\varepsilon, V+dV)$ applies instead of $T(\varepsilon, V)$ and the corresponding effective barrier height changes to $\Phi_{\text{eff}} = \bar{\Phi} + \frac{e(V+dV)}{2} - \varepsilon$. Thus, due to the increase in the tunneling voltage by dV the barrier height increases by $\frac{1}{2}edV$.Using (21.8) the current change due to the different

transmission factor integrated over the bias window is

$$dI_2 \propto \int_0^{eV} \rho_{\text{sample}}(\varepsilon)\ [T(\varepsilon, V + dV) - T(\varepsilon, V)]\, d\varepsilon, \tag{21.13}$$

leading to a contribution to the differential conductance of

$$\frac{dI_2}{dV} \propto \int_0^{eV} \rho_{\text{sample}}(\varepsilon) \frac{\partial T(\varepsilon, V)}{\partial V} d\varepsilon. \tag{21.14}$$

This contribution to the differential conductance corresponds to the last term in (21.11).

Summarizing the previous discussion, we obtain the following relation for the differential conductance if the tip density of states is constant

$$\begin{aligned} \frac{dI}{dV}\frac{\hbar}{4\pi e} &= e\rho_{\text{tip}}\rho_{\text{sample}}(eV)T(eV, V) \\ &\quad + \int_0^{eV} \rho_{\text{tip}}\rho_{\text{sample}}(\varepsilon) \frac{\partial T(\varepsilon, V)}{\partial V} d\varepsilon. \end{aligned} \tag{21.15}$$

with the transmission factor

$$T(\varepsilon, V, d) \propto \exp\left[-2d\sqrt{\frac{2m}{\hbar^2}\left(\bar{\Phi} + \frac{eV}{2} - \varepsilon\right)}\right]. \tag{21.16}$$

With some parameters like the (average) barrier height and the barrier thickness obtained from additional experiments and a sample density of states obtained for instance from theory calculations, the differential conductance can be calculated using the above equations, and compared to the experimentally observed differential conductance. If a model is available for the sample density of states, the parameters of this model can be fitted in order to match the obtained differential conductance to the measured dI/dV signal.

21.5 Recovery of the Density of States

In the following, we will show how it is possible to solve (21.15) analytically for the sample density of states (if the tip density of states is constant and with another approximation applied subsequently). The aim is to obtain an analytic expression for

the sample density of states as a function of the measured dI/dV signal. Details and extensions of this approach can be found in [46].

The derivative of the transmission factor occurring in (21.15) can be calculated using (21.16) (since the T is an exponential function, its derivative is T times the inner derivative) resulting in the following expression for the differential conductance

$$\frac{dI}{dV} = \frac{4\pi e}{\hbar}\left[e\rho_{\text{tip}}\rho_{\text{sample}}(eV)T(eV,V) - \int_0^{eV} \rho_{\text{tip}}\rho_{\text{sample}}(\varepsilon)T(\varepsilon,V)\frac{ed\frac{\sqrt{2m}}{\hbar}}{2\sqrt{\bar{\Phi}+(eV/2)-\varepsilon}}d\varepsilon\right] \tag{21.17}$$

The second term in (21.17) is (without the fraction) the tunneling current according to (21.1). If this fraction is replaced (as an approximation) by its value at the middle of the bias window ($\varepsilon = eV/2$), this term becomes a constant not dependent on either ε, or V. Thus the differential conductance can be written as

$$\frac{dI}{dV} = \frac{4\pi e^2}{\hbar}\rho_{\text{tip}}\rho_{\text{sample}}(eV)T(eV,V) - \frac{ed\sqrt{2m}}{2\hbar\sqrt{\bar{\Phi}}}I(V). \tag{21.18}$$

Finally, we obtain for the desired density of states of the sample

$$\rho_{\text{sample}}(eV) = \frac{\hbar}{4\pi e^2\rho_{\text{tip}}T(eV,V)}\left[\frac{dI}{dV} + \frac{ed\sqrt{2m}}{2\hbar\sqrt{\bar{\Phi}}}I(V)\right]. \tag{21.19}$$

This expression can be used to relate the sample density of states to the measured differential conductance dI/dV and the measured tunneling current $I(V)$, with the average barrier height $\bar{\Phi}$ and the tip-sample distance d determined independently.

The first term in (21.19) leads us back to the original conclusion that the sample density of states is proportional to the differential conductance dI/dV. The second term in (21.19) is proportional to the tunneling current. If we evaluate this second term for usual tunneling conditions ($\bar{\Phi} \approx 4\,\text{eV}$, $d \approx 1\,\text{nm}$, and $I = 1\,\text{nA}$) this term evaluates to 1.3 nA/V. Since the first term dI/dV also has values in the nA/V range, this means that the second term which is proportional to the current is usually not negligible compared to the differential conductance term. If $I(V)$ and dI/dV are both measured, a quantity proportional to the density of states can be calculated with the help of (21.19).

If we suspend the approximation that the fraction in (21.17) is replaced by a constant, this equation can be solved for the density of states numerically. If we would suspend the other approximation that the tip density of states is constant, an additional term would enter into (21.17). Due to the fact that tip and sample density of states enter symmetrically into the model, there is no unique result for the

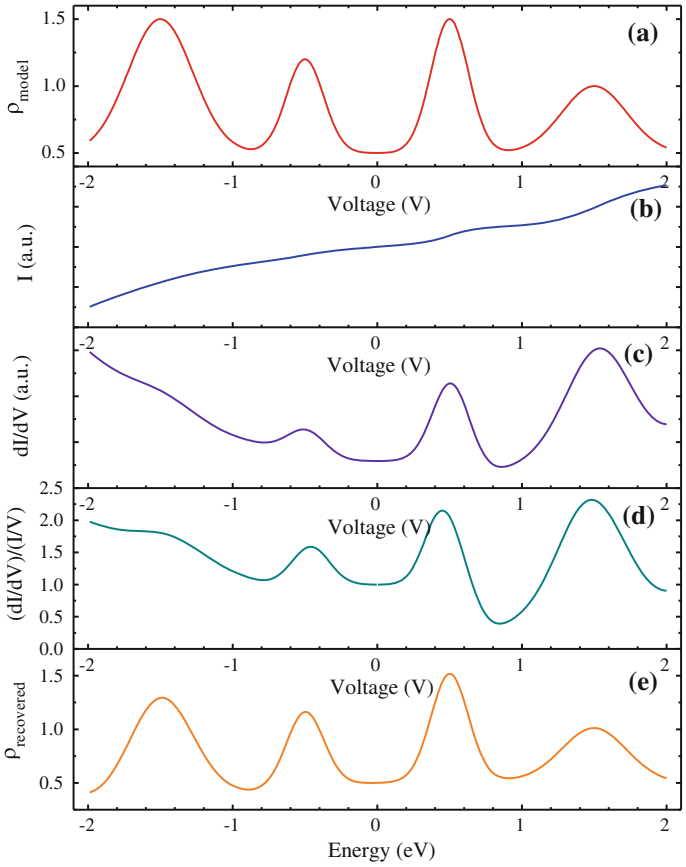

Fig. 21.5 **a** Model LDOS assumed for the subsequent calculation of **b** the current, **c** the conductance, **d** the normalized conductance, and **e** the recovered LDOS. Positive bias voltages correspond to empty sample states and negative bias to filled states, respectively

tip and sample density of states, even with numerical methods. More information can be obtained by measuring the dI/dV signal at different tip-sample distances, as described in [46], in order to disentangle the tip and sample density of states.

In order to explore how close the different approximations presented above come to the original sample density of states (LDOS), a model LDOS can be chosen and the (normalized) differential conductance and the recovered LDOS can be calculated from the initial model LDOS. Such a procedure is shown in Fig. 21.5. From a model DOS shown in Fig. 21.5a the current is calculated from (21.8) and shown in Fig. 21.5b. The differential conductance calculated according to (21.15) is shown in Fig. 21.5c. The normalized differential conductance $(dI/dV)/(I/V)$ calculated according to (21.15) and (21.8) is shown in Fig. 21.5d. The recovered LDOS according to (21.19) is shown in Fig. 21.5e. Peaks at nearly the same position as in the model

LDOS are observed, however, the intensities of the peaks are quite different than in the model LDOS. Unoccupied states of the sample (positive sample bias voltages) are observed much more clearly and with higher intensity than the occupied states. A steeply rising background leads to the result that the LDOS peak at $-1.5\,\mathrm{V}$ is only observed as a weak shoulder in the (normalized) conductance. This numerical simulation demonstrates that the intensities of the peaks observed in the normalized conductance are not proportional to the density of states, particularly at negative sample bias voltages. The recovered LDOS (Fig. 21.5e) reproduces the intensities of the starting LDOS best. The calculations were performed for $d = 0.7\,\mathrm{nm}$ and $\bar{\Phi} = 4\,\mathrm{eV}$.

21.6 Asymmetry in the Tunneling Spectra

In Fig. 21.5c, we have already noticed an asymmetry in the STS spectra. The dI/dV signal represents the DOS of the empty sample states (positive bias voltages) reliably, while the dI/dV signal for filled sample states is superimposed by a large background signal. Here we will explain this asymmetry. Moreover, we will also see that a structured tip density of states contributes in an asymmetric way to the differential conductance.

As shown in Fig. 21.6a, for increasing positive bias voltages new empty sample states enter the bias window at the top. The smallest barrier is present at the top of the bias window Thus those states contribute with the maximal transmission factor, leading to a maximal contribution to the current and to dI/dV. For negative bias voltages, on the other hand, new filled sample states enter the bias window from the bottom (Fig. 21.6b). Due to the reduced transmission factor at the bottom of the bias window (larger tunneling barrier) those states contribute with a much smaller weight to the current and to dI/dV.

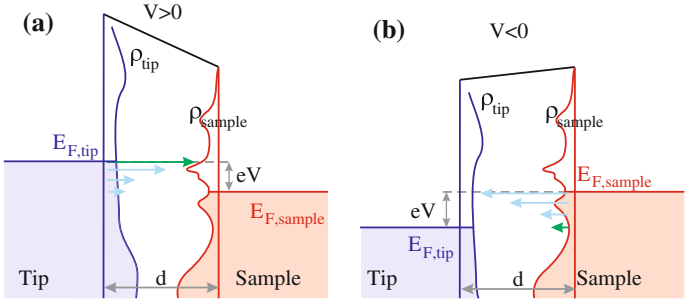

Fig. 21.6 Asymmetry in the tunneling spectra between positive and negative voltages. **a** For positive sample bias voltages new sample states entering the bias window at the *top* are probed with a large transmission factor (*long green arrow*). **b** For negative sample voltages new filled sample states enter the bias window from the *bottom* with a very small transmission factor (*short green arrow*)

For usual tunneling conditions ($d = 0.7\,\text{nm}$, $\Phi = 4\,\text{eV}$, and $V = 1\,\text{V}$) the transmission factor decreases by one half within 0.3 eV. This means the exponentially decreasing transmission factor has noticeable values only in a small energy range below the highest Fermi level. For the case of positive sample bias voltages (Fig. 21.6a), this means that electrons from the tip tunnel (due to the transmission factor) only in a relatively narrow energy range of $\sim$300 meV below the top of the bias window as indicated by the arrows in Fig. 21.6a. In terms of the general expression for the differential conductance, the first term in (21.15) is proportional to the (empty) sample density of states at eV above the sample Fermi level and contributes with the largest transmission factor. The second term in (21.15) depends on the derivative of the transmission factor (which is itself proportional to the transmission factor (21.16), due to its exponential dependence). Thus the exponential decrease of the transmission factor reduces any contributions of more than 300 meV below the top of the bias window.

For negative sample bias voltages an increase of the tunneling voltage by a small value dV gives rise to new filled sample states entering the bias window at its bottom (Fig. 21.6b). This leads to an additional current from filled sample states at energies of $|eV|$ below the top of the bias window. Tunneling from these states is related to the largest tunneling barrier, i.e. to the smallest transmission factor within the bias window. Thus the contribution from filled sample states is quite small for negative sample voltages. This means that a peak in the occupied density of states will only lead to a relatively small peak in the measured dI/dV. This explains the insensitivity of STS to the occupied density of states which we saw in Fig. 21.5c, d, where the peaks of the model LDOS for negative voltages (Fig. 21.5a) are hardly visible on top of the exponentially rising background. This exponentially increasing background comes from the second term in (21.15). The change of the transmission factor with increasing voltage (proportional to the transmission factor itself) is integrated over the bias window up to the sample Fermi level. The largest transmission factor acts at the sample Fermi level, and increases exponentially with more negative bias voltages. This leads to the large increasing background in the dI/dV signal at negative bias voltages, as observed in Fig. 21.5c.

There is also another kind of asymmetry in the tunneling spectra. If, in contrast to what we assumed before, the density of the states of the tip is not constant, this has a strong influence on the dI/dV for the occupied sample states but a smaller influence for the spectroscopy of the empty sample states. We assume that the tip density of states has a peak and consider tunneling into the empty sample states ($V > 0$). If the peak in the tip DOS is outside the region close to the tip Fermi energy it does not contribute too much to the current since it is "damped" by the transmission factor. If the peak is close to the Fermi level, it will give rise to an additional contribution, which shifts *together* with the tip Fermi level and probes the empty sample states. Thus, when probing empty sample states, a non-flat tip DOS is not a major problem.

For the case of sampling the filled sample states ($V < 0$), the situation is different. The filled sample states close to the sample Fermi level make the highest contribution to the current and probe the empty tip states. Therefore, dI/dV spectra are very sensitive to structures in the empty tip density of states for negative sample bias

voltages and thus, at negative bias voltages the dI/dV spectra can be greatly influenced by the tip density of states and care has to be taken that the tip density of states is constant.

21.7 Beyond the 1D Barrier Approximation

Up to now, we have mostly studied one-dimensional models where the whole momentum of an electron is considered to be perpendicular to the surface of the electrodes. In reality, the participating states have momentum components in all three spatial directions.

In the free electron model, the momentum parallel to the surface is included by writing

$$E = \frac{\hbar^2}{2m}(k_x^2 + k_y^2 + k_z^2) = \frac{\hbar^2}{2m}(k_{||}^2 + k_{\perp}^2) = E_{||} + E_{\perp}. \tag{21.20}$$

In the simplest extension of the 1D model to three dimensions only the energy component perpendicular to the surface is considered as "effective" for tunneling. Thus the energy component entering into the transmission factor is the perpendicular component $E_{\perp} = E - E_{||}$, and the transmission factor becomes

$$T(E, V, d) \propto \exp\left(-2d\sqrt{\frac{2m}{\hbar^2}\bar{\Phi} + \frac{eV}{2} - \left(E - E_{||}\right)}\right). \tag{21.21}$$

If we consider states with a given energy E, the effective tunneling barrier will be smallest for the states with $E_{||} = 0$. For surface states, this condition can be expressed as: the transmission factor is largest for the states at the Γ-point of the two-dimensional surface Brillouin zone.

21.8 Energy Resolution in Scanning Tunneling Spectroscopy

The energy resolution in scanning tunneling spectroscopy is determined by the range of energies which contribute to the tunneling current, and is usually considered in two limits. In the first limit large tunneling voltages (several volt) are considered. In this limit the thermal broadening due to the Fermi functions is neglected. In this limit the main contribution to the energy resolution comes from the transmission factor. The transmission factor according to (21.16) is plotted in Fig. 21.7 for usual tunneling parameters. The transmission factor decreases to one half for a decrease of the voltage of 300 meV. This means that electrons within this energy range contribute

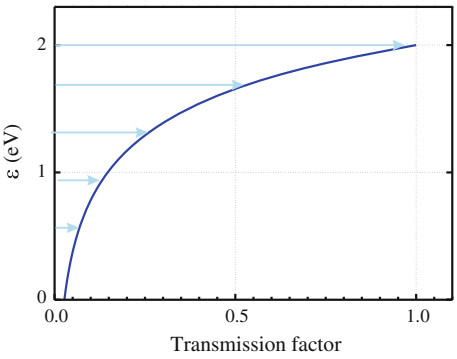

Fig. 21.7 Energy resolution of tunneling electrons visualized by the energy dependence of the transmission factor for usual tunneling parameters ($d = 0.7\,\text{nm}$, $\Phi = 4\,\text{eV}$). The width of the exponential function is about 300 meV

to the tunneling current. Thus the energy resolution in STS using larger voltages has a resolution of only several hundred meV.

In the second limit of small voltages and low temperatures used simultaneously, two effects limit the energy resolution in STS: the thermal broadening of the Fermi functions, as well as the modulation amplitude used in the lock-in detection scheme.

First, we discuss the energy resolution limited due to the thermal broadening of the Fermi functions. We lift the zero temperature approximation in which the Fermi functions were considered as step functions. In this case, we use (20.65), which reads in the limit of constant tip density of states as

$$I = \frac{4\pi e}{\hbar}\rho_{\text{tip}} \int_{-\infty}^{\infty} \{f(\varepsilon - eV) - f(\varepsilon)\}\, \rho_{\text{sample}}(\varepsilon) T(\varepsilon, V, d) d\varepsilon. \tag{21.22}$$

The thermal broadening of the Fermi functions extends over an energy range in the order of kT, i.e. in the meV range. On this scale, the transmission factor which changes at a range of hundreds of meV can be considered to be approximately constant. With this approximation, the differential conductance results as

$$\begin{aligned} \frac{dI}{dV} &= \frac{4\pi e}{\hbar} T(d)\rho_{\text{tip}} \frac{d}{dV} \int_{-\infty}^{\infty} \{f(\varepsilon - eV) - f(\varepsilon)\}\, \rho_{\text{sample}}(\varepsilon) d\varepsilon \\ &= \frac{4\pi e}{\hbar} T(d)\rho_{\text{tip}} \int_{-\infty}^{\infty} \frac{\partial f(\varepsilon - eV)}{\partial V} \rho_{\text{sample}}(\varepsilon) d\varepsilon. \end{aligned} \tag{21.23}$$

The slope of the Fermi function at $\varepsilon = eV$ is $-1/(2\,\text{kT})$, as visualized in Fig. 21.8a. Thus the Fermi function drops from one (occupied) to zero (empty) within a range of $\sim 4\,\text{kT}$ around $\varepsilon = eV$. The (negative) derivative of $f(\varepsilon)$ is plotted in Fig. 21.8b.

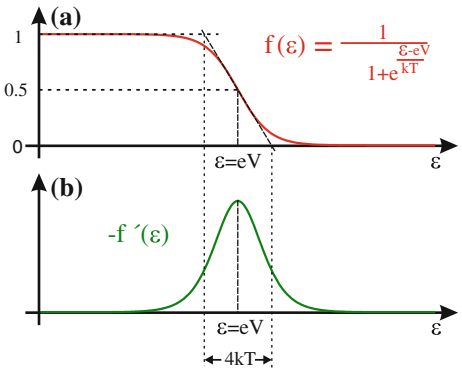

Fig. 21.8 **a** Fermi function as a function of energy. The transition from occupied states to empty states occurs in a range of ∼4 kT around $\varepsilon = eV$. **b** df/dE is zero, apart from a peak around E_F with a width of 4 kT. This means that only those states contribute to the differential conductance which are in the range of ∼4 kT around E_F

It is zero everywhere, except for a peak with a width of ∼4 kT around $\varepsilon = eV$. Since the derivatives of f with respect to ε and V differ only by the factor $-e$, the same arguments also apply to the derivative of f with respect to V, which enters in (21.23).

Therefore, according to (21.23), dI/dV can also be considered as the convolution of the density of states with the "thermal resolution function" $\partial f(\varepsilon - eV)/\partial V$ of width 4 kT. The higher the temperature, the larger the broadening of the Fermi function and the worse the energy resolution in STS. Assuming a delta function in the density of states of the sample, this would lead to a peak in dI/dV with a width of ∼4 kT. Here we were assuming the simple tangent approximation for the width of the Fermi function. A more rigorous evaluation of the thermal broadening results in a thermal broadening of 3.2 kT (FWHM) for Gaussian peaks in the density of states. This corresponds to a peak width (due to thermal broadening) of about 0.28 meV per K. At room temperature, this leads to a peak width of 83 meV. At 4 K the energy resolution in STS drops to 1.2 meV.

The modulation voltage used in the lock-in detection also leads to an (instrumental) broadening of the energy resolution. According to Fig. 21.1, a modulation voltage V_{mod} leads to an averaging (broadening) over a voltage range of about $2V_{\text{mod}}$. The average broadening is roughly two times the RMS value of the modulation voltage $V_{\text{mod,RMS}}$.

Taking both independent effects into account (energy resolution due to the modulation voltage and due to the broadening of the Fermi functions), the energy resolution ($\Delta E = e\Delta V$) becomes

$$\Delta E = \sqrt{(2eV_{\text{mod,RMS}})^2 + (0.28\,\text{meV/K} \cdot T)^2}. \tag{21.24}$$

Usually the electronics is tuned in such a way that the contribution due to the modulation voltage is not the limiting contribution to the energy resolution.

For a superconductor the density of states has a sharp peak on both sides of the superconducting gap. Therefore, superconductors provide a good benchmark to study the energy resolution as a function of temperature. In Fig. 21.9 STS data

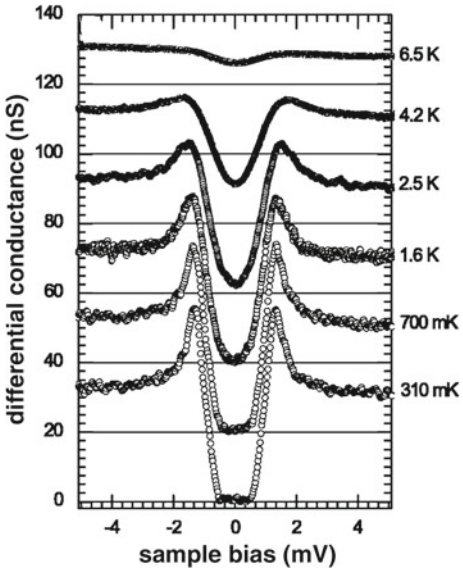

Fig. 21.9 Differential conductance measured by STS around the superconducting gap of $NbSe_2$. The energy resolution in STS increases with decreasing temperature (*sharper peaks*). A modulation voltage of 24 μV RMS was used (reproduced with permission from [47])

measured on the superconductor $NbSe_2$ are shown for different temperatures (below the transition temperature). For the lowest temperature the dI/dV signal shows two peaks at the superconduction band edges and a vanishing dI/dV signal inside the band gap. Towards higher temperatures the width of the peaks broadens. According to the energy resolution of 0.28 meV/K, mentioned before, energy resolution at 4.2 K is around 1.2 meV, while it is about 0.1 meV at 310 mK. At such low temperatures the energy resolution is also limited by the modulation voltage.

21.9 Barrier Height Spectroscopy

In conventional STS, the tunneling current and dI/dV is measured as a function of the applied voltage, providing a measure of the sample density of states. One of the uncertainties in such experiments is that the effective height and width of the tunneling barrier are generally unknown. However, it is possible to measure the tunneling barrier height experimentally. If the tunneling current is measured as a function of the distance between tip and sample, the average barrier height $\bar{\Phi}$ can be determined from the exponential decrease of the tunneling current with increasing tip-sample distance d due to the transmission coefficient $T(\varepsilon, V, d)$. At low bias voltages ($\bar{\Phi} \gg V$), the transmission factor does not depend on the tunneling voltage (as shown in (20.59)) and can be approximated as

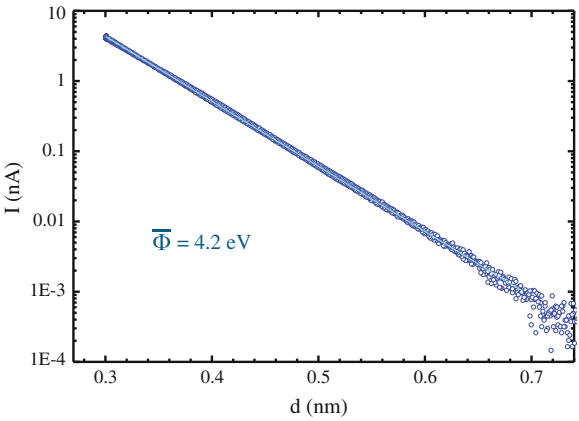

Fig. 21.10 $I(z)$ spectroscopy on a Au(111) sample with a Au tip [48]. The feedback is stopped and the tunneling current is measured as function of the relative tip-sample distance. From the measured exponential behavior the effective barrier height can be determined. The bias voltage is only 2 meV, and thus the condition $\bar{\Phi} \gg V$ is fulfilled

$$T(\bar{\Phi}, d) \propto \exp\left(-2d\sqrt{\frac{2m}{\hbar^2}\bar{\Phi}}\right). \tag{21.25}$$

Thus the transmission factor can be written as a constant in front of the integral in the expression for the tunneling current (21.1). In this case, the tunneling current is proportional to the transmission factor. The average tunneling barrier $\bar{\Phi} = (\Phi_{\text{tip}} + \Phi_{\text{sample}})/2$ can be fitted from the experimentally measured exponential dependence of the tunneling current on the tip-sample separation d.

In experiments, the feedback is stopped and the tunneling current is measured as a function of the tip-sample distance d as $I(d)$. A plot of the tunneling current for varying tip-sample distances is shown in Fig. 21.10. The tunneling barrier $\bar{\Phi}$ (average barrier height) is obtained by an exponential fit to the data. In this analysis only the relative change of the tip-sample distance enters and not the absolute tip-sample distance, which is more difficult to obtain. The absolute tip-sample distance can be measured by approaching the tip closer to the sample until mechanical contact is made. This is indicated by the initially very high tunneling resistance in the order of GΩ dropping approximately to the inverse of the conductance quantum of about 12 kΩ. Since this transition to the low resistance state is quite sharp, the point of contact can be identified. However, it must be considered that at such close tip-sample distances forces between tip and sample lead to an elastic extension of the tip length [49].

The tunneling gap has to be very stable in order to perform $I(d)$ spectroscopy or barrier height spectroscopy in practice. Any drift of the tip-sample distance would change the value measured for the barrier height.

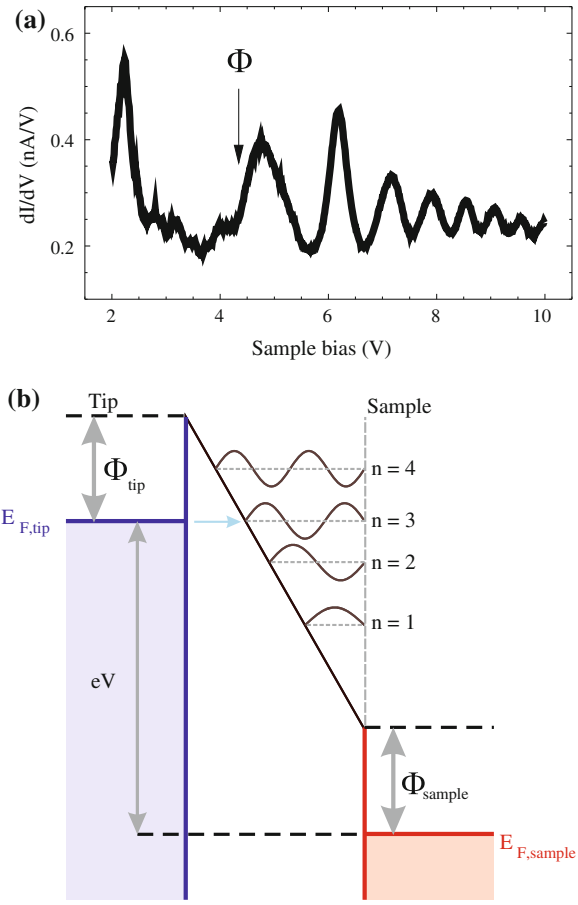

Fig. 21.11 **a** Oscillations in the *dI*/*dV* signal occur on a (bismuth-covered) Si surface if the bias voltage is larger than the work function (indicated by Φ) [50]. **b** The transmission is particularly large if an integer multiple of nodes of the wave function fits between the potential step at the sample surface and the potential wall given by the vacuum level

21.10 Barrier Resonances

If the tunneling voltage exceeds the barrier height, oscillations in the *dI*/*dV* signal are observed, as shown in Fig. 21.11a (cf. Fig. 20.3). In some cases, the feedback is kept active during the measurement of the *dI*/*dV* signal, which maintains the measured current in a desired regime of high sensitivity. In this mode, the tip-sample distance will automatically increase towards larger voltages in order to maintain a constant tunneling current. The *dI*/*dV* signal is measured using the previously described modulation technique with a modulation frequency above the feedback bandwidth of the current regulation. In the following, we will discuss the reason for the occurrence of these barrier resonances.

When discussing the one-dimensional potential barrier model, we have seen that the transmission factor oscillates if the energy of the electron is larger than the barrier height. The transmission factor of one is reached only if an integer multiple of nodes of the wave function fits in the barrier width, as shown in Fig. 20.3. Here, similarly,

the resonances are caused by standing electron waves, which are reflected back and forth between the potential step at the sample surface and the potential wall given by the vacuum level, as shown schematically in Fig. 21.11b. It should be remembered that in quantum mechanics a (partial) reflection of the wave function also occurs at a downward potential step.

21.11 Spectroscopic Imaging

In point spectroscopy considered so far, the dI/dV signal is recorded as a function of voltage at one specific point above the surface. More information about the variation of the electronic structure at the surface can be obtained by the spatial mapping of a specific spectroscopic feature. A spectroscopic image yielding the map of the local density of states (LDOS) at a certain energy can be obtained by performing a slow constant current (topographic) STM scan at a certain bias voltage V and simultaneously recording the (normalized) dI/dV signal at the voltage V. Using a modulation technique, an image of the differential conductance is acquired for a certain bias voltage called spectroscopic image (or LDOS map). States of different energies can be mapped separately.

21.11.1 Example: Spectroscopy of the Si(7 × 7) Surface

In this section, spectroscopic measurements of the Si(111)-(7 × 7) are presented as an example of the application of the previously described spectroscopic techniques. Before we come to the spectroscopic data, the structure of this surface is explained. Surface reconstruction is the rearrangement of the surface atoms due to the termination of the bulk structure at the solid vacuum interface. The (7 × 7) reconstruction of the Si(111) surface is a complex but at the same time very frequently studied structure. One unit cell of the reconstruction is shown in Fig. 21.12. It consists of two triangular half-unit cells (HUC). The top silicon layers (first and second layer in Fig. 21.12) are stacked with the normal (bulk) sequence in one half-unit cell, while in the other HUC there is a stacking fault present relative to the bulk structure. Because of this stacking fault, the two half-unit cells are not equivalent and are referred to as faulted (F) and unfaulted (U) half-unit cells. The reconstruction is terminated by 12 adatoms shown as red balls in Fig. 21.12 (resting on the rest atoms). The rest atoms are in the first layer. Another obvious structural element of this reconstruction is the corner hole present at the corners of the unit cell. The adatoms located close to the corner holes are called corner adatoms, while the remaining adatoms are called center adatoms. One reason for the formation of the (7 × 7) reconstruction is the reduction of the number of unsaturated, energetically costly (dangling) bonds, originating from the breaking of the bulk to form a surface. For the bulk terminated surface there would be one dangling bond per (1 × 1) unit cell which would result in 49 dangling

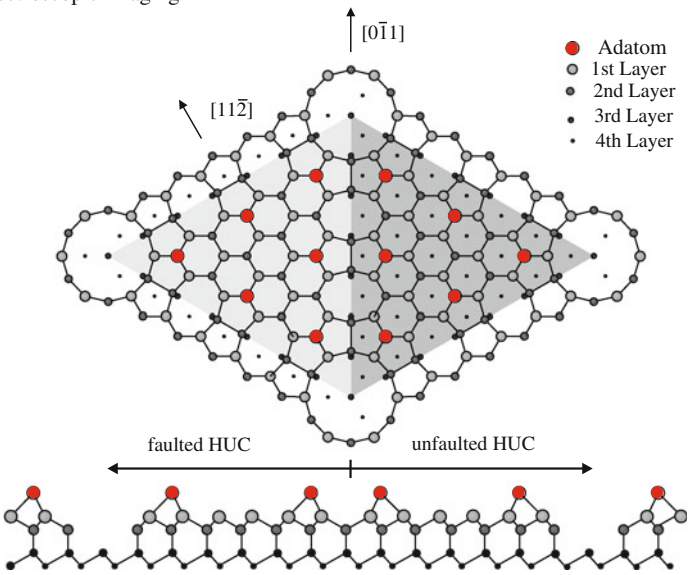

Fig. 21.12 Schematic *top* and *side view* of one unit cell of the Si(111)-(7 × 7) reconstruction. The half-unit cells with and without stacking fault are labeled *F* and *U* and shaded in *light* and *dark gray*, respectively

bonds on the area of the (7 × 7) unit cell. The (7 × 7) reconstruction reduces this number of dangling bonds to 19 dangling bonds on the reconstructed unit cell. The formation of the (7 × 7) reconstruction therefore leads to a clear reduction in energy, compared to the hypothetical unreconstructed surface. The reduction of the number of dangling bonds is responsible for the stability of the (7 × 7) reconstruction under vacuum conditions.

In STS the normalized conductance $(dI/dV)/(I/V)$ represents approximately a measure of the local density of electron states (LDOS) on the sample surface. Figure 21.13c–e shows point spectra, i.e. normalized conductance as a function of the bias voltage at specific points over the surface. In Fig. 21.13a a constant current STM image of the reconstructed surface is shown at $V = +2\,\mathrm{V}$ and $I = 0.1\,\mathrm{nA}$. Only the adatoms are seen in the image and all appear equivalent (positive sample bias voltage). In Fig. 21.13b a corresponding schematic shows the positions of corner adatoms (circles), center adatoms (crosses), and rest atom dangling bonds (dots) in the faulted (F) and unfaulted (U) halves of the surface unit cell. Figure 21.13c–e shows point STS spectra of the Si(111)(7 × 7) surface measured at $T = 7\,\mathrm{K}$ on corner adatoms, center adatoms, and rest atoms in between the adatoms. The spectral features at specific energies can be assigned to specific positions as indicated in Fig. 21.13c–e. For instance the S_2 peak at $-1.3\,\mathrm{eV}$ can be assigned to the rest atoms while the S_1 and the $S_1^{'}$ peaks can be assigned to the corner adatoms. This assignment is also made using the spectroscopic images shown below. In order to acquire these point spectra the technique of variable tip-sample separation was used. Starting at the original tunneling voltage $V_S = 2\mathrm{V}$ (also sometimes called stabilization voltage),

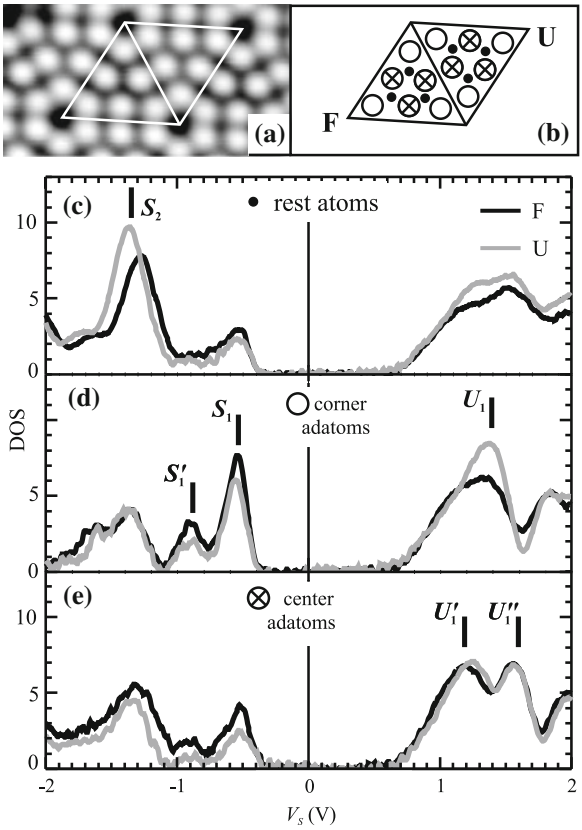

Fig. 21.13 **a** STM topography of the Si(111)(7 × 7) surface unit cell. **b** Schematic of the surface atoms and orientation of the faulted (F) and unfaulted (U) half-unit cells (also used in Fig. 21.14). **c–e** STS spectra measured on the surface atoms at $T = 7\,\mathrm{K}$ and assignment of the Si(111)(7 × 7) surface electronic features to rest atoms [S_2, (**c**)], corner adatoms [S_1', S_1, U_1, (**d**)], and center adatoms [U_1', U_1'', (**e**)]. $V_S = 2\,\mathrm{V}$, and (stabilization) tunneling current before the feedback is switched off at $I_t = 0.1\,\mathrm{nA}$ (reproduced with permission from [51])

the tip was approached towards the surface by $0.4\,\text{Å}\mathrm{V}^{-1}$ for smaller voltages in order to enhance the signal at lower bias voltages. This also helps to obtain a reasonable sensitivity for the occupied states.

A complementary type of spectroscopic technique is spectroscopic imaging.[1] LDOS maps of the Si(111)(7 × 7) surface are shown in Fig. 21.14. From measurements at various voltages, we obtain an unambiguous assignment of the spectral features in Fig. 21.13c–e. Figure 21.14a–d shows LDOS maps of the surface electrons with ascending energy. The lowest energy dangling bonds belong to rest atoms at −1.3 V (Fig. 21.14a), in accordance with the position of the S_2 state in the point spectra. These dangling bonds are fully occupied. Adatom dangling bonds have a higher energy. Corner adatom dangling bonds are imaged at −0.5 V (Fig. 21.14b) and at +1.4 V (Fig. 21.14c) in accordance with the point spectra , while at +1.6 eV the electronic states of the center adatoms are imaged (Fig. 21.14d). This set of images

[1] In this case, for spectroscopic imaging the STM feedback was switched off, and the measurement of the DOS was performed during scanning the tip at a plane parallel to the surface, yielding a map.

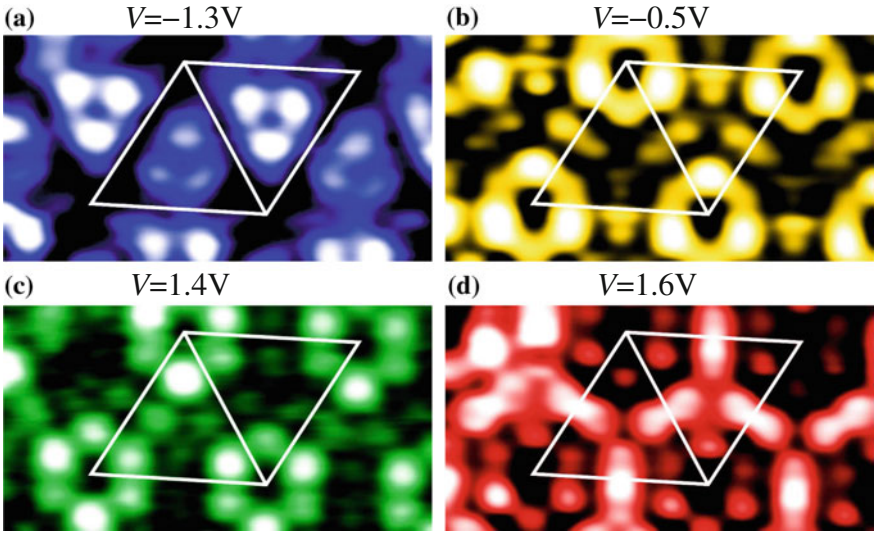

Fig. 21.14 By acquiring the normalized conductance $(dI/dV)/(I/V)$ at each point of an image details of the electronic structure of the Si(111)(7 × 7) surface can be measured. This was done for four different voltages shown in (**a**) to (**d**). At −1.3 V the electronic states of the rest atoms are imaged, at −0.5 and +1.4 V the states of the corner adatoms are imaged, while at +1.6 eV the electronic states of the center adatoms are imaged. The measurements were performed at 7 K

shows clearly that STS is capable of imaging individual localized electron states, both their density distributions in space and their energy levels.

21.12 Summary

- Scanning tunneling spectroscopy allows to obtain spectroscopic data with atomic resolution. In the simplest approximation, the local density of states of the sample is proportional to the differential conductance dI/dV.
- In experiments the derivative of the I-V curve is measured using a modulation technique. The nth derivative of the I-V curve at voltage V is proportional to the AC amplitude of the current signal at n-times the modulation frequency.
- The desired signal of the LDOS is often buried by a large background resulting from the transmission factor. Using the normalized differential conductance $(dI/dV)/(I/V)$ captures the LDOS better than dI/dV.
- In a more rigorous treatment the differential conductance is related to the sample density of states $\rho_{\text{sample}}(eV)$ (for a constant tip density of states ρ_{tip}) as

$$\frac{dI}{dV}\frac{\hbar}{4\pi e} = e\rho_{\mathrm{tip}}\rho_{\mathrm{sample}}(eV)T(eV,V) + \int_0^{eV} \rho_{\mathrm{tip}}\rho_{\mathrm{sample}}(\varepsilon)\frac{\partial T(\varepsilon,V)}{\partial V}d\varepsilon. \tag{21.26}$$

- At tunneling voltages of several volts, the energy resolution in STS is given by the energy width of the transmission factor (about 0.3 eV). If small voltages in the millivolt range and low temperatures are considered, the energy resolution is limited by the thermal broadening of the Fermi function and the amplitude of the modulation voltage.
- From measurements of the tunneling current as a function of tip-sample distance (with feedback switched off), the apparent barrier height $\bar{\Phi}$ can be extracted.
- If the bias voltage is larger than the apparent barrier height, standing electron waves occur in the barrier for particular energies. This results in oscillations in the dI/dV signal.
- In spectroscopic imaging, maps of the local density of states (LDOS) at a certain energy are obtained by performing a slow constant current (topographic) STM scan at a certain bias voltage V and simultaneously recording the (normalized) dI/dV signal.

Chapter 22
Vibrational Spectroscopy with the STM

Vibrational spectroscopy provides a fingerprint of the identity of molecular species. A molecule at a surface can be identified by a very characteristic set of vibrational modes. Vibrational spectroscopy at surfaces was performed in the past with spatially averaging techniques using light (infrared spectroscopy and Raman spectroscopy) or electrons (electron energy loss spectroscopy EELS) in order to excite the vibrations [52].

Also the method of inelastic tunneling spectroscopy (IETS) was used at planar buried interfaces long before the invention of the STM. Here two metal electrodes are separated by an oxide barrier. When electrons tunnel through this barrier they can excite vibrations. In this way, vibrations of molecular species present at the electrode—tunneling barrier interface can be measured.

Soon after the invention of the STM, it was proposed that similar mechanism of vibrational excitation should be possible in STM. Here, the two metal electrodes are replaced by the tip and sample, and the oxide layer by a tip-sample vacuum gap.

Using inelastic scanning tunneling spectroscopy (IETS) it is possible to excite vibrations of a single atom or molecular species. The usual energy range for observed molecular vibrations ranges from several meV up to several hundred meV. If we remember that the energy resolution in scanning tunneling spectroscopy is only in the order of 80 meV at room temperature, operation at low temperatures is necessary in order to achieve the required energy resolution of a few meV in scanning tunneling microscopy IETS.

22.1 Principles of Inelastic Tunneling Spectroscopy with the STM

The principle of inelastic spectroscopy with STM is illustrated in Fig. 22.1a, b. The tip is positioned over the molecule, the z-feedback is disabled, and the bias voltage between tip and sample is ramped. In the limit of low bias voltages, assuming a constant transmission factor and a constant density of states (Tersoff-Hamann

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_22

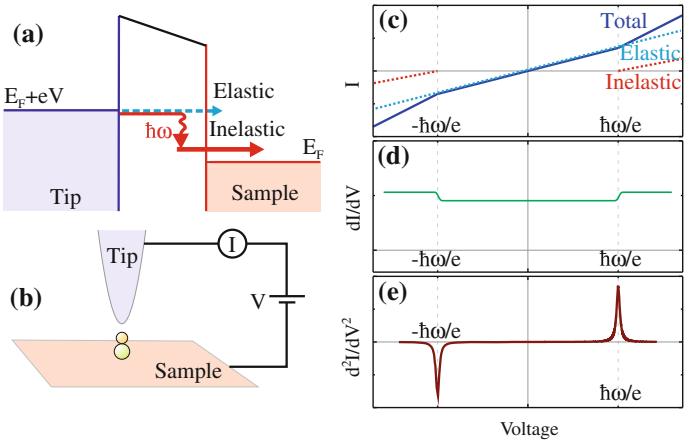

Fig. 22.1 **a** Schematic potential diagram of the STM illustrating the principle of inelastic spectroscopy. Molecular vibrations are excited by the inelastically tunneling electrons. The inelastic channel opens when the bias voltage exceeds the threshold energy of the vibrational mode $\hbar\omega/e$. **b** Schematic of an inelastic STS measurement probing the vibrational spectrum of a molecule at the surface. **c** Effect of the inelastic process on the current-voltage characteristics. The inelastic channel opens at the threshold voltage $V \geq \hbar\omega/e$. **d** This results in a step-like increase of the differential conductance in the dI/dV curve. **e** In the d^2I/dV^2 spectrum, a characteristic peak and dip signature appears symmetrically around zero bias, at bias voltages corresponding to the energy of vibrational mode

approximation), the current is a linear function of voltage. The elastic tunneling channel is indicated in Fig. 22.1a by a horizontal arrow. The excitation of a molecular vibration only becomes possible when the energy of the tunneling electrons exceeds the energy of the vibrational mode. In this case, the tunneling electron can excite a vibration of an atom or molecule residing on the (metal) surface. This tunneling channel is called the inelastic channel. A minimum energy amount of is needed in order to excite a quantized vibration of $\hbar\omega$ as

$$e\,|V| \geq \hbar\omega. \tag{22.1}$$

When the inelastic electron tunneling channel opens above the threshold voltage, a slight increase in the tunneling current occurs because the tunneling current now receives contributions from two channels: the elastic and the inelastic channel Fig. 22.1b. Since most electrons still tunnel elastically, only a small inelastic current is added (usually a few percent). This inelastic current is also proportional to the voltage and added to the elastic current Fig. 22.1c. This slight increase in the slope of the I-V curve is usually too small to be detected directly. The increase of the slope in the I-V curve leads to a (small) jump in the dI/dV signal at the threshold voltage. Even this change in the dI/dV is usually too small to be detected, so information about vibrational modes is extracted from the d^2I/dV^2 signal measured by the lock-in technique (Chap. 6 and Sect. 21.2), where a peak appears at the threshold

voltage. The effect of the opening of the inelastic channel on the I-V, dI/dV and d^2I/dV^2 spectra is shown schematically in Fig. 22.1c–e.

There is an additional specific signature in the vibrational spectra which is used to confirm that a specific feature in the second derivative is really present due to vibrational excitations at this energy. The vibrational mode can be excited by electrons tunneling in both directions, from the tip to the sample and from the sample to the tip. This implies an important feature of the STM vibrational spectrum. A dip occurs in the negative bias voltage, (anti) symmetrically to the peak in the positive bias. In this way, vibrational spectra of single molecules at surfaces can be identified unambiguously, as shown schematically in Fig. 22.1c.

22.2 Examples of Vibrational Spectra Obtained with the STM

The second derivative of the I-V curve is measured using the lock-in technique. Even with the lock-in technique the measured vibrational spectra of molecules tend to be quite noisy, as seen in the example shown in Fig. 22.2. Here a vibrational spectrum of C_{60} (measured using the lock-in technique) is shown. Several peaks can be observed in this spectrum. However, a clear signature of a peak and a corresponding dip at the same negative voltage is only observed in two cases, at 53 and at 138 meV. Due to this signature the vibrational mode is identified unequivocally in spite of the noise present in the d^2I/dV^2 spectrum shown in Fig. 22.2. Also the comparison with the reference spectrum taken on the clean metal surface helps to identify the vibrational peaks. If peaks are present on the bare surface as well as above the molecule, these peaks may not be vibrational peaks of the molecule but induced by the substrate. Therefore, it is always important to measure a reference spectrum close to the molecule under study.

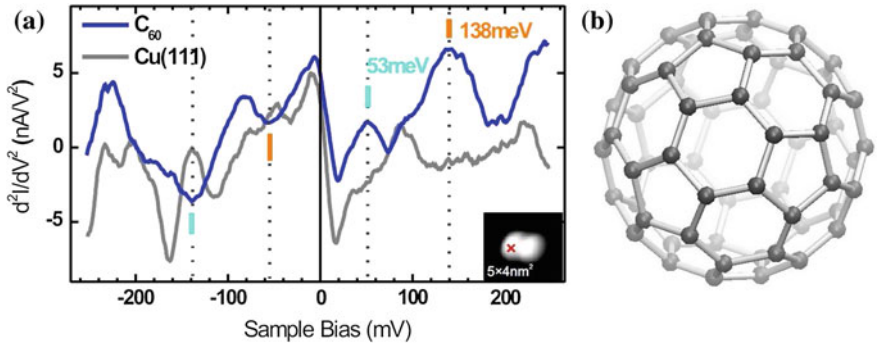

Fig. 22.2 **a** STM-IETS spectrum measured for C_{60} on Cu(111). Vibrational spectra are often noisy and a characteristic peak/dip structure at positive/negative voltages can be used to recognize the vibrational peaks. The spectrum taken on the clean surface is shown for comparison. ($I = 0.1$ nA, $V = 250$ mV, $V_{\mathrm{mod}} = 6$ mV RMS). **b** Schematic representation of a C_{60} molecule

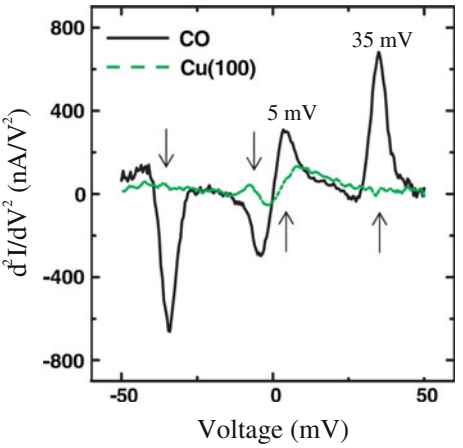

Fig. 22.3 STM-IETS spectrum measured for CO on Cu(100). The inelastic tunneling process occurs in both directions at positive and negative tunneling voltages. This leads to a clear signature of a symmetric peak and corresponding dip for positive and negative voltages, respectively (reproduced with permission from [53])

For instance, the prominent structure at −240 meV in Fig. 22.2 is clearly identified as due to the substrate.

In Fig. 22.3, the vibrational spectrum (d^2I/dV^2) of a CO molecule is shown. The green line shows a reference spectrum on the clean Cu(100) substrate close to the molecule. For the CO, two pairs of antisymmetric peak/dip are observed at 35 and 5 meV. The two vibrations correspond to the CO bending vibrational mode and the frustrated translation, respectively. One disadvantage of the vibrational spectroscopy performed with the STM is that usually only very few of the vibrational modes of a molecule can be measured. For instance, in the case of the CO molecule only two of the total of six vibrational modes are observed. Also the selection rules of STM vibrational spectroscopy are not very simple. The great advantage of vibrational spectroscopy with an STM is the ability to measure vibrational excitations with ultimate spatial resolution, i.e. at a single atom or molecule. The vibrational spectra carry chemical information, which is lacking in STM imaging or dI/dV scanning tunneling spectroscopy. With STM vibrational spectroscopy it is possible to resolve aspects of the local bonding of molecules at surfaces and modifications of intermolecular interactions due to surface bonding. When single molecule reactions are induced, STM-IETS can be used to characterize the nature of the reaction products.

Another way to identify d^2I/dV^2-peaks clearly as vibrations is the isotope test. If one chemical element in a molecule is replaced by an isotope with a different mass, the energies of the corresponding vibrations change. If for instance in a molecule the hydrogen atoms are replaced by the heavier deuterium isotope, the vibrational energies are reduced. If only the stretching vibration of the hydrogen (deuterium) is considered, the frequency shifts can be understood using a simple spring model. The vibration frequency ω as function of the spring constant k and the mass of the atom m is

$$\omega = \sqrt{\frac{k}{m}} \propto \sqrt{\frac{1}{m}}. \tag{22.2}$$

Since the chemical bonds are the same for different isotopes the atomic spring constant k stays the same for different isotopes. Therefore, the frequency (and correspondingly the energy) of the hydrogen stretch vibration should be reduced by a factor of $\sqrt{2}$ when hydrogen is replaced by deuterium in a molecule.

An example of this is shown in Fig. 22.4. The bonding configuration of a C_2H_2 molecule on Cu(100) is shown in Fig. 22.4a. With the STM image in Fig. 22.4b alone it is impossible to distinguish between the different isotopes present on the surface. An STM d^2I/dV^2 spectrum is shown in the upper trace in Fig. 22.4c. To prove that the peak observed at 358 meV corresponds to the C-H stretch vibration, an isotope was used in which the hydrogen was replaced by deuterium. The corresponding spectrum is shown as the middle trace in Fig. 22.4c displaying an energy reduction of the peak from 358 to 266 meV corresponding approximately to a factor of $\sqrt{2}$. If in a C_2H_2 molecule only one hydrogen atom is replaced by deuterium, the lower spectrum in Fig. 22.4c results, which shows both peaks. With this information the molecules in Fig. 22.4b can be individually identified as C_2H_2, C_2D_2 and C_2HD, respectively as indicated. If the isotope exchange is possible, it is a very useful test, since also electronic states lead to structures (peaks and dips) in the d^2I/dV^2 spectrum. Such peaks do not shift upon isotope exchange, since the electronic structure is not changed by the higher mass.

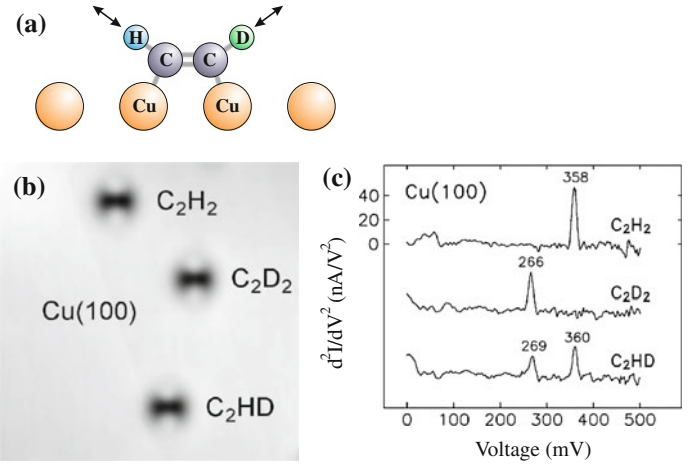

Fig. 22.4 **a** Schematic of the bonding of a C_2H_2 molecule on a Cu substrate. The C-H stretch vibrations are indicated by *arrows*. **b** By the STM image alone the C_2H_2 molecule cannot be distinguished from the C_2D_2 and the C_2HD isotopes. **c** Using STM vibrational spectroscopy, the different molecules can be identified via their corresponding vibrational energies (∼360 meV for the C-H stretch and ∼269 meV for the D-H stretch) (reproduced with permission from [54])

22.3 Summary

- In inelastic tunneling spectroscopy with the STM (STM-IETS) vibrations of molecules in the tunnel junction are excited.
- In addition to the elastic channel an inelastic channel for the tunneling current is opened if the tunneling voltage exceeds the energy of the vibration $\hbar\omega$.
- A vibrational excitation leads to a peak at $\hbar\omega$ and a corresponding dip at $-\hbar\omega$ in the d^2I/dV^2 signal.

Chapter 23
Spectroscopy and Imaging of Surface States

The metals Cu, Ag and Au have surface states at their low index surfaces with energies around the Fermi level. On the (111) surfaces the surface state band is parabolic around the center of the surface Brillouin zone ($k_{||} = 0$). Figure 23.1 shows schematically the energy as function of the wave vector parallel to the surface at the Cu(111) surface. The projected bulk band is shown as a shaded area in Fig. 23.1, while the surface state is shown as a green line (cf. Chap. 10).

These surface states, which extend over the whole surface, are sensitive to defects on the surface. Steps, islands or adatoms act as potential barriers for the surface state wave functions. Because of the confinement of the wave functions at these potential barriers, the eigenstates become standing waves. The wave functions of the surface states can be probed by scanning tunneling microscopy and spectroscopy with high spatial resolution as well as high energy resolution.

23.1 Energy Dependence of the Density of States in Two, One and Zero Dimensions

Quasi-free electrons obey a parabolic band dispersion, as

$$E = \frac{\hbar^2 \boldsymbol{k}^2}{2m}, \tag{23.1}$$

where E is the energy relative to the bottom of the band E_0, and m is the effective mass of the electron. The energy dependence of the density of states can be calculated by integrating the states in k-space up to a certain energy. Due to the different integration conditions for the k-space of different dimensionality, different dependences of the density of states result for different dimensionalities [55] as

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_23

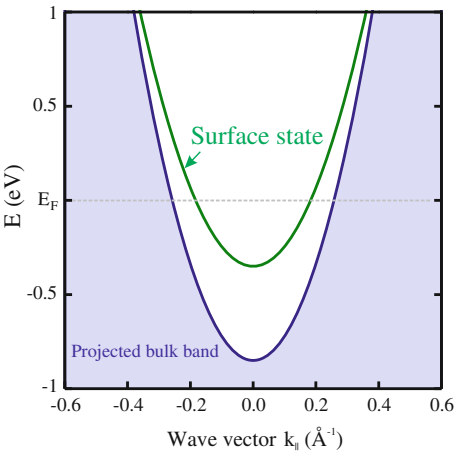

Fig. 23.1 Energy as a function of the wave vector parallel to the surface at the Cu(111) surface. The surface state is located outside the projected bulk band

$$\rho(E) \propto \sqrt{E} \quad \text{in 3D} \tag{23.2}$$

$$\rho(E) = const. \quad \text{in 2D (for every subband)} \tag{23.3}$$

$$\rho(E) \propto \frac{1}{\sqrt{E}} \quad \text{in 1D (peak for every subband)} \tag{23.4}$$

$$\rho(E) = \delta(E - E_i) \quad \text{in 0D (peak for each state)} \tag{23.5}$$

The above given dependence of the density of states $\rho(E)$ is shown graphically in Fig. 23.2 for different dimensions.

In the direction in which the dimensionality is reduced by confinement of the electrons, modes of different energy occur known as subbands. The occurrence of

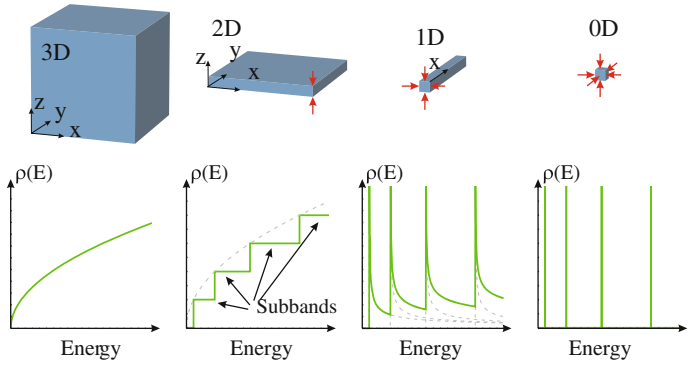

Fig. 23.2 Energy dependence of the density of states for a free electron gas of different dimensionality

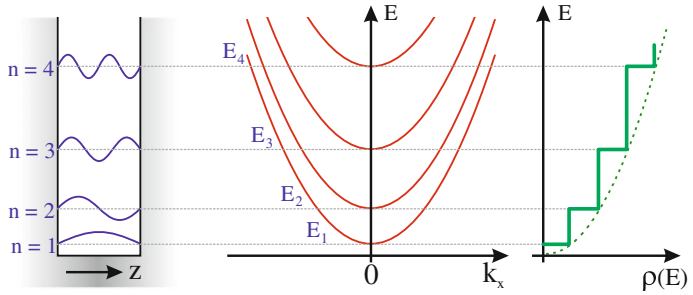

Fig. 23.3 In the 2D confinement, different parabolic subbands occur due to different modes $n = 1, 2, 3, \ldots$ of the wave functions existing in the confined direction

subbands with parabolic dispersion, in directions in which no confinement occurs, is shown in Fig. 23.3.

As shown in Chap. 21, the density of states at the surface can be probed by scanning tunneling spectroscopy. In the first approximation the density of states can be expressed as

$$\rho_{\mathrm{surf}}(eV) \propto \frac{dI}{dV}. \tag{23.6}$$

Experimental data for the energy dependence of the density of states obtained with scanning tunneling spectroscopy (STS) are shown in Fig. 23.4 for different dimensionalities. Figure 23.4a shows the dI/dV signal, corresponding to the 2D density of states of the surface states (Fig. 23.1) on the Cu(111) surface. At the onset of the surface state (−460 meV), an abrupt jump in the density of states is observed in accordance with the expected step function for the 2D density of states as displayed in Fig. 23.2. At step edges there are sometimes states which extend only along the step edges (1D states). The dI/dV spectrum taken at a step edge on an InAs surface is shown in Fig. 23.4b compared to the spectrum on the terrace [56]. For the spectrum at the step edge, peaks with a sharp onset and a characteristic $1/\sqrt{E}$ are observed (as shown by the turquoise line in Fig. 23.4b), indicative of the density of states of 1D states. On the terrace, the dI/dV spectrum does not show these peaks. In Fig. 23.4c the dI/dV spectrum taken on a small (∼6 nm diameter) InAs cluster shows sharp peaks indicative of a 0D density of states corresponding to quantized electron energies [57]. Due to the confinement in all three spatial directions, these clusters are also often called artificial atoms or quantum dots. An STM image of one InAs cluster is shown in the inset in Fig. 23.4c. In order to obtain a good energy resolution, all these measurements were performed at liquid helium temperatures.

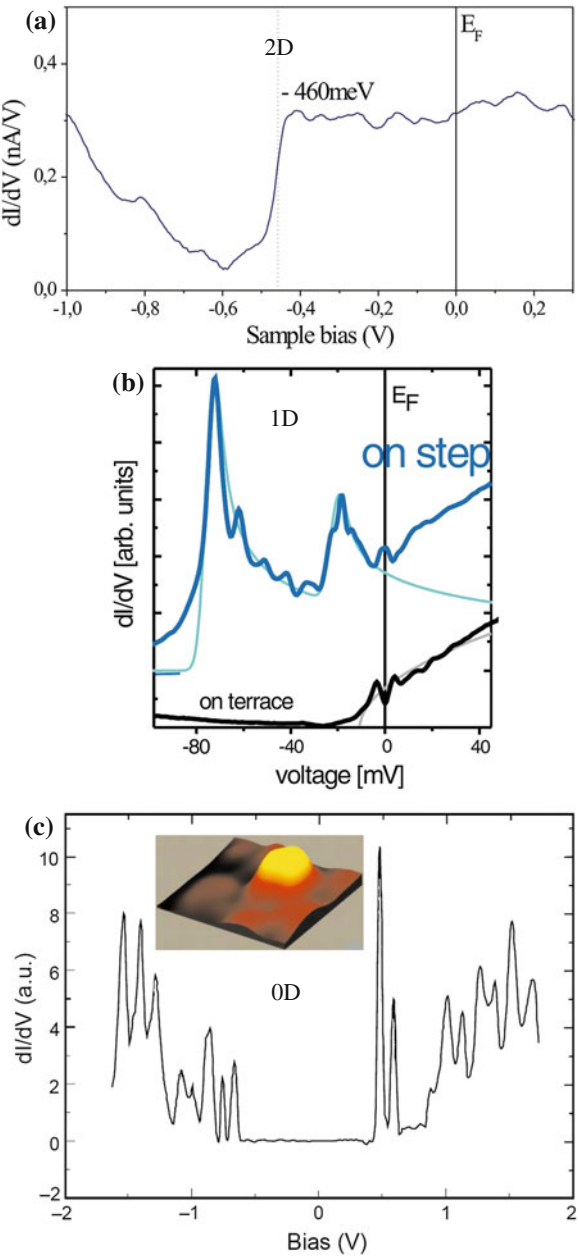

Fig. 23.4 Density of states (DOS) measured via dI/dV spectra. **a** A 2D Cu(111) surface state leads to a step-like increase in the density of states beyond the onset energy of the surface state at −460 meV. **b** Two 1D states at the step edge on an InAs surface with the characteristic $1/\sqrt{E}$ energy dependence (reproduced with permission from [56]). **c** Spectrum taken on an InAs cluster shows peaks for the individual states of the quantum dot (reproduced with permission from [57]). The inset shows an STM image of an InAs quantum dot (image size 10 nm × 10 nm). Each of the DOS spectra in (**a**)–(**c**) shows the energy dependence characteristic of its dimensionality as shown schematically in Fig. 23.2

23.2 Scattering of Surface State Electrons at Surface Defects

Apart from the energy dependence also the spatial dependence of the local density of states at surfaces can be studied. On clean low-index (especially (111)) metal surfaces the atomic corrugation is usually very small in STM images. This means that the lattice periodic modulation function $u_k(r)$ of the (Bloch) wave is nearly a constant with a lattice periodic modulation of only a few percent. At defects a different electrostatic potential can lead to an enhanced or decreased amplitude of the wave function around a defect. STM and STS are ideal tools to study such wave functions of electrons confined at defects such as steps or adatoms.

If electron waves of a 2D electron gas are confined at defects, a traveling wave is no longer a solution compatible with this boundary condition. First we consider the case of a step edge where we assume that the wave function is confined by *one* boundary condition. We assume in the following the extreme case that the amplitude of the wave function goes to zero at the boundary. In this case, a standing wave is a solution of the Schroedinger equation. However, since the confinement is only at one side, the states still have a continuous range of energies. Only if the electrons are confined by two boundary conditions do standing waves with quantized energies and wave vectors result.

In a 1D model incoming and reflected waves can be considered as

$$\psi_1(x) = u_k(x)e^{ikx} \tag{23.7}$$

and

$$\psi_2(x) = u_k(x)e^{-ikx}. \tag{23.8}$$

Superposition results in a stationary solution of a standing wave. If we assume as a simple and extreme boundary condition $\psi(x = 0) = 0$ so that the wave function vanishes at the defect site (step edge), the difference of the wave functions above satisfies this boundary condition.

$$\psi_{1-2}(x) = \frac{1}{\sqrt{2}}u_k(x)(e^{ikx} - e^{-ikx}) = -i\sqrt{2}u_k(x)\sin(kx). \tag{23.9}$$

For the absolute square of this wave function the factor $\sin^2(kx)$ induces a modulation from zero to one, instead of the small modulation induced by $u_k(x)$. Such standing waves in the 2D electron states at surfaces are observed in STM, either directly in the topographic images, for instance at step edges (Fig. 23.5a), or more pronounced in dI/dV maps such as the line scans shown in Fig. 23.5b. For the case of the dI/dV maps, the signal is approximately proportional to the density of states at the voltage eV. In a simple analysis, the component of the k-vector along the surface can be obtained from a measurement of the the wavelength λ of the standing wave oscillations in Fig. 23.5b as $k_{||} = 2\pi/\lambda$. Plotting the energy as a function of $k_{||}$ results in a dispersion relation.

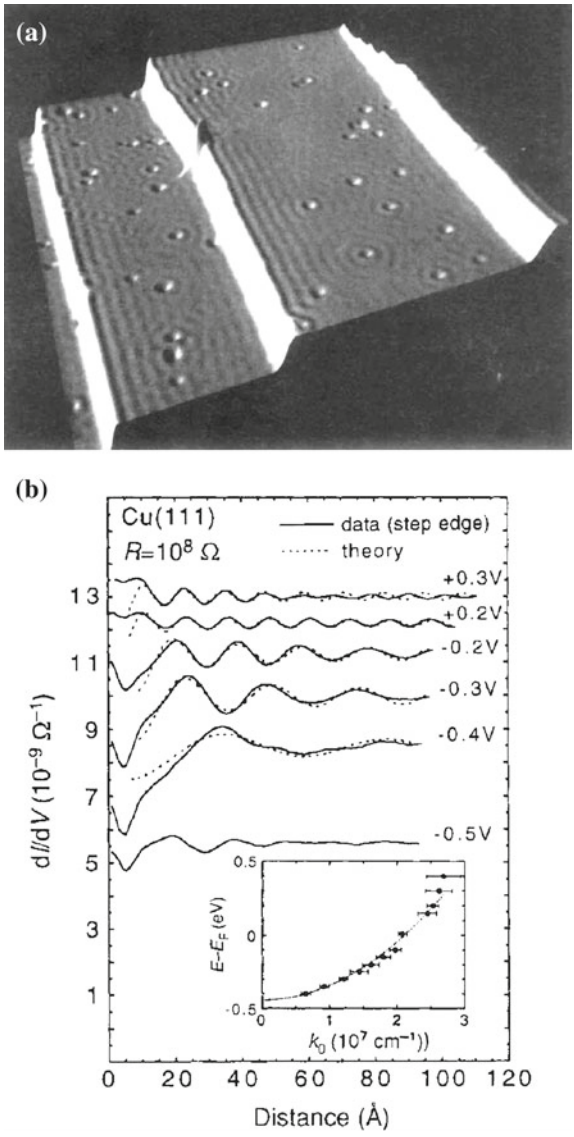

Fig. 23.5 **a** Constant current image of the Cu(111) surface (image size 500 Å × 500 Å, $V = 0.1$ V, $I = 1.0$ nA). Three monatomic steps and several point defects are visible. Spatial oscillations at the step edges and around defects with a periodicity of ∼15 Å are clearly visible. **b** Spatial dependence of dI/dV, measured as a function of distance from the step edge for different bias voltages (*solid lines*). The inset shows the dispersion of the surface state as obtained from the experimental data (reproduced with permission from [58])

In a more advanced analysis the density of states can be calculated for the 2D case. The following density of states is found [58]

$$\rho(E, x) = \rho_0 \left(1 - J_0(2k_{||}x)\right), \tag{23.10}$$

where J_0 is the zero order Bessel function, ρ_0 is the density of states of a 2D electron gas in the absence of any scattering, and $E = \frac{\hbar^2 k_{||}^2}{2m}$. For larger distances from the step edge, it follows from (23.10) that the density of states decays $\sim 1/\sqrt{x}$. This is a result which is qualitatively different from the 1D case, where the amplitude does not decay with the distance from the step edge. As in the case of the pure energy dependence of the density of states shown in Fig. 23.2, also here the integration in the k-space leads to qualitatively different results for a different dimensionality of the problem (1D or 2D).

In Fig. 23.5b, the spatial dependence of dI/dV is shown as function of the distance from the step edge and compared to the results obtained by the model outlined above (dotted lines) [58]. A value of the effective mass m can be obtained from a fit of (23.10) to the dispersion relation (dotted line in the inset of Fig. 23.5b). An effective mass $m = 0.38\, m_e$ is obtained for the electrons in the 2D surface state on Cu(111).

Step edges are not the only type of defects at which surface state electrons can be confined or scattered. Scattering at point defects is observed in Fig. 23.5a. The scattering at point defects can be considered by taking an incoming and a scattered wave into account. In the isotropic case s-wave scattering occurs with the local density of states decreasing with $1/r$ as a function of the distance from the point defect [58]. In a more general treatment, also a scattering phase and a scattering amplitude are included in the model.

23.3 Summary

- The characteristic signature of the energy dependence of the density of states in 2D, 1D and 0D is measured in STS.
- Step edges and adatoms act as potential barriers and scatterers for electrons in 2D surface states, leading to standing electron waves at surfaces.

Chapter 24
Building Nanostructures Atom by Atom

The scanning tunneling microscope, initially used to image surfaces down to the atomic scale, has been further developed into an operative tool, with which atoms and molecules can be positioned at will in order to create and investigate artificial structures. In this chapter, we will see that two-dimensional quantum systems can be built and modified, exploiting the ability of the STM to move atoms and molecules on the surface. The building of dedicated potential barriers on the atomic scale results in a playground to explore the wave nature of electrons. Such experiments are usually performed at low temperatures (for instance liquid helium), because at room temperatures single atoms would diffuse quite rapidly on metal surfaces. Furthermore, the tunneling current can be used to selectively break chemical bonds, but also to induce chemical bonds. These possibilities give rise to new opportunities to study chemistry on the level of the single atom and single molecule.

24.1 Positioning of Single Atoms and Molecules by STM

The procedure for lateral positioning of single atoms is illustrated schematically in Fig. 24.1. In the STM imaging process the tip is scanned at distances of a few atomic diameters above the surface (a) and follows contours of constant local electronic density of states (in the constant current mode). For the movement of atoms the tip is brought close to the atom to be moved. Subsequently, the tip-sample distance is reduced from the imaging distance (a) to the positioning distance (b). This reduced tip-sample distance brings the tip into stronger interaction with the atom/molecule to be moved. This stronger interaction can be either attractive or repulsive. Then the tip is moved parallel to the surface either in constant current or in constant height mode (c) to the predetermined place whereby the atom or molecule is pulled (or pushed) to the desired location (d). The tip is then withdrawn to the scanning distance, i.e. at a distance where no motion of the atom is induced (e). Then a new STM image is scanned to check the result of the atom positioning. A threshold tunneling resistance

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0_24

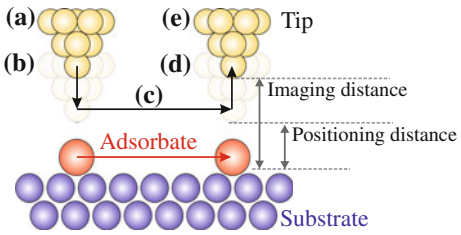

Fig. 24.1 (*a*) Principle of the lateral positioning of single atoms by the STM tip. The tip-adsorbate distance is decreased from the imaging distance to the positioning distance at which stronger tip-sample forces are present

corresponding to a certain tip height and correspondingly to a certain force on the atom/molecule is necessary to move the atoms across the surface.

The movement of the atom/molecule is driven by the tip-adsorbate interaction, i.e. either attractive chemical forces, or repulsive forces. More information about the nature of this interaction can be gained by recording the response of the STM z-feedback (in constant current positioning mode) or the tunneling current (in constant height positioning mode). Depending on the type of forces involved between the moved atom/molecule and the tip, two different positioning mechanisms are observed: pushing and pulling. The processes are shown schematically in Fig. 24.2a, b together with the characteristic plots of the tip height Fig. 24.2c, d recorded in the constant current mode during positioning.

Pulling takes place when the adsorbate experiences an attractive interaction with the tip. If the tip is initially directly above the molecule no sidewise force acts. When it is moved to the side (in constant current mode) a lateral force builds up. At a certain sidewise motion, the lateral tip-sample force overcomes the threshold for motion to the next adsorption site. The molecule follows the tip by hopping from one adsorption site to the next, associated with an instantaneous upward jump of the tip which can be observed in the height trace in Fig. 24.2c.

The *pushing* mechanism involves a repulsive interaction between the adsorbate and the tip. In the initial part, the tip starts to move up the contour of the molecule. At some point, due to the increasing repulsive forces, the atom or molecule jumps forward, which is seen as a sudden decrease in the tip height in Fig. 24.2d.

As an example, the lateral positioning of a single Pb atom along an intrinsic Cu(211) step edge is shown. The force which induces the positioning of the Pb atom is attractive, as will be explained now. Figure 24.3a shows the height of the tip during the actual positioning process while keeping the tunneling current constant. When positioning a Pb atom a low tunneling resistance of 120 kΩ (corresponding to a closer tip-sample distance than during imaging) is used and the tip is moved from left to right. The jumps in the tip height clearly indicate a discontinuous upward movement of the tip height, corresponding to the signature of the pulling mode. This indicates that there is an attractive force between the tip and the atom during the positioning process. Since the force between a metal tip apex and a metal atom is always attractive in the tunneling regime, one has to use a different adsorbate to show the effect of repulsive forces.

Here we show an example where the positioning is performed in constant current mode to precisely position single C_{60} molecules on Cu(111). The molecules were

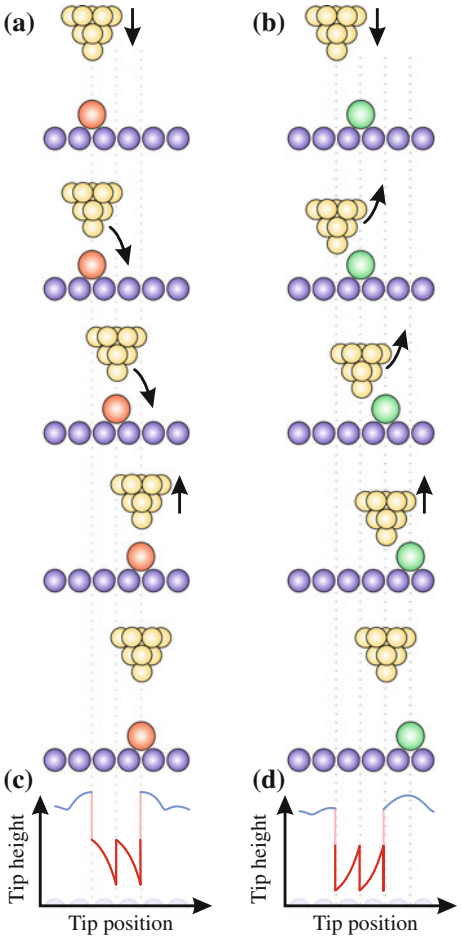

Fig. 24.2 Principle of the pulling mechanism (**a**) and the pushing mechanism (**b**). Corresponding tip height profiles during constant current positioning in the pulling (**c**) and pushing mode (**d**). The *blue lines* correspond to traces at the imaging distance, while the *red* traces correspond to the positioning distance

deposited on Cu(111) at a temperature of $T = 25\,\mathrm{K}$, in order to obtain single C_{60} molecules on the substrate instead of larger agglomerates. To move the molecule, the tip was brought close to the sample, decreasing the resistance of the tunneling junction to 0.5 MΩ. Subsequently, the tip was moved laterally towards the molecule at constant current. The positioning of individual C_{60} is shown in Fig. 24.3c, d. The molecules were moved along the surface in a controlled way. Characteristic tip height plots found for C_{60} moved on Cu(111) are shown in Fig. 24.3e, f. The characteristic shape of the curves indicates that the molecule has been moved in the pushing mode. Figure 24.3e shows the tip height plot during repositioning of the C_{60} along the [110] direction of the Cu(111) substrate. The length of a single jump (2.5 Å) is close to the nearest neighbor distance on Cu(111). Figure 24.3f shows the tip trajectory during positioning along the [211] direction. The average hopping distance corresponds to the distance between two equivalent adsorption sites along [211] direction. The model of Cu(111) surface with crystal directions is shown in Fig. 24.3g. The distances

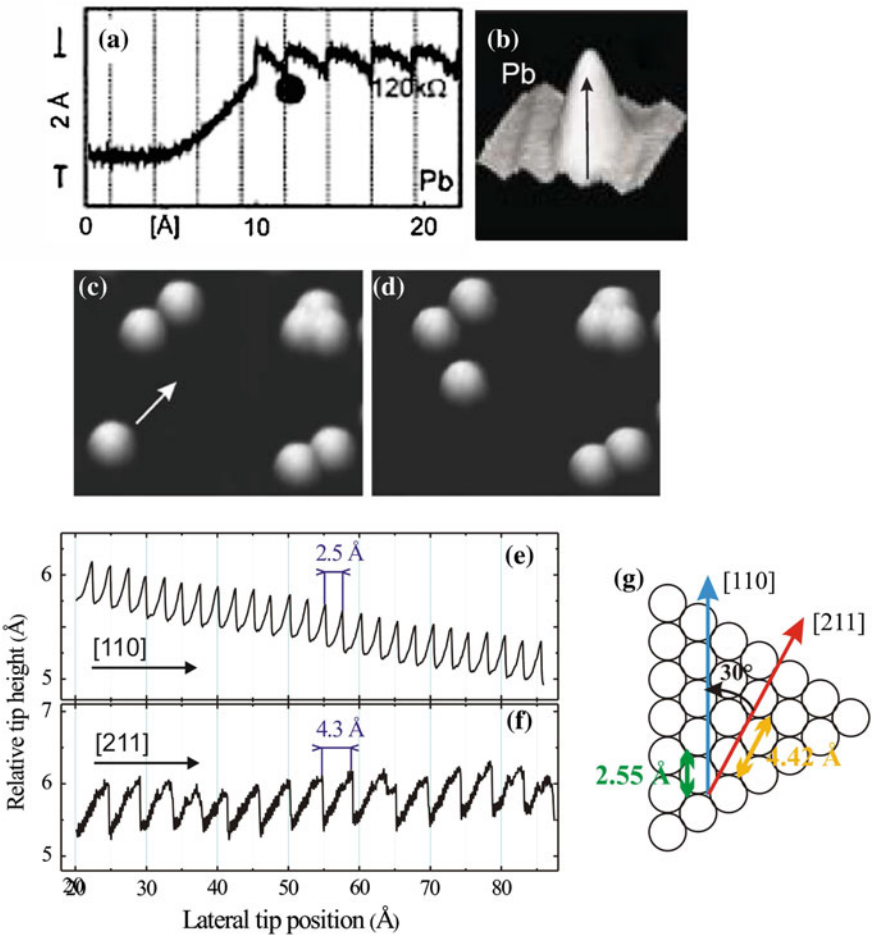

Fig. 24.3 **a** Characteristic height trace observed when pulling a Pb atom (reproduced with permission from [59]). **b** Shows a 3D representation of an STM image of a Pb atom on the copper surface. Positioning of a single C_{60} on a Cu(111) surface: **c** before and **d** after the positioning process (reproduced with permission from [60]). Displacement of the molecule is indicated by an *arrow*. Positioning was performed in constant current mode. **e** and **f** show records of the tip height during the positioning of a single C_{60} molecule on Cu(111) along different crystallographic directions. The shape of the tip height trace shows that the positioning occurred due to the pushing mechanism. **g** Model of the Cu(111) surface with crystal directions along which the motion was performed

between equivalent adsorption sites along [110, 211] are indicated in this image. Figure 24.4 shows the word NANO written in the way described above with single C_{60} molecules. Each letter has a height of 15 nm.

Up to now we considered lateral positioning of atoms/molecules along the surface. Another kind of positioning is the vertical positioning. Here an atom/molecule is in a first step transferred from the surface to the tip and in a second step transferred to the surface at another position. In Fig. 24.5a two Xe atoms (smaller circular protrusions)

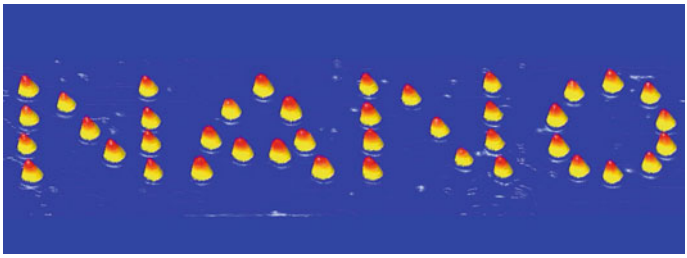

Fig. 24.4 The word NANO assembled from single C_{60} molecules by lateral motion (letter size: $15 \times 15\,\mathrm{nm}^2$)

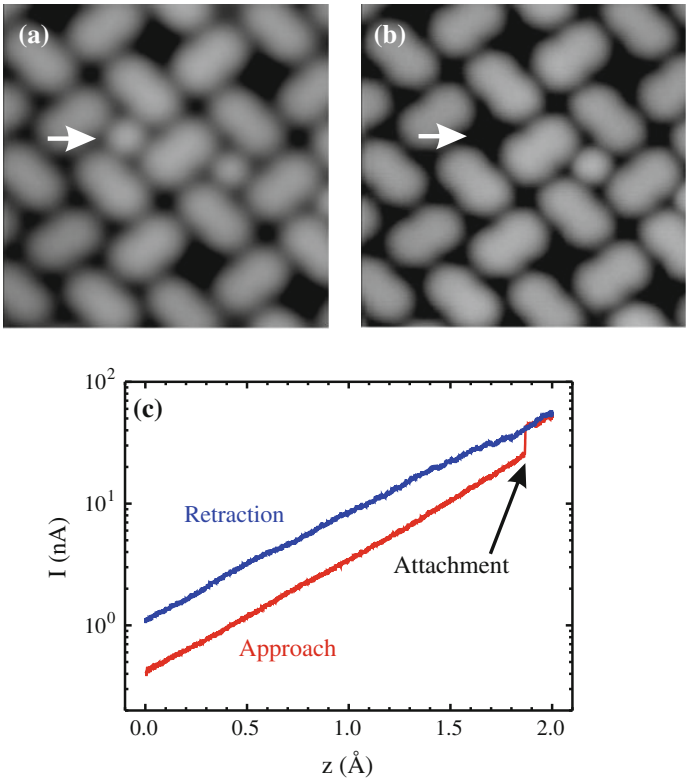

Fig. 24.5 **a** STM image of two Xe atoms embedded in a layer of PTCDA molecules on a Au(111) surface (imaging parameters: $I = 5 \times 10^{-11}$ A, $V = -10$ mV [61]). **b** Same location on the surface after vertical transfer of a Xe atom to the tip. The left Xe is missing. **c** Current distance dependence during approach and retraction. The tip is approached until an instantaneous jump of the current indicates a transfer of the Xe atom to the tip (sample bias voltage: 0.1 V)

are embedded in an ordered array of PTCDA molecules (larger elongated protrusions) [61]. The left Xe atom in Fig. 24.5a is removed by attaching it to the tip. The tunneling current as function of the approach distance is shown in Fig. 24.5c. At a certain approach distance the measured current increases abruptly, which indicates that the Xe atom jumps from the surface to the tip. In this modified junction geometry (Xe on the tip) the tunneling current is larger than with the Xe atom at the surface. This attachment of the Xe atom to the tip can be confirmed by subsequent imaging of the same area on the surface, which shows that the Xe atom was removed (Fig. 24.5b). The image of the molecular layer taken with a Xe atom attached to the tip (Fig. 24.5b) has a more pronounced and sharper contrast than the image taken without the Xe atom. Tips functionalized with an atom/molecule often lead to contrast mechanisms different from the metal tip. The Xe atom can be attached back to the surface at a desired position by approaching the surface with a bias voltage of opposite polarity.

24.2 Electron Confinement in Nanoscale Cages

If electrons are confined completely to a certain region in space this leads to wave functions of the standing wave type with discrete energies and wave vectors. Using the ability of the scanning tunneling microscope to move single atoms or molecules along the surface, dedicated structures can be built in order to act as barriers for the surface state electrons. In this way, the surface state electrons can be confined to cage structures of particular shapes. An example of an artificial nanostructure prepared on a copper surface is shown in Fig. 24.6a. Iron atoms are placed on a Cu(111) surface at low temperature (4 K). First the iron atoms are deposited randomly on the Cu surface, then they are moved to the desired positions using the STM tip, as described in the previous section. The STM images in Fig. 24.6b show the steps in the formation of a "quantum corral". Forty eight iron atoms were positioned into a circular ring Fig. 24.6a in order to form a cage for surface state electrons and to force them into quantum states with circular boundaries.

A standing wave pattern of electron waves confined inside the corral can be observed in Fig. 24.6a; similar to the standing wave patterns observed, for instance, on a drumhead. These ripples in the ring of atoms correspond to the density distribution of a particular set of quantum states of the corral. Since the STM probes the electronic wave functions, the standing wave within the corral must be due to electrons located at the Cu surface. As we have seen in Chap. 10, there are electronic surface states which have a free electron character on Cu(111) (parabolic dispersion). Electrons occupying these states are located at the surface and their motion parallel to the Cu(111) surface is essentially that of free electrons. For these electrons, the Fe atoms forming the circular corral are strong scattering centers, such that the electrons are confined by the circular barrier. The spatial variation of the electronic density of states can be described, even quantitatively, by the distribution of round-box eigenstates (Bessel functions) of electrons within Cu surface states near the Fermi level. Depending on

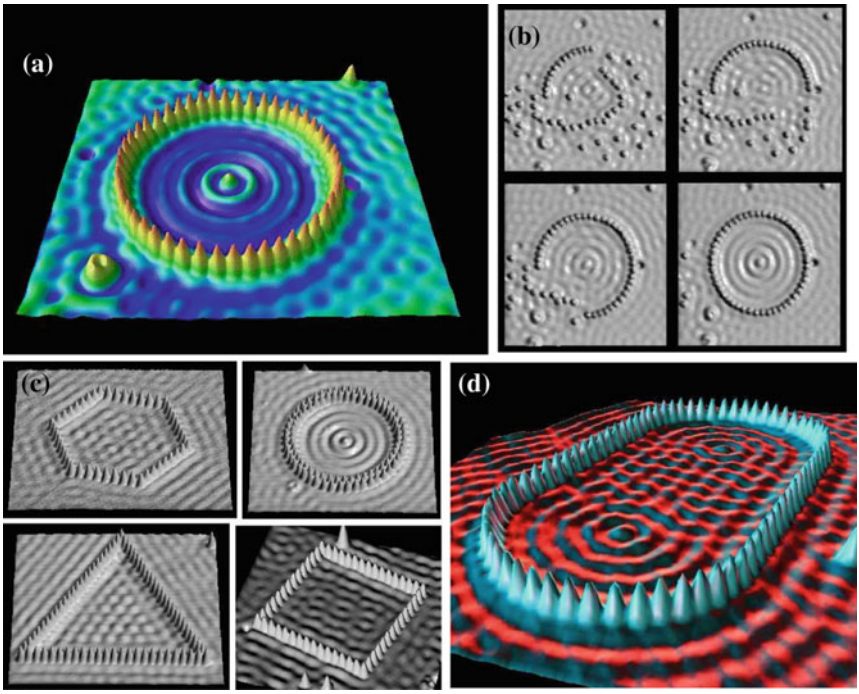

Fig. 24.6 Surface state electrons on Cu(111) are confined to closed structures (corrals) defined by adatoms forming barriers for the electron waves. The barriers were assembled by individually positioning Fe adatoms using the tip of a low-temperature scanning tunneling microscope. **a** A circular corral of radius 71 Å of 48 Fe adatoms was constructed in this way [62]. This STM image shows the direct observation of standing-wave patterns in the local density of states of the Cu(111) surface. These spatial oscillations are quantum-mechanical interference patterns caused by scattering of the two-dimensional electron gas at the Fe adatoms. **b** Formation of the corral structure from single Fe atoms. **c**, **d** Different shapes of the corrals give rise to different patterns of the standing electron waves [62]. (Images originally created by IBM Corporation)

the shape of the outer boundary the wave patterns are quite different, as seen in Fig. 24.6c, d.

Spectroscopic data (dI/dV) as a function of the bias voltage show that the energy levels of the confined electrons are discrete. Good agreement with quantum mechanical calculations was found, as shown in Fig. 24.7. From a detailed analysis of the standing waves of such "quantum corrals" it was determined that Fe atoms reflect about 25 % of the incident wave, while 25 % are transmitted and 50 % are absorbed (scattered into bulk states) [63].

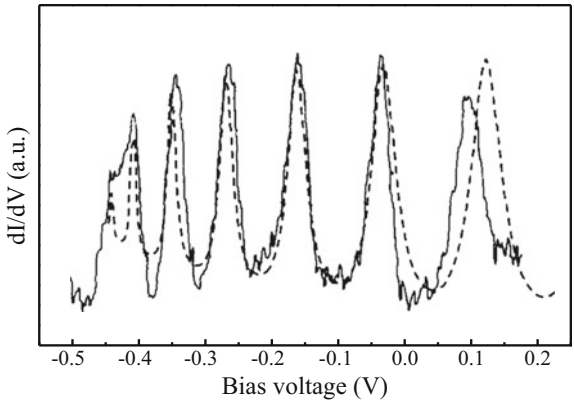

Fig. 24.7 Experimental (*solid line*) and theoretical (*dashed line*) voltage dependence *dI*/*dV* with the tip of the STM located at the center of the circle of Fe atoms on a Cu(111) surface. The peaks show the energies of the confined electron states (reproduced with permission from [63])

24.3 Inducing a Single Molecule Chemical Reaction with the STM Tip

One advanced application of the capability of the STM to position single molecules is inducing a chemical reaction and following all the steps of the reaction on the single molecule level. In the Ullmann reaction, iodine has to be split off from the iodobenzene parent molecules to form the phenyl reactants which then combine to form a biphenyl molecule. Tunneling electrons temporarily populate the iodine-phenyl anti-bonding level, thus causing the dissociation step. Subsequently the iodine is moved away from the phenyl ring by STM positioning (Fig. 24.8a–c).

Both iodine and phenyl fragments are found on the surface at a step edge (d). Subsequently iodine was transferred to the terrace (Fig. 24.8a). To bring two phenyls together, lateral motion in the pulling mode is employed Fig. 24.8e, f. In a chemical reactor working at elevated temperatures, this step would be performed by thermal diffusion. At the low temperatures of the Cu(111) substrate (4 K), the proximity of the two phenyls is not sufficient to induce the association to biphenyl. If a pulling procedure is applied to the phenyl couple from one end, the phenyl at the rear does not move together (not shown in Fig. 24.8). This proves that the two phenyl rings are still separate and no reaction between them has occurred. Only after the injection of electrons by the STM tip is the synthesis step performed (Fig. 24.8f), which can be proven by pulling the product from one end and observing that the entire molecule follows the tip (Fig. 24.8g, h). The synthesis of the two phenyls to biphenyl is probably enabled by the local excitation of vibrational modes in the phenyl groups, enabling the two open bonds to find the proper relative orientation for bond formation. This is an example of a single molecule reaction induced and followed in detail by STM.

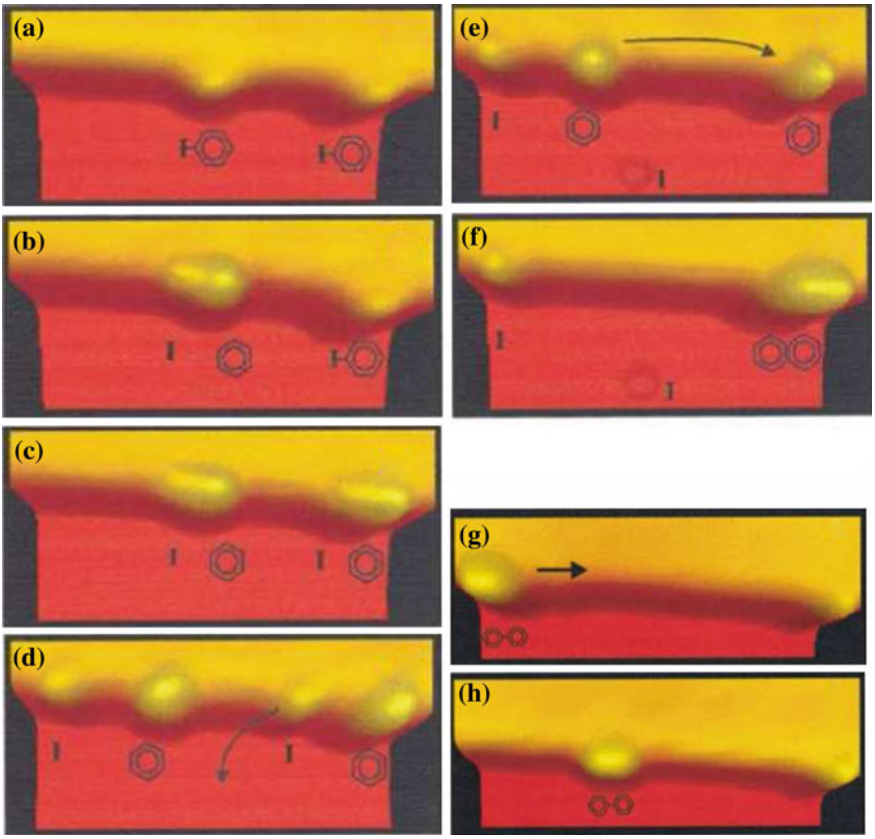

Fig. 24.8 Steps in the STM-tip-induced single-molecule Ullmann reaction of two iodobenzene parent molecules to a biphenyl molecule on Cu(111) (reproduced with permission from [64])

24.4 Summary

- STM can also be used to position atoms and molecules deliberately at desired locations.
- It can be distinguished by the height trace of the tip, if positioning occurs by pushing or pulling of the molecule.
- Electron wave functions can be confined to nanoscale corrals built from single atoms.
- Using the STM tip, a chemical reaction can be induced and followed step by step on the single molecule level.

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 [7].

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.

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

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0

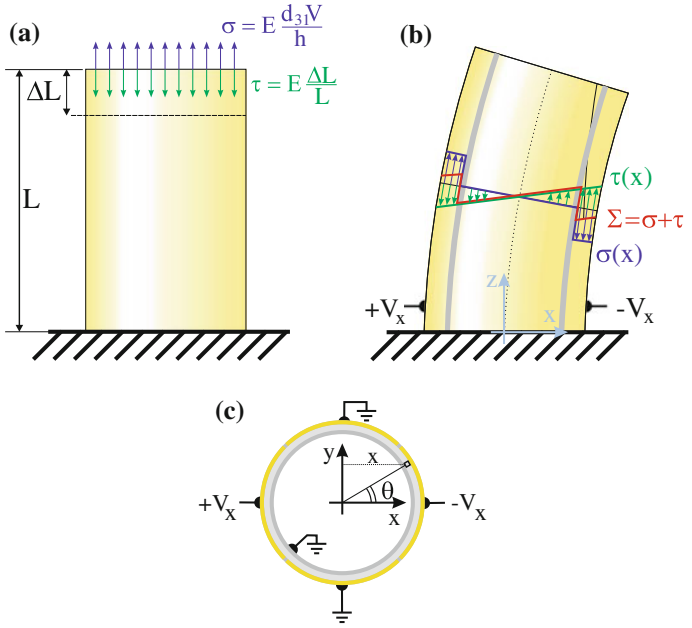

Fig. A.1 **a** When pulled with an external stress σ at an elastic object (piezo tube) the object extends by ΔL and an inner stress τ builds up as a response. In equilibrium the two stresses compensate each other. **b** In the case of a bending of the tube due to voltages on the x-electrodes, the externally 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

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 ($45° < \theta < 90°$), 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 ($0° < \theta < 45°$), 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

$$\begin{aligned} &\int_0^{90°} \Sigma(\theta) \cos\theta d\theta \\ &= \int_0^{45°} (\tau_{\max} \cos\theta - \sigma_{\max}) \cos\theta d\theta + \int_{45°}^{90°} \tau_{\max} \cos\theta \cos\theta d\theta = 0. \end{aligned} \tag{A.3}$$

The evaluation of these integrals leads to the equilibrium condition

$$\tau_{\max} = \frac{2\sqrt{2}}{\pi} \sigma_{\max}. \tag{A.4}$$

Replacing $\sigma_{\max} = E d_{31} V/h$ and $\tau_{\max} = E\Delta L/L$, results in

$$\Delta L = \frac{2\sqrt{2}}{\pi} \frac{d_{31} L V}{h}. \tag{A.5}$$

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.

Appendix B
Fermi's Golden Rule and Bardeen's Matrix Elements

In the following, we derive a variant of Fermi's golden rule which applies to the case of STM. Subsequently, we also derive Bardeen's expression for the tunneling matrix elements occurring in the equation of Fermi's golden rule.

Fermi's Golden Rule for Scanning Tunneling Microscopy

It is too complicated to solve the Schrödinger equation for the complete system of tip, sample and barrier (as shown schematically in Fig. B.1a), even in the single electron approximation. Bardeen's approach was to split the problem (initially) into two independent subsystems (tip and sample). The solutions for the independent systems can be found more easily and this knowledge can be exploited when attempting the solution of the complete system by the time-dependent perturbation theory.

A separation surface ($S_{\mathrm{tip/sample}}$) dividing the tip from the sample system is chosen somewhere inside the tunneling gap, as shown in Fig. B.1a. The potential of the complete system (Fig. B.1b) is decomposed into a tip and a sample potential shown in Fig. B.1c, d, respectively. In doing so the following conditions should be fulfilled. First the potential of the total system is built up as $V_{\mathrm{total}}(\boldsymbol{r}) = V_{\mathrm{tip}}(\boldsymbol{r}) + V_{\mathrm{sample}}(\boldsymbol{r})$, and the tip potential $V_{\mathrm{tip}}(\boldsymbol{r})$ is zero inside the sample system and vice versa, as also indicated in Fig. B.1b, c. With this choice, the origin of the energy scale is the vacuum energy.

The tip and sample potentials can be more complicated than simple rectangular shapes. Generally, a three-dimensional potential including the actual atomistic structure of tip and sample (within the single electron approximation) can be used, as schematically indicated in Fig. B.1a. The potential of the tip could also be considered in the presence of the sample and vice versa. The time-dependent Schrödinger equations for the tip system and the sample system can be written as

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0

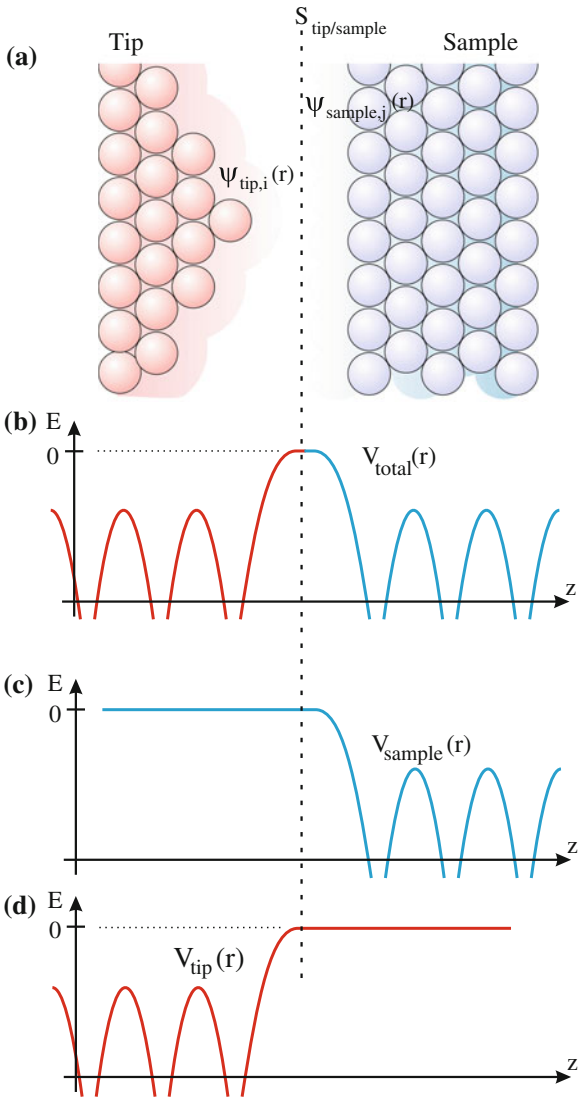

Fig. B.1 **a** Schematic of the tip and sample system considered on the atomic level. The complete system is split up into a tip and a sample system separated by a separation surface $S_{\text{tip/sample}}$ (*dashed line*). The total potential (**b**) is composed of the disjunct tip and sample potentials (**c**) and (**d**) as $V_{\text{total}}(\boldsymbol{r}) = V_{\text{tip}}(\boldsymbol{r}) + V_{\text{sample}}(\boldsymbol{r})$

$$
\begin{aligned}
i\hbar\frac{\partial\Psi_{\text{tip},i}(\boldsymbol{r},t)}{\partial t} &= \left[\frac{-\hbar^2}{2m}\Delta + V_{\text{tip}}(\boldsymbol{r})\right]\Psi_{\text{tip},i}(\boldsymbol{r},t), \text{ and} \\
i\hbar\frac{\partial\Psi_{\text{sample},j}(\boldsymbol{r},t)}{\partial t} &= \left[\frac{-\hbar^2}{2m}\Delta + V_{\text{sample}}(\boldsymbol{r})\right]\Psi_{\text{sample},j}(\boldsymbol{r},t),
\end{aligned}
\tag{B.1}
$$

respectively. The solution of the time-dependent Schrödinger equation for the state i of the tip plus vacuum system with a potential $V_{\text{tip}}(\boldsymbol{r})$ can be written as

$$\Psi_{\mathrm{tip},i}(\boldsymbol{r},t) = \psi_{\mathrm{tip},i}(\boldsymbol{r}) \exp\left(-\frac{iE_i t}{\hbar}\right), \tag{B.2}$$

with $\psi_{\mathrm{tip},i}(\boldsymbol{r})$ being the solution of the time-independent Schrödinger equation of the tip plus vacuum system. A corresponding equation applies for the sample states j. The time-independent Schrödinger equations for the tip states i and the sample states j can be written as

$$\left(-\frac{\hbar^2}{2m}\Delta + V_{\mathrm{tip}}(\boldsymbol{r})\right)\psi_{\mathrm{tip},i}(\boldsymbol{r}) = E_i\psi_{\mathrm{tip},i}(\boldsymbol{r}), \tag{B.3}$$

and

$$\left(-\frac{\hbar^2}{2m}\Delta + V_{\mathrm{sample}}(\boldsymbol{r})\right)\psi_{\mathrm{sample},j}(\boldsymbol{r}) = E_j\psi_{\mathrm{sample},j}(\boldsymbol{r}). \tag{B.4}$$

Considering only the separate systems, an electron in a specific tip state would remain in this state forever. Now we consider a transition (scattering or tunneling) of an electron from its initial state i to the final states j of the sample. Tunneling leads only to a small perturbation of the initial state, so that the time-dependent wave function of the complete system can be written as follows

$$\Psi_{\mathrm{final}}(\boldsymbol{r},t) = \Psi_{\mathrm{tip},i}(\boldsymbol{r},t) + \sum_j a_j(t)\Psi_{\mathrm{sample},j}(\boldsymbol{r},t), \tag{B.5}$$

where the sum extends over all final states j. The final wave function will almost be the same as the initial one $\Psi_{\mathrm{tip},i}(\boldsymbol{r},t)$ plus a sum over the sample stationary wave functions $\Psi_{\mathrm{sample},j}(\boldsymbol{r},t)$ with a_j being the probability amplitude. The square of a_j describes the probability of the electron being in state $\Psi_{\mathrm{sample},j}(\boldsymbol{r},t)$. At time $t = 0$ all the a_j are zero, since the wave function is still the initial tip wave function. Since all the a_j are assumed to be small for small times, the wave function $\Psi_{\mathrm{final}}(\boldsymbol{r},t)$ remains normalized in the first order. The aim in the following is to calculate the time-dependence of the a_j which correspond to the transfer rate (tunneling rate) into the final states.

The wave function $\Psi_{\mathrm{final}}(\boldsymbol{r},t)$ written in (B.5) is a solution to the time-dependent Schrödinger equation of the complete system

$$i\hbar\frac{\partial\Psi_{\mathrm{final}}(\boldsymbol{r},t)}{\partial t} = \left[\frac{-\hbar^2}{2m}\Delta + V_{\mathrm{tip}}(\boldsymbol{r}) + V_{\mathrm{sample}}(\boldsymbol{r})\right]\Psi_{\mathrm{final}}(\boldsymbol{r},t). \tag{B.6}$$

Inserting the wave function (B.5) into the time-dependent Schrödinger equation of the complete system (B.6), this results in

$$i\hbar\frac{\partial\Psi_{\mathrm{tip},i}(\boldsymbol{r},t)}{\partial t} + i\hbar\sum_j\frac{da_j(t)}{dt}\Psi_{\mathrm{sample},j}(\boldsymbol{r},t) + i\hbar\sum_j a_j(t)\frac{\partial\Psi_{\mathrm{sample},j}(\boldsymbol{r},t)}{\partial t}$$

$$= \left[-\frac{\hbar^2}{2m}\Delta + V_{\text{tip}}(\boldsymbol{r}) \right] \Psi_{\text{tip},i}(\boldsymbol{r}, t) + V_{\text{sample}}(\boldsymbol{r}) \Psi_{\text{tip},i}(\boldsymbol{r}, t)$$
$$+ \left[-\frac{\hbar^2}{2m}\Delta + V_{\text{tip}}(\boldsymbol{r}) + V_{\text{sample}}(\boldsymbol{r}) \right] \sum_j a_j(t) \Psi_{\text{sample},j}(\boldsymbol{r}, t). \tag{B.7}$$

Since the Schrödinger equations are given for the tip and the sample systems separately (B.1), the first term in the first line cancels out against the first term in the second line. Also the last term in the first line cancels out against large parts of the last line, due to the Schrödinger equation for the sample states (B.1). In the last line, only the term proportional to $V_{\text{tip}}(\boldsymbol{r})$ survives. In total we can rewrite (B.7) as

$$i\hbar \sum_j \frac{da_j(t)}{dt} \Psi_{\text{sample},j}(\boldsymbol{r}, t)$$
$$= V_{\text{sample}}(\boldsymbol{r}) \Psi_{\text{tip},i}(\boldsymbol{r}, t) + V_{\text{tip}}(\boldsymbol{r}) \sum_j a_j(t) \Psi_{\text{sample},j}(\boldsymbol{r}, t). \tag{B.8}$$

If we now replace $\Psi_{\text{tip},i}(\boldsymbol{r}, t)$ and $\Psi_{\text{sample},j}(\boldsymbol{r}, t)$ according to (B.2) we obtain

$$i\hbar \sum_j \frac{da_j(t)}{dt} \psi_{\text{sample},j}(\boldsymbol{r}) \exp\left(-\frac{iE_j t}{\hbar}\right) = V_{\text{sample}}(\boldsymbol{r}) \psi_{\text{tip},i}(\boldsymbol{r}) \exp\left(-\frac{iE_i t}{\hbar}\right)$$
$$+ V_{\text{tip}}(\boldsymbol{r}) \sum_j a_j(t) \psi_{\text{sample},j}(\boldsymbol{r}) \exp\left(-\frac{iE_j t}{\hbar}\right). \tag{B.9}$$

Now we evaluate the matrix elements as usual in quantum mechanics, by multiplying (B.9) by the wave function $\psi^*_{\text{sample},f}$ of a specific sample state f, and subsequently perform a spatial integration. If we consider that the final states are normalized and orthogonal ($\int \psi^*_{\text{sample},f} \psi_{\text{sample},j} d^3\boldsymbol{r} = \delta_{fj}$) most terms in the sums vanish from (B.9) to give

$$i\hbar \frac{da_f(t)}{dt} \exp\left(-\frac{iE_f t}{\hbar}\right)$$
$$= \int \psi^*_{\text{sample},f}(\boldsymbol{r}) V_{\text{sample}}(\boldsymbol{r}) \psi_{\text{tip},i}(\boldsymbol{r}) d^3\boldsymbol{r} \exp\left(-\frac{iE_i t}{\hbar}\right)$$
$$+ V_{\text{tip}}(\boldsymbol{r}) a_f(t) \exp\left(-\frac{iE_f t}{\hbar}\right). \tag{B.10}$$

Since the a_f terms are considered to be small we neglect them on the right side of (B.10) as usually done in perturbation theory in quantum mechanics. With this (finally) the following differential equation for the time-dependence of the coefficients a_f is obtained

$$\frac{da_f(t)}{dt} = \frac{1}{i\hbar}\int \psi^*_{\text{sample},f}(\boldsymbol{r})V_{\text{sample}}(\boldsymbol{r})\psi_{\text{tip},i}(\boldsymbol{r})d^3\boldsymbol{r}\exp\left[\frac{i(E_f - E_i)t}{\hbar}\right]. \tag{B.11}$$

The expression

$$M_{fi} = \int\limits_{\Gamma_{\text{sample}}} \psi^*_{\text{sample},f}(\boldsymbol{r})V_{\text{sample}}(\boldsymbol{r})\psi_{\text{tip},i}(\boldsymbol{r})d^3\boldsymbol{r} \tag{B.12}$$

occurring in (B.11) is called the (transition) matrix element with the integral originally extending over the whole space. Since, however, the sample potential vanishes in the tip region, the integration can be limited to the sample space Γ_{sample}.

Equation (B.11) can be integrated and results in

$$a_f(t) = \frac{1}{i\hbar}M_{fi}\int\limits_0^t \exp\left[\frac{i(E_f - E_i)t'}{\hbar}\right]dt'. \tag{B.13}$$

This is the probability amplitude of state f at time t. It is assumed that the transition rate is small, so that the initial state can always be taken as nearly full and the final states as nearly empty. The integral in (B.13) can be evaluated as[1]

$$\begin{aligned} a_f(t) &= -M_{fi}\frac{\exp\left[i(E_f - E_i)t/\hbar\right] - 1}{(E_f - E_i)} \\ &= -2iM_{fi}\exp\left[i(E_f - E_i)t/2\hbar\right]\frac{\sin\left[(E_f - E_i)t/2\hbar\right]}{E_f - E_i}. \end{aligned} \tag{B.15}$$

Using (B.15), the probability of finding an electron which was originally in tip state i in the sample state f is

$$\left|a_f(t)\right|^2 = 4\left|M_{fi}\right|^2\frac{\sin^2\left[(E_f - E_i)t/2\hbar\right]}{(E_f - E_i)^2} = \left|M_{fi}\right|^2\frac{2t}{\hbar}\frac{\sin^2\left[(E_f - E_i)t/2\hbar\right]}{(E_f - E_i)^2 t/2\hbar}. \tag{B.16}$$

[1] The last equality arises since

$$\begin{aligned} &\exp(ia) - 1 = \cos a + i\sin a - 1 \\ &= 1 - 2\sin^2\frac{a}{2} + 2i\sin\frac{a}{2}\cos\frac{a}{2} - 1 \\ &= 2\sin\frac{a}{2}\left(i\cos\frac{a}{2} - \sin\frac{a}{2}\right) \\ &= 2i\sin\frac{a}{2}\left(\cos\frac{a}{2} + i\sin\frac{a}{2}\right) \\ &= 2i\exp\left(i\frac{a}{2}\right)\sin\frac{a}{2}. \end{aligned} \tag{B.14}$$

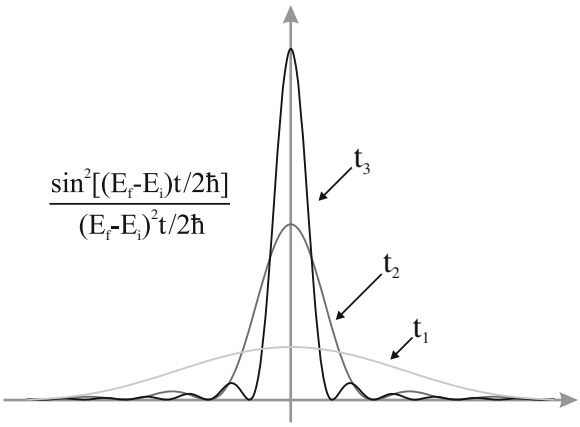

Fig. B.2 Probability of being found in the final state as a function of the difference in energy between initial and final state for three times. This function becomes infinitely *high* and *narrow* as time increases, which corresponds to energy conservation for the (*tunneling*) transitions from an initial to a final state

Equation (B.16) is plotted as function of $E_f - E_i$ in Fig. B.2. The probability of being found in any particular final state f peaks for E_f close to E_i. The function $\left|a_f(t)\right|^2$ becomes infinitely high and narrow, with increasing t. If we use the following representation of the Dirac delta function

$$\delta(x) = \lim_{a\to\infty} \frac{1}{\pi} \frac{\sin^2 ax}{ax^2}, \tag{B.17}$$

and identify $x = E_f - E_i$ and $a = t/2\hbar$, $\left|a_f(t)\right|^2$ results in the limit of large times as

$$\left|a_f^{\infty}(t)\right|^2 = \frac{2\pi}{\hbar} \left|M_{fi}\right|^2 \delta(E_f - E_i)t. \tag{B.18}$$

Thus the transition rate (electrons per time) from the initial tip state i to a final sample state f is given by

$$w_{if} = \frac{2\pi}{\hbar} \left|M_{fi}\right|^2 \delta(E_f - E_i). \tag{B.19}$$

Thus Fermi's golden rule applied by Bardeen to the case of tunneling results in a transition rate proportional to the matrix element. Energy conservation (elastic tunneling) is obeyed by the delta function.

Bardeen's Expression for the Tunneling Matrix Elements

A disadvantage of the expression of the matrix element M_{fi} in (B.12) is that it depends not only on the wave functions of tip and sample but also explicitly on the sample potential $V_{\text{sample}}(\boldsymbol{r})$. In the following, we derive Bardeen's expression of the matrix

element which does not explicitly depend on the sample potential, but on the values of the wave functions and their derivatives on a certain surface. We start with (B.12) as

$$M_{fi} = \int\limits_{\Gamma_{\text{sample}}} \psi_{\text{tip},i}(\boldsymbol{r}) V_{\text{sample}}(\boldsymbol{r}) \psi^*_{\text{sample},f}(\boldsymbol{r}) d^3\boldsymbol{r}. \tag{B.20}$$

Bardeen showed that, using the time-independent Schrödinger equations for tip and sample states, the matrix element can be rewritten to an expression which depends (explicitly) only on the values of the wave functions of the tip and sample states on the separation surface $S_{\text{tip/sample}}$.

In the following, we will use the time-independent Schrödinger equations for the tip and sample states (B.3) and (B.4). If we insert the expression for $V_{\text{sample}}(\boldsymbol{r})$ $\psi^*_{\text{sample},f}(\boldsymbol{r})$ obtained from (B.4) into the expression for the matrix element (B.20), we obtain

$$M_{fi} = \int\limits_{\Gamma_{\text{sample}}} \psi_{\text{tip},i}(\boldsymbol{r}) \left(E_f + \frac{\hbar^2}{2m}\Delta \right) \psi^*_{\text{sample},f}(\boldsymbol{r}) d^3\boldsymbol{r}. \tag{B.21}$$

Since we know from (B.19) that tunneling transitions occur only if $E_i = E_f$, we replace E_f in (B.21) by E_i and obtain

$$M_{fi} = \int\limits_{\Gamma_{\text{sample}}} \left(E_i \psi_{\text{tip},i}(\boldsymbol{r}) \psi^*_{\text{sample},f}(\boldsymbol{r}) + \psi_{\text{tip},i}(\boldsymbol{r}) \frac{\hbar^2}{2m} \Delta \psi^*_{\text{sample},f}(\boldsymbol{r}) \right) d^3\boldsymbol{r}. \tag{B.22}$$

If we use (B.3) to replace $E_i \psi_{\text{tip},i}(\boldsymbol{r})$ in (B.22), this results in

$$\begin{aligned} M_{fi} = \int\limits_{\Gamma_{\text{sample}}} \Bigg[& \left(-\frac{\hbar^2}{2m}\Delta + V_{\text{tip}}(\boldsymbol{r}) \right) \psi_{\text{tip},i}(\boldsymbol{r}) \psi^*_{\text{sample},f}(\boldsymbol{r}) \\ & + \psi_{\text{tip},i}(\boldsymbol{r}) \frac{\hbar^2}{2m} \Delta \psi^*_{\text{sample},f}(\boldsymbol{r}) \Bigg] d^3\boldsymbol{r}. \end{aligned}$$

Since the integration extends over the sample volume and $V_{\text{tip}}(\boldsymbol{r})$ is zero here, we can skip this term resulting in

$$M_{fi} = \frac{\hbar^2}{2m} \int\limits_{\Gamma_{\text{sample}}} \left[\psi_{\text{tip},i}(\boldsymbol{r}) \Delta \psi^*_{\text{sample},f}(\boldsymbol{r}) - \Delta \psi_{\text{tip},i}(\boldsymbol{r}) \psi^*_{\text{sample},f}(\boldsymbol{r}) \right] d^3\boldsymbol{r}. \tag{B.23}$$

The volume integral in (B.23) can be converted to a surface integral using Green's second identity, which results in

$$M_{fi} = \frac{\hbar^2}{2m} \int\limits_{S_{\text{tip/sample}}} \left[\psi_{\text{tip},i}(\boldsymbol{r}) \nabla \psi^*_{\text{sample},f}(\boldsymbol{r}) - \psi^*_{\text{sample},f}(\boldsymbol{r}) \nabla \psi_{\text{tip},i}(\boldsymbol{r}) \right] \cdot d\boldsymbol{S}. \quad \text{(B.24)}$$

The integral extends over the separation surface $S_{\text{tip/sample}}$. The parts closing the surface integral are considered to be at infinity and to add a negligible contribution. Therefore, the matrix element depends only on the values of the tip and sample wave functions on the separation surface. The dependence on the potential is implicit, since the wave functions depend via the Schrödinger equation on the potential.

Appendix C
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 [37].

The oscillation of the cantilever at its shifted resonance frequency ω_0' is written (neglecting an offset phase) as

$$z(t) = A\sin(\omega_0' t). \tag{C.1}$$

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

$$\omega(t) = \omega_0' + \omega_\Delta \cos(\omega_{\mathrm{mod}} t), \tag{C.2}$$

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 $\phi = \int \omega(t)dt$. With this the oscillation coordinate can be written as

$$z(t) = A\sin\left(\int \omega(t)dt\right) = A\sin\left(\omega_0' t + \frac{\omega_\Delta}{\omega_{\mathrm{mod}}}\sin(\omega_{\mathrm{mod}} t)\right). \tag{C.3}$$

This expression can be written as an infinite sum over Bessel functions. However, in the limit that $\omega_\Delta \ll \omega_{\mathrm{mod}}$, the oscillation of the cantilever can be approximated as

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0

$$z(t) = A \sin \omega_0' t + \frac{A\omega_\Delta}{2\omega_{\mathrm{mod}}} \left(\sin \left[\left(\omega_0' + \omega_{\mathrm{mod}} \right) t \right] - \sin \left[\left(\omega_0' - \omega_{\mathrm{mod}} \right) t \right] \right). \quad \text{(C.4)}$$

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

$$z(t) = A \sin \omega_0' t + \sqrt{2}\delta A_+ \sin \left[\left(\omega_0' + \omega_{\mathrm{mod}} \right) t + \phi_0 \right]. \quad \text{(C.5)}$$

Using the mathematical identity $\sin(\alpha + \beta) = \sin\alpha \cos\beta + \cos\alpha \sin\beta$, the following expression results

$$\begin{aligned} z(t) = A \sin \omega_0' t \left[1 + \frac{\sqrt{2}\delta A_+}{A} \cos\left(\omega_{\mathrm{mod}} t + \phi_0 \right) \right] \\ + \sqrt{2}\delta A_+ \cos \omega_0' t \sin\left(\omega_{\mathrm{mod}} t + \phi_0 \right). \end{aligned} \quad \text{(C.6)}$$

Since $\delta A_+ \ll A$, the second term in the square brackets can be neglected, which results in

$$\begin{aligned} z(t) = A \sin \omega_0' t \cos\left(\frac{\sqrt{2}\delta A_+}{A} \sin\left(\omega_{\mathrm{mod}} t + \phi_0 \right) \right) \\ + A \cos \omega_0' t \sin\left(\frac{\sqrt{2}\delta A_+}{A} \sin\left(\omega_{\mathrm{mod}} t + \phi_0 \right) \right). \end{aligned} \quad \text{(C.7)}$$

In order to apply the above-mentioned identity for trigonometric functions in the next step, we included the factor $\cos \frac{\sqrt{2}\delta A_+}{A} (\omega_{\mathrm{mod}} t + \phi_0)$, which is very close to one, since $\delta A_+ \ll A$. Further, we also replaced the small term $\frac{\sqrt{2}\delta A_+}{A} \sin(\omega_{\mathrm{mod}} t + \phi_0)$ by its sinus. Due to this we can apply the above-mentioned identity in the reverse direction, which results in

$$z(t) = A \sin\left(\omega_0' t + \frac{\sqrt{2}\delta A_+}{A} \sin\left(\omega_{\mathrm{mod}} t + \phi_0 \right) \right). \quad \text{(C.8)}$$

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

$$\omega(t) = \omega_0' t + \frac{\sqrt{2}\delta A_+}{A} \omega_{\mathrm{mod}} \cos\left(\omega_{\mathrm{mod}} t + \phi_0 \right). \quad \text{(C.9)}$$

Thus the RMS displacement noise δA_+ at the frequency $\omega_0 + \omega_{\mathrm{mod}}$ translates into a RMS frequency noise $\delta\omega_+$, as

$$\delta\omega_+ = \frac{\omega_{\mathrm{mod}}}{A}\delta A_+ \qquad \text{or} \qquad \delta f_+ = \frac{f_{\mathrm{mod}}}{A}\delta A_+, \tag{C.10}$$

correspondingly for the natural frequencies.

If we additionally consider a second independent noise component of the same magnitude from the lower side band at $\omega_0' - \omega_{\mathrm{mod}}$, 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

$$N_f(f_{\mathrm{mod}}) = \frac{\sqrt{2}f_{\mathrm{mod}}}{A}N_z(f_0 + f_{\mathrm{mod}}), \tag{C.11}$$

where N_z is the spectral displacement noise density and N_f is the spectral frequency noise density.

References

1. W.D. Pilkey, *Formulas for Stress, Strain and Structural Matrices* (Wiley, New York, 1994)
2. PI Ceramic GmbH, Lindenstrasse, 07589 Lederhose, Germany www.piceramic.com
3. K.G. Vandervoort, R.K. Zasadzinski, G.G. Galicia, G.W. Crabtree, Rev. Sci. Instrum. **65**, 3862 (1994)
4. J.P. Spatz, S. Sheiko, M. Moller, R.G. Winkler, P. Reineker, O. Marti, Nanotechnology **6**, 40 (1995)
5. aixACCT Systems GmbH, Talbotstr. 25, 52068 Aachen, Germany www.aixacct.com
6. J.P. Ibe, P.P. Bey Jr, S.L. Brandow, R.A. Bizzolara, N.A. Burnham, D.P. DiLella, K.P. Lee, C.R.K. Marrian, R.J. Colton, J. Vac. Sci. Technol. A **8**, 3570 (1990)
7. C. Julian Chen, *Introduction to Scanning Tunneling Microscopy* (Oxford University Press, New York, 2008)
8. B. Drake, R. Sonnenfeld, J. Schneir, P.K. Hansma, Surf. Sci. **181**, 92 (1987)
9. Patent DE 40 23311 C2
10. Patent DE 3610540 C2
11. Patent WO 93/19494
12. J. Johnson, Phys. Rev. **32**, 97 (1928)
13. H. Nyquist, Phys. Rev. **32**, 110 (1928)
14. H. Ibach, H. Lüth, *Solid-State Physics* (Springer, Heidelberg 2009)
15. S.G. Davison, M. Steslicka, *Basic Theory of Surface States* (Oxford University Press, Oxford, 1996)
16. H. Lüth, *Solid Surfaces, Interfaces and Thin Films* (Springer, Heidelberg, 2010)
17. J. Israelachvili, *Intermolecular and Surface Forces* (Academic Press, London, 2011)
18. M. Saint Jean, S. Hudlet, C. Guthmann, J. Berger, J. Appl. Phys. **86**, 5245 (1999)
19. J.L. Hutter, Langmuir **21**, 2630 (2005)
20. J.E. Sader et al., Rev. Sci. Instr. **70**, 3967 (1999)
21. M. Godin et al., Appl. Phys. Lett. **79**, 551 (2001)
22. J.E. Sader et al., Rev. Sci. Instr. **83**, 103705 (2012)
23. J.L. Hutter, H. Bechhoefer, Rev. Sci. Instrum **64**, 1868 (1993)
24. S. Cook, T.E. Schaffer et al., Nanotechnology **17**, 2135 (2006)
25. H.-J. Butt, M. Jaschke, Nanotechnology **6**, 1 (1995)
26. R. Proksch, T.E. Schäffer, J.P. Cleveland, R.C. Callahan, M.B. Viani, Nanotechnology **15**, 1344 (2004)
27. M.J. Higgins et al., Rev. Sci. Instr. **77**, 013701 (2006)
28. J.P. Cleveland, B. Anczykowski, A.E. Schmid, V.B. Elings, Appl. Phys. Lett. **72**, 2613 (1998)

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0

29. M.V. Salapaka, D.J. Chen, J.P. Cleveland, Phys. Rev. B **61**, 1106 (2000)
30. L.D. Landau, E.M. Lifshitz, *Mechanics*, (Volume 1 of A Course of Theoretical Physics) (Pergamon Press, Oxford, 1969)
31. H. Hölscher, U.D. Schwarz, Non-linear Mech. **42**, 608 (2007)
32. J.E. Sader, T. Uchihashi, M.J. Higgins, A. Farrell, Y. Nakayama, S.P. Jarvis, Nanotechnology **16**, S94 (2005)
33. S. Morita, F.J. Giessibl, R. Wiesendanger, *Non-Contact Atomic force Microscopy*, vol. 2 (Springer, Heidelberg, 2009)
34. F.J. Giessibl, H. Bielefeld, Phys. Rev. B **61**, 9968 (2000)
35. J.E. Sader, S.P. Jarvis, Appl. Phys. Lett. **84**, 1801 (2004)
36. H. Hölscher et al., Phys. Rev. B **64**, 075402 (2001)
37. K. Kobayashi, H. Yamada, K. Matsushige, Rev. Sci. Instr. **80**, 043708 (2009)
38. F.J. Giessibl, F. Pielmeier, T. Eguchi, T. An, Y. Hasegawa, Phys. Rev. B **84**, 12540968 (2011)
39. F.J. Giessibl, S. Hembacher, M. Herz, Ch. Schiller, J. Mannhart, Nanotechnology **15**, S79 (2004)
40. I. Morawski, B. Voigtländer, Rev. Sci. Instr. **81**, 033703 (2010)
41. G.H. Simon, M. Heyde, H.-P. Rust, Nanotechnology **18**, 255503 (2007)
42. P. Hofmann, *Solid State Physics* (Wiley, Weinheim, 2008)
43. R.M. Feenstra, Joseph A. Stroscio, J. Tersoff, A.P. Fein, Phys. Rev. Lett. **58**, 1192 (1987)
44. R.M. Feenstra, J.A. Stroscio, A.P. Fein, Surf. Sci. **181**, 295 (1987)
45. Scanning Tunneling Microscopy, ed. by J.A. Stroscio, W.J. Kaiser (Academic Press, Boston, 1993)
46. B. Koslowski, C. Dietrich, A. Tschetschetkin, P. Ziemann, Phys. Rev. B **75**, 035421 (2007)
47. J. Wiebe, A. Wachowiak, F. Meier, D. Haude, T. Foster, M. Morgenstern, R. Wiesendanger, Rev. Sci. Instrum. **75**, 4871 (2004)
48. R. Temirov, personal communication
49. R. Temirov, A. Lassise1, F.B. Anders, F.S. Tautz, Nanotechnology **19**, 065401 (2008)
50. J. Myslivecek, F. Dvorak, A. Strozecka, B. Voigtländer,Phys. Rev. B **81**, 245427 (2010)
51. J. Myslivecek, A. Strozecka, J. Steffl, P. Sobotik, I. Ostadal,B. Voigtländer, Phys. Rev. B **73**, 161302(R) (2006)
52. K. Oura, V.G. Lifshits, A. Saranin, A.V. Zotov, M. Katayama, *Surface Science* (Springer, Heidelberg, 2003)
53. J.I. Pascual, Eur. Phys. J. D **35**, 327 (2005)
54. B.C. Stipe, M.A. Rezaei, W. Ho, Phys. Rev. Lett. **82**, 1724 (1999)
55. R. Waser (ed.), *Nanoelectronics and Information Technology* (Wiley, Weinheim, 2012)
56. Chr. Meyer, J. Klijn, M. Morgenstern, R. Wiesendanger, Phys. Rev. Lett. **91**, 076803 (2003)
57. U. Banin, Y.W. Cao, D. Katz, O. Millo, Nature **400**, 542 (1999)
58. M.F. Crommie, C.P. Lutz, D.M. Eigler, Nature **363**, 524 (1993)
59. L. Bartels, G. Meyer, K.-H. Rieder, Phys. Rev. Lett. **79**, 697 (1997)
60. A. Strozecka, J. Myslivecek,B. Voigtländer. Appl. Phys. A **87**, 475 (2007)
61. G. Kichin, C. Weiss, C. Wagner, F.S. Tautz, R. Temirov, J. Am Chem. Soc. **133**, 16847 (2011)
62. M.F. Crommie, C.P. Lutz, D.M. Eigler, E.J. Heller, Surf. Rev. Lett. **2**, 127 (1995)
63. E.J. Heller, M.F. Crommie, C.P. Lutz, D.M. Eigler, Nature **369**, 464 (1994)
64. S.W. Hla, L. Bartels, G. Meyer, K.-H. Rieder, Phys. Rev. Lett. **85**, 2777 (2000)
65. H. Schönherr, G.J. Vancso, *Scanning Force Microscopy of Polymers* (Springer, Berlin, 2010)
66. F.J. Giessibl, Science **267**, 68 (1995)
67. M.F. Crommie, C.P. Lutz, D.M. Eigler, Science **262**, 218 (1993)
68. B.C. Stipe, M.A. Rezaei, W. Ho, Science **279**, 1907 (1998)
69. C.B. Duke, *Tunneling in Solids*, (Solid State Phys. Suppl. 10), ed. by F. Seitz, D. Turnbull, H. Ehrenreich (Academic Press, New York, 1969)
70. G. Meyer, N.M. Amer, Appl. Phys. Lett. **53**, 1045 (1988)

Index

B. Voigtländer, *Scanning Probe Microscopy*,
NanoScience and Technology, DOI 10.1007/978-3-662-45240-0