INSTRUMENT FOR MEASURING pH WITH OPTICAL INDICATOR THIN FILM

Instrument for Measuring pH with Optical Indicator Thin Film

Ville Voipio

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission for public examination and debate in Auditorium S4 at Helsinki University of Technology (Espoo, Finland) on the 7th of December, 2001, at 12 o'clock noon.

Book distribution by the author (e-mail: ville.voipio@iki.fi).

ISBN 951-22-5726-2 (book)
ISBN 951-22-5727-0 (pdf)
ISBN 951-22-5728-9 (CD-ROM)

HELSINKI UNIVERSITY OF TECHNOLOGY P.O.Box 1000, FIN-02015 HUT http://www.hut.fi/	ABSTRACT OF DOCTORAL DISSERTATION

Author	Ville Voipio		
Name of the dissertation	Instrument for Measuring pH with Optical Indicator Thin Film		
Date of manuscript	August 21, 2001	Date of the dissertation	December 7, 2001
☒ Monograph		☐ Article dissertation (summary + original articles)	
Department	Department of Electrical and Communications Engineering		
Laboratory	Metrology Research Institute		
Field of research	Measurement science and technology		
Opponent(s)	Prof. Vahid Sandoghdar, Dr. Juha Rantala		
Supervisor	Prof. Erkki Ikonen		
Instructor(s)			

Abstract

The aim of this thesis is to describe a novel pH measurement instrument and demonstrate its performance by measurements carried out with a prototype. pH measurement is one of the most common measurements in process industry, but despite the wide-spread use of the measurement, common pH measurement methods have several deficiencies.

The new instrument is based on measuring color changes of pH sensitive dye molecules which are trapped in a thin porous glass membrane. The main advantage of the dye measurement is the strong measurement signal, there are no noisy low-level electrochemical signals to be measured.

The dye color measurement is realized by depositing a porous dielectric mirror on top of the dye film. This mirror can be manufactured with the same methods as the indicator film itself. The reflection spectrum of this indicator film and dielectric mirror combination is then dependent on the external pH, and the pH can then be determined by a simple three-wavelength measurement.

A prototype was constructed to prove the viability of this optical measurement concept. New optical constructions were designed to meet the requirements of this instrument. The data processing methods required in the measurement interpretation were also developed.

This thesis describes a part of an industrial research and development project. The ultimate aim of the project is to produce a commercial measurement instrument which is manufactured in large volumes. Thus, economic viability and manufacturability play an important role in the constructions and methods described in this work.

Prototype measurements show the measurement principle is accurate enough to be used in an industrial instrument. The results also show that the optical properties of the medium under measurement can be compensated for with signal processing.

In addition to the pH measurement instrument itself, some new sol-gel thin film research methods were developed. Also, to enable accurate manufacturing control, a signal processing method for improving the resolution of a commercially available position sensor was developed.

Keywords	pH, sol-gel, thin film, dielectric mirror, process instrument		
UDC	541.132.3:53.08:681.2.082	Number of pages	194 p. + app. 14 p.
ISBN (printed)	951-22-5726-2	ISBN (pdf)	951-22-5727-0
ISBN (CD-ROM)	951-22-5728-9	ISSN	
Publisher	Janesko Oy		
Print distribution	Ville Voipio (vv@iki.fi)		
☒ The dissertation can be read at http://lib.hut.fi/Diss/			

Contents

Preface

Five years ago my boss, Jan Kåhre, came to me and asked if I knew anything about measuring pH. He had visited the Westinghouse Savannah River Company and seen there some thin glass films which changed their color in response to external acidity. This gave him the idea of making a pH measurement instrument around this idea. He had also some ideas about the basic principles of the instrument and just left it to me to design the instrument.

I thought it should be rather straightforward to develop an instrument around this idea. After all, it was only a matter of making some glass materials with known recipes, and then adding some standard electronics and optics to the system.

I was wrong. I underestimated the complexity and challenges of the project completely. In a way this was fortunate, as had I known the real extent of the project, I would have kept it impossible.

After I started working with the project, difficulties started to emerge. The recipes we had for the glass materials did not produce desired results, or anything useful at all. The first ideas how to make the measurement turned out to be impractical. Several times during the project almost everything seemed to go wrong.

According to an old metaphor, a large project is like eating an elephant. You cannot eat the whole elephant at once, you have to eat it bit by bit. And, naturally, it is very important to eat all the bits from the same elephant.

The road to a working prototype was bumpy. Advancing steps were initially very short. However, one thing lead to another, and finally it became possible to verify our visions by constructing a prototype .

At an early stage of the project I realized I may never be able to eat the elephant all by myself, and so research chemist Katri Vuokila started to eat the elephant from the other end by developing the materials and syntheses required in manufacturing the sensor films. Without her knowledge and willingness to share it I would still be in lab trying to understand sol-gel chemistry.

In industrial research and development long-term work is very often interrupted and distracted by everyday tasks. I owe Jan a thank you for keeping me to the same elephant, and also keeping other elephants from approaching me. While this prototype is my design, it is unarguably originally Jan's vision.

Jan and Katri have been great co-workers in this project. I am grateful for their comments concerning this thesis. In addition to the academic content of our long and numerous discussions, I would like to thank both of them for all the fun we have had working together.

Professor Erkki Ikonen has been a great instructor for this work. He has offered all the help I have needed but has still let me do my work the way I have wanted to do it. A thank you belongs also to professor Folke Stenman and doctor Juha Rantala for their valuable and to-the-point comments concerning my manuscript.

Most of the project has been privately funded by my employer, Janesko Oy. In addition to this, the National Technology Agency (TEKES) has funded a significant part of the project. Without TEKES funding this project would have been very difficult to carry out in a small company. A special thank you is due to Aila Maijanen who believed in us and our strange visions right from the beginning.

Last, but certainly not least, I would like to warmly thank my parents for all the discussions, food, and support they have given me throughout this process.

I am afraid I have omitted many people who have helped me in making this research. A huge thank you to all of you, this would not have happened without your support!

Ville Voipio
Vantaa
November 2001

Abbreviations

AD	Analog/digital [conversion/converter]
AR	Anti-reflection [coating]
BPB	Bromophenol blue, the dye used in the pH sensitive films
BS	Borosilicate, low refractive index film material
BST	Borosilicate/titanate, high refractive index film material
DC	Direct current (more generally: constant component of a signal)
LED	Light Emitting Diode
PTFE	Polytetrafluoroethylene, plastic material, also known as Teflon
PZT	Lead Zirconate Titanate, a piezoelectric thin film material
SAR	Successive approximation [AD converter]

1 Introduction

This thesis forms a part of an industrial project whose goal is to develop a pH measurement instrument which is suitable for in-line[1] applications in process industry, and which is better than currently used instruments.

1.1 MOTIVATION

The mainstream pH measurement instruments (electrochemical sensors) require a lot of care and calibration when used in industrial application. While the instruments are relatively inexpensive, their servicing and calibrating makes pH measurement expensive. Yet, pH measurement is one of the most commonly used process measurements and it cannot be easily replaced by any other measurement.

Electrochemical sensor technology has already matured during the decades, and advances on that field are small and slow. So, a completely different approach has been taken to find solutions to the pH measurement problem.

The new solution—or, rather, an old solution in another form—is to use indicator dyes to measure pH. They have several advantages, a dye molecule never gets old and slow; if it does not disintegrate, it always has the same response. However, the most common form of an indicator, liquid, is not very useful in process measurements as such. Hence, the indicator has to be bound to a matrix.

The instrument described in this thesis uses dye molecules bound to a thin glass film (in the order of a few hundred nanometers). This is a known technique but its usability has been very limited due the lack of suitable reflective color measurement methods. A new reflective dye color measurement method avoiding the main problems of earlier methods has been developed in this project. The main goal of this thesis

[1] In-line instrument is one which measures the main process stream directly. On line instruments measure a small side stream and off-line instruments measure discrete samples taken from the process.

is to prove the usability of this method in pH measurements and its applicability to an industrial measurement instrument.

The use in process industry sets several requirements. The instrument has to be rugged enough to withstand the harsh environmental conditions, i.e. heat, dust, vibration, and corrosive environment. There are also several electrical limitations, especially if the instrument is to be taken to explosion hazard environments.

A laboratory instrument has to be as accurate as possible for a short period of time after calibration. Laboratory instruments may be calibrated daily or even before each measurement. On the other hand, process instruments have to be able to keep their calibration for a long period of time, preferably several months, or otherwise incorporate an integral calibration system.

The instrument introduced in this doctoral thesis is based on the reflective measurement of optical indicator molecules trapped into a thin film membrane. The main advantages of this approach are the lack of any buffer or reference liquids, and very small drift compared to glass electrodes. The instrument may also be used in high-pressure conditions where the glass membrane electrodes are difficult or impossible to use.

This thesis describes a part of an industrial research and development project. The main emphasis has been on finding solutions which are both manufacturable and economically viable, rather than finding the ultimately best solutions at any cost. This should be borne in mind throughout this thesis; there is most probably a plethora of ways to make the measurement more accurate, but most of those ways lead to solutions which either do not meet the process requirements or are simply too expensive.

This thesis does not describe a complete process instrument. The work presented here has aimed at proving the viability of the measurement concept. Further research and development is still needed to make this instrument a serially manufactured industrial product.

1.2 AUTHOR'S CONTRIBUTION

My personal contribution to the project is concentrated on the following fields: thin film mirror design, film manufacturing method development (film dipping mechanics, film thickness control), measurement optics development, and result analysis. I have designed and built all measurement electronics used in this work. The signal processing methods and software used in this work are my own contribution, as well. I have also designed and built all the manufacturing equipment presented in this work.

This work presents a new approach to measuring pH with optical indicator dyes. The dielectric mirror based reflective measurement method forms the core of this

work. The feasibility of this measurement method has been demonstrated by measurements carried out with a prototype. The prototype has some new optical solutions which make it potentially inexpensive and simple to manufacture.

A new three-wavelength measurement algorithm has been developed for indicator dye measurements. This algorithm is tolerant against dye leaching from the thin film. This is an important feature because it decreases the need for film replacement.

Some analysis, which is believed to be new, is carried out on cylindrical rod optical components which are utilized in the prototype construction. Also, some new control methods are introduced to thin film wet deposition (dipping) control, most importantly a method for depositing a large number of different thin film mirrors on a single substrate. Further, a method of achieving the high precision position measurement required in manufacturing method development has been developed.

Three patents have been applied to protect the innovations presented in this thesis. One of these patents covers the main measurement principle (see appendix A), and this patent has already been issued. The other two patents applications cover dipping methods described in chapter 7 and reference measurement methods described in chapter 9. In these two applications I am the only inventor.

1.3 ORGANIZATION OF THIS THESIS

The concept of pH is a nontrivial one. While the simple definition of pH gives an illusion of a simple concept, the practical situation is rather different. The theoretical background to pH and different ways of measuring it are given in chapter 2.

Chapter 3 discusses the sol-gel process used in producing the thin films required in the sensor structures. The basic chemistry is introduced, but the main emphasis is set on the physical and chemical properties of sol-gel glass and possible applications of the sol-gel process.

Thin film indicator measurements are introduced in chapter 4. The chapter discusses different possibilities of measuring the color of a thin film in a process instrument. The core innovation presented of this work is described in this chapter.

Chapter 5 introduces the optical theory behind dielectric mirrors required in this application. Analysis is carried out on porous thin film mirror stacks in varying refractive index environment and on highly absorbing dielectric films.

Chapter 6 takes a sidestep to manufacturing technology and introduces a high resolution position measurement method required in controlling the thin film manufacturing process. A method for significantly improving the accuracy of a commercial position sensor is shown in this chapter.

The measurement methods and equipment required to characterize the deposited dielectric mirrors are discussed in chapter 7. A new and simple method of depositing a large number of mirrors with different characteristics on a single substrate is demonstrated.

Chapter 8 returns to the indicator pH measurement. A novel method of determining the pH from an indicator film by using three monochromatic light sources is developed. A short discussion of the optical considerations concerning the instrument is also included with an outline of the electronic construction. Possible error sources of this measurement are identified and discussed.

The actual optical construction of the instrument is presented in chapter 9. The theory of short light mixing rods is developed and simulation results are shown. As to my knowledge, similar simulations have not been presented before.

Chapter 10 describes the prototype which has been constructed to demonstrate the possibilities of the new pH measurement method. The mechanical, electronic, and optical constructions are shown in detail along with the signal processing methods.

Measurement results obtained by the prototype are presented in chapter 11 along with discussion. Results are shown on the measurement accuracy and sensitivity to optical properties of the surrounding medium. Some interesting dynamic phenomena in the indicator response are also demonstrated.

Chapter 12 concludes the thesis and discusses future prospects of the new measurement method.

Some supplementary information has been attached to this thesis as appendices. Appendix A contains the patent which covers the sensing element structure. Appendix B shows the theoretical background of the signal obtained by the position measuring instrument used in dipper position measurements (chapter 6).

2 About pH

2.1 DEFINITION OF pH

The common schoolbook definition of pH usually tells pH is an acidicity scale where 7 is neutral, smaller values acidic, and higher values basic [1, p.646]. While this definition is practical in many laboratory applications, it is not generally valid.

The definition of pH was originally proposed by the Danish chemist Sørenson [2] in 1909. The original form of the definition was that pH is the negative Briggs (i.e. base ten) logarithm of the hydrogen ion concentration:

$$\mathrm{pH} = -\log_{10} C_{\mathrm{H}^+} \tag{2.1}$$

where C_{H^+} denotes the concentration of H^+ ions in the solution.

Sørenson's definition has later been found to be slightly inaccurate; if the ion concentration is high, ion-ion interactions and other phenomena will decrease the number of hydrogen ions available to reactions. So, the modern definition of pH is that it is a measure of the number of active $\mathrm{H_3O^+}$ (hydronium, oxonium) ions in a solution, more precisely a negative base-ten logarithm of the concentration of active $\mathrm{H_3O^+}$ ions: [3]

$$\mathrm{pH} = -\log_{10} a_{\mathrm{H_3O^+}} \tag{2.2}$$

where a_{X} denotes the concentration of active X. In this definition H^+ ions are replaced by hydronium ions as the H^+ ions in the solution are bonded to water molecules ($\mathrm{H_3O^+}$, $\mathrm{H_2O_5^+}$). This does not, however, make any difference to the qualitative treatment below.

This definition has its shortcomings. Strictly speaking, equation (2.2) is mathematically wrong, as a logarithm cannot be taken from anything else than a bare number,

so this equation should actually be expressed in the form (2.3) to be mathematically more precise.

$$\mathrm{pH} = \log_{10} \frac{1\ \mathrm{mol/dm^3}}{a_{\mathrm{H_3O^+}}} \tag{2.3}$$

This form requires one "magic number", a constant just taken out of nowhere. This indicates that the pH scale has been defined rather arbitrarily; there could be several equally good definitions or even better definitions. The reason why Sørenson chose the scale was its logarithmic nature. Almost all solutions needed by Sørenson did give nice positive pH values, and the definition is easy to remember.

The number of active H_3O^+ ions is not enough to define acidicity of a solution. To understand the dynamics, it is important to understand what happens in pure water. Liquid water consists of water molecules, H_2O, which are more or less grouped together but can still slide past each other [1, p.446]. Water molecules are rather tightly bonded, but they do occasionally undergo autoprotolysis (autoionization), i.e. a single proton (H^+, hydrogen ion) is removed from the water molecule and joined to another water molecule:

$$2H_2O \rightleftharpoons OH^- + H_3O^+ \tag{2.4}$$

At fixed temperature the probability of a molecule gaining enough energy to undergo protolysis is constant. Similarly, when a hydronium ion meets a hydroxide ion, they unite with a certain probability.

In a constant volume the number of molecules splitting during a constant period of time depends on the splitting probability (p_{split}) multiplied by the total number of molecules. Similarly, the number of reuniting ions depends on the probability of uniting when the ions meet (p_{unite}) multiplied by the number of each type of ions.

$$\begin{cases} n_{\mathrm{split}} = p_{\mathrm{split}}\, n_{\mathrm{H_2O}} \\ n_{\mathrm{unite}} = p_{\mathrm{unite}}\, n_{\mathrm{H_3O^+}}\, n_{\mathrm{OH^-}} \end{cases} \tag{2.5}$$

In order to preserve dynamic balance in the system, n_{unite} must be equal to n_{split}. This gives the following result:

$$p_{\mathrm{split}}\, n_{\mathrm{H_2O}} = p_{\mathrm{unite}}\, n_{\mathrm{H_3O^+}}\, n_{\mathrm{OH^-}} \tag{2.6}$$

$$\frac{p_{\mathrm{split}}}{p_{\mathrm{unite}}} = \frac{n_{\mathrm{OH^-}}\, n_{\mathrm{H_3O^+}}}{n_{\mathrm{H_2O}}} \tag{2.7}$$

So, the product of number of hydroxide ions and number of hydrogen ions divided by the number of water molecules is a constant. This can be written with active ion

concentrations:

$$\frac{a_{\mathrm{OH^-}}\, a_{\mathrm{H_3O^+}}}{a_{\mathrm{H_2O}}} = K_a \tag{2.8}$$

where K_a is the acid dissociation constant of water ($\approx 1.8 \times 10^{-16}$ mol/dm^3). This equation is in accordance with the law of mass action.

As the protolysis does not significantly change the H_2O concentration (the number of protolyzed molecules is small compared to the total number of molecules), this may be simplified:

$$a_{\mathrm{OH^-}}\, a_{\mathrm{H_3O^+}} = K_W \tag{2.9}$$

where K_W is the ionic product of water ($\approx 1.008 \times 10^{-14}$ mol/dm^3). If there is nothing else than water in the solution, the number of H_3O^+ and OH^- ions must be equal. This way the concentration of H_3O^+ ions can be solved:

$$a^2_{\mathrm{H_3O^+}} = a_{\mathrm{H_3O^+}}\, a_{\mathrm{OH^-}} \tag{2.10}$$

$$a_{\mathrm{H_3O^+}} = \sqrt{K_W} \tag{2.11}$$

By application of (2.3) this number gives pH = 7.00 for a neutral solution. Unfortunately, this number is valid only at 25 °C. Furthermore, with one more decimal the figure is 6.998, so neutral pH is not exactly 7 even at this temperature. As a historical side note, Sørensen's value for K_W was 0.72×10^{-14} at 18 °C which gives neutral pH = 7.07 at that temperature.[1] Thus, the fact that the pH of a neutral solution is almost exactly 7 at 25 °C is just a coincidence.

At higher temperatures the thermal energy increases and the probability of protolysis increases accordingly. This will increase the ionic product, and that way decrease the pH for neutral solution.

Assuming there is some certain threshold energy which must be exceeded to protolyze a molecule, the probability of protolysis is given by the Arrhenius equation:

$$p_{\mathrm{split}} \propto e^{-\frac{E_a}{RT}} \tag{2.12}$$

where E_a is the activation energy required in the protolysis reaction.

This, in turn, is directly proportional to the ionic constant. By combining equations (2.3), (2.10), and (2.12) the neutral pH changes as the inverse of the absolute temperature:

$$\mathrm{pH} = A + \frac{B}{T} \tag{2.13}$$

where A and B are constants.

[1] Modern value for neutral pH at 18 °C is 7.12.

While this calculation is rather simplified, it gives a fair qualitative description of what happens when the temperature changes. The tabulated value for pH at 100 °C is in fact 6.13, so the effect is by no means insignificant. For pure water a commonly accepted empirical equation for K_W is [4]:

$$K_W = 10^{14.00 - 0.0331\,^\circ\mathrm{C}^{-1}(t - 25\,^\circ\mathrm{C}) + 0.00017\,^\circ\mathrm{C}^{-2}(t - 25\,^\circ\mathrm{C})^2} \tag{2.14}$$

where t is the temperature in °C.

The situation becomes even more complicated when there are other ions present in the solution. The H_3O^+ and OH^- ions of the water participate in other processes taking place in the solution. Their ionic product is maintained by (2.9) but there are other similar dynamic equilibria in the solution. As different reactions have different threshold energies, their thermal behavior is not uniform. This way both the pH and its effects may change in a complicated pattern.

In many cases the H_3O^+ and OH^- trigger the same reactions to opposite directions. This may be thought of as a competition over protons between OH^- and other negative ions in the solution. Thus it might have been a better choice to define pH as the quotient of H_3O^+ and OH^- ions in the solution. This would have kept the neutral point fixed.

On the other hand, not all processes are such that both OH^- and H_3O^+ may participate. For example, there may be a membrane which permits the passing of H_3O^+ ions but not that of OH^- ions. In this case the absolute pH may be a better measurement unit.

It should also be borne in mind that the definition of pH is not limited to aqueous solutions, pH can be calculated for any solution with H_3O^+ ions. However, the practical significance of pH is almost exclusively limited to aqueous solutions, and pH measurements in non-aqueous solutions are more difficult due to the small number of ions. For example, the pH of ethanol at 25 °C is approximately 9.55. [6]

2.2 SIGNIFICANCE OF pH

pH measurements find numerous applications in process industry. The simplest application is a straightforward pH adjustment by adding a base or acid to a solution. This adjustment may be important in, e.g., avoiding scaling of surfaces by solid deposits, avoiding corrosion, or changing the physical properties of the process medium.

Many organic syntheses are sensitive to pH. pH may change the time required by the process or even the products from the process. The same applies to biochemical and biological processes. In the wrong pH the bacteria may die or undesired bacteria may start to grow.

While the pH is known to affect many chemical processes, the actual mechanisms are not always known. Especially with more complicated syntheses or biochemical processes, the actual processes are not well understood.

2.3 PRACTICAL CONSIDERATIONS

Due to its rather complicated behavior, measuring the pH is difficult in industrial processes. Most industrial processes run at a high temperature, and most of the time the desired pH values have been determined experimentally. This empirical approach includes the instrument. Thus, the knowledge is not actually "at this pH our process runs well" but rather "when this pH sensor indicates this pH, the process runs well". [5]

The routine calibration of industrial pH sensors is carried out by taking samples from the process and analyzing them in the laboratory. If the sample temperature changes on its way from the process pipe to the laboratory, the results may be erroneous. There may even be some irreversible processes, such as polymerization, so that reheating the sample does not bring it back to its original state.

As long as the handling of the calibration sample is the same from one calibration to another, the measurement may function well. The results given by the laboratory are not measured under the same conditions, but they still give an indication of the process conditions. This way the instrument may be verified and even adjusted to give satisfactory—while not correct in the absolute sense—results.

In industrial applications the control loops try to keep the process conditions constant. Thus it is important the instrument gives correct readings at one single set-point. At other points the most important thing is to know if the parameter is too high or too low and to have some idea if it is slightly too high or slightly too low.

Neutralization processes form one important application field for the pH measurement instruments. They are rather difficult from the control loop point of view [6]. Even rather large errors in the neutralizer feed give rather small pH errors. A ten percent error in the H^+ concentration is only 0.05 in pH. So, when the instruments note the pH change, the neutralizer feed has to be changed a lot. This makes it necessary to use large gain in the feedback, which may make the process unstable.

The absolute accuracy of industrial pH instruments is rather modest. The manufacturers may claim high accuracies, almost all instruments give two decimals, some even three. These claims are not in accord with industrial experience, accredited pH measurement laboratories have estimated the measurement uncertainty (95 %) to be approximately ±0.2 units for industrial measurement instruments [7]. Well-calibrated

laboratory instruments may reach ±0.02 pH unit repeatability and ±0.05 pH unit accuracy [8].

It should be noted that for any industrial process repeatability is more important than the absolute measurement result. This is especially true with pH measurements, where drift is a much worse problem than any constant error. However, long term repeatability and accuracy are close to each other, highly repeatable instrument can be made accurate by adjusting the output.

2.4 MEASUREMENT METHODS

Several methods of pH measurement have been developed over the years. The most common method is the use of glass membrane electrodes. Some newer electronic methods have been developed with semiconductors, and the ISFET sensors have some applications especially in biomedicine.

There are also some optical indicator methods. Optical indicators have been and are used a lot in laboratory applications, especially in quick manual applications they are fast, inexpensive, and reliable.

An ideal pH measurement system should measure the pH directly from the bulk of the medium, not from an interface between the medium and some measuring surface. Optical indicators do this, but as there is no way to collect the molecules back from the solution, they contaminate the solution and are thus not useful in in-line use as such.

Another desired property of a pH measurement instrument is selectivity. It should respond only to H_3O^+ ions, not to any other chemicals in the solution. In practice, this goal cannot be achieved by any known method, at least, there are always chemicals which attack and damage the measurement surface or break optical indicator molecules.

In principle, certain spectroscopic methods might be useful in non-contact bulk measurement of pH. In practice, spectroscopic methods are very good at finding trace amounts of certain chemicals, but their quantitative properties are difficult to control. Also, spectroscopy is often too expensive and sensitive to environmental conditions to be used in process instrumentation.

2.4.1 Glass membrane sensors

When a liquid comes in touch with a specially composed glass surface, an electromotive force (voltage) arises across the surface. This e.m.f. is given by the Nernst

equation:

$$E = E_0 - \frac{RT}{nF} \ln Q \tag{2.15}$$

where E_0 is a constant, Q is the reaction quotient from the mass action law (e.g., K_a of water in (2.8)), F is the Faraday constant (96 485 As/mol), R the molar gas constant (8.314 J/mol/K), and n is the number of electrons transferred in the process. For the H_3O^+sensitive glass electrode this can be written as:

$$E = E_0 - \frac{RT}{F} \ln a_{H_3O^+} \tag{2.16}$$

where $a_{H_3O^+}$ is the activity of the hydronium ions. This can also be written as:

$$E = E_0 - \frac{RT \ln 10}{F} \log_{10} a_{H_3O^+} = E_0 - \frac{RT \ln 10}{F} \mathrm{pH} \tag{2.17}$$

i.e. a voltage which has some constant bias and then a variable part which depends on the pH.

Unfortunately, this potential cannot be measured directly. Any galvanic connections to either the liquid or the glass membrane will introduce another electrochemical half-cell which has its own Nernst potential.

In practice, there is another liquid (reference solution) on the other side of the glass surface. Thus two potentials form, and as these potentials are in series, they are subtracted from each other to give potential

$$E_g = E_0 - \frac{RT}{F} \ln a_{H_3O^+} - (E_0 - \frac{RT}{F} \ln a'_{H_3O^+}) = \frac{RT \ln 10}{F} (\mathrm{pH}' - \mathrm{pH}) \tag{2.18}$$

across the membrane. $a'_{H_3O^+}$ is the H_3O^+ ion activity in the reference solution and pH′ the pH of the reference solution. The constant part of the potential is cancelled out, so a voltage results which is proportional to the difference between pH on either side of the membrane.

This voltage can be measured between the liquids on either side of the glass membrane. However, if an electrode is placed to either side, there will be different potentials between the electrode and the liquid surrounding it due to different composition of the liquids. This can be avoided by using a reference electrode configuration (figure 2.1).

The solution in the reference electrode is similar to that inside the pH sensitive electrode. Also, the metal electrodes interfacing with the liquid are similar. The only difference between the two electrodes is that one has the ion-selective glass membrane,

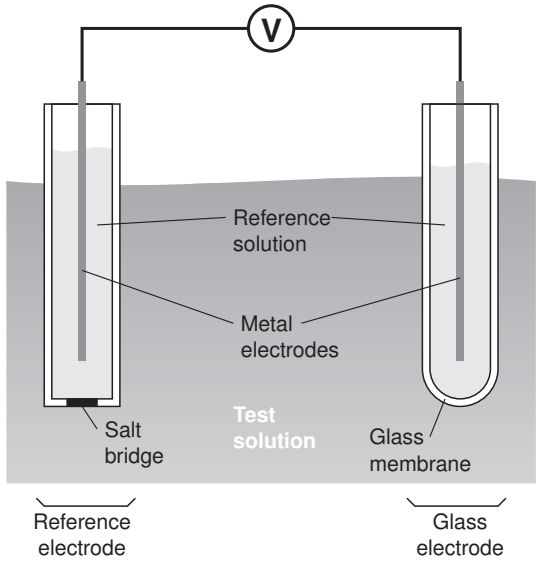


Figure 2.1 A glass membrane pH sensor

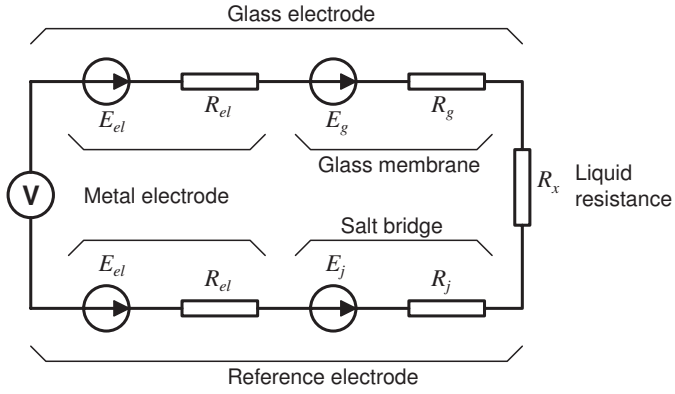


Figure 2.2 Equivalent circuit of the glass membrane pH sensor

and another has direct contact between the two liquids in the form of a salt bridge (i.e. porous material which is permeable to ions).

Figure 2.2 shows the electric circuit with its main components. From this diagram the measured potential can be calculated:

$$V = -E_{el} - E_j + E_g + E_{el} = E_g - E_l \tag{2.19}$$

The liquid junction potential E_j is a diffusion potential which depends on the ion transfer in the salt bridge. This potential is usually rather small compared to the potential E_g created by the pH difference across the membrane. However, the junction potential may be highly variable, and as it depends on the liquid flow in the salt bridge, it may depend on process liquid flow around the sensor [9].

Usually, the reference liquid used in the electrodes is a saturated KCl solution. The use of KCl is advantageous because K^+ and Cl^- ions have similar mobilities and thus they do not introduce any additional error in the salt bridge. The galvanic contact to the reference solution is usually made with a Ag/AgCl electrode. From (2.18) and (2.19) the measured potential may be written as:

$$V = E_C + 59.2\text{mV} \times \text{pH} \tag{2.20}$$

where E_C is a constant. The temperature is taken to be 25 °C, and the small e.m.f. of the liquid junction has been omitted for clarity.

In practice, reference and measurement electrodes can be combined to one electrode as shown in figure 2.3. This does not change the operational principle explained above.

The actual process how the glass membrane potential forms is rather a complicated one. Outer layers of glass hydrolyse, and there is ion exchange between the alkali ions in the glass and hydrogen ions in the surrounding solution. The process takes place in a very thin layer on the glass surface. Damaging this layer will deteriorate the measurement. Also, when the glass surface is depleted of alkali ions, the electrode has to be replaced. [5]

In practice, the voltage output of the sensor is not very accurate. The sensors exhibit wear, and their response deteriorates over time. This wear is not uniform, i.e. it is most pronounced in the pH region where the electrode is used. Also, if the sensor is cleaned thoroughly or let dry, its outmost hydrated glass layer is removed, and it takes several hours or days for a new layer to form. This makes it more difficult to clean the sensor surface if there is some scaling.

As the current path goes through the glass and through the process liquid (figure 2.2), the measurement impedance is large, from hundreds of megaohms to gigaohms.

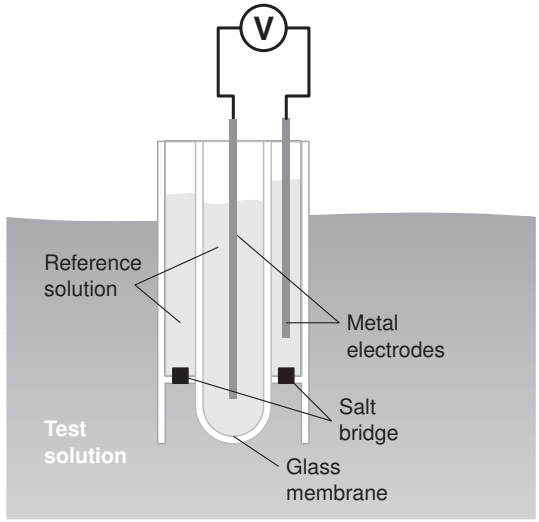


Figure 2.3 A glass membrane pH sensor with integrated reference electrode

Small voltages and high impedances make the measurement sensitive to small error currents, and capacitive coupling of noise is a significant problem.

The impedance of the measurement depends on the process medium conductivity. If the liquid is nearly neutral and has few other ions in it, there is very little ion exchange in the hydrated layer, and the measurement is sensitive to errors. Water cleaning applications are among the most difficult pH measurement applications.

Glass electrodes are not very welcome in the food industry because they may break and glass shards may end up in the final product. Also, high pressure environment is difficult as the reference electrode depends on the liquid junction. High pressures may slow the response or even lead to reference electrode contamination as ions from the process medium enter the reference electrode through the porous interface. One solution is to pressurize the reference electrode, which adds complexity to the system.

Almost all commercially available glass membrane pH measurement instruments are claimed to be temperature compensated. This temperature compensation refers to the compensation of the electrode, i.e. the T term in equation (2.18). This does not mean the electrodes would be compensated so that they give the same reading for neutral in all temperatures, which is a common misconception.

Glass electrodes are not strictly selective, they are sensitive to other ions, as well.

The most well-known errors are associated with small positive ions, e.g. Na^+. Solutions which have a high concentration of these ions will give erroneous readings from the electrode.

Despite their peculiarities, glass membrane sensors are the mainstream of pH measurement. They are relatively inexpensive and rather repeatable, at least in the short term.

2.4.2 ISFET sensors

In 1970 Bergveld introduced the ISFET (Ion-Sensitive Field Effect Transistor), a field-effect transistor (FET) where the gate connection was replaced by liquid (figure 2.4). In a FET the gate electrode attracts free charge to the surface of the semiconductor substrate and hence a conductive channel is formed. In the ISFET the charge collected on the surface of the insulating gate oxide induces the channel formation.

The reaction between the oxide surface and the surrounding liquid is similar to that in the glass membrane sensor, the gate oxide is in fact silica glass (SiO_2). So, the gate voltage is proportional to the pH by the Nernst equation. While Bergvald originally thought no reference electrode is required with ISFET, it was soon noticed that also this construction does need a reference electrode [10].

In a FET the drain-source current depends on the gate voltage. There are, however, several parameters which affect this relation. In practice, the ISFET measurement is easiest made by keeping the drain-source current of the ISFET constant by adjusting the voltage between the reference potential and the semiconductor substrate. This voltage is then similar to the voltage given by glass membrane electrodes (2.20).

There have been several attempts to replace the reference electrode by another semiconductor device. One possibility would be to use two ISFETs with different sensitivities; then the potential difference between the two devices would indicate the pH. These reference FETs (REFET) have been investigated during the last years [11], but there do not seem to be industrial applications yet.

As ISFET sensors have a bare semiconductor die exposed to the liquid under measurement, they are not very rugged, and they are usually not suitable for continuous measurements [12]. ISFET sensors have found their way to analytical systems especially in the biomedical arena, where small size and simple structure are important.

There are several variants of the ISFETs (e.g., MEMFET, CHEMFET, SURFET), and the technology was adopted to measurement of other ions quite early [10]. It seems that the main application development of FET-based chemical sensors is more in the microanalysis and biochemical/biomedical field (e.g., [13]) than in industrial measurements. The only advantage of FET-based pH sensors in industrial applications

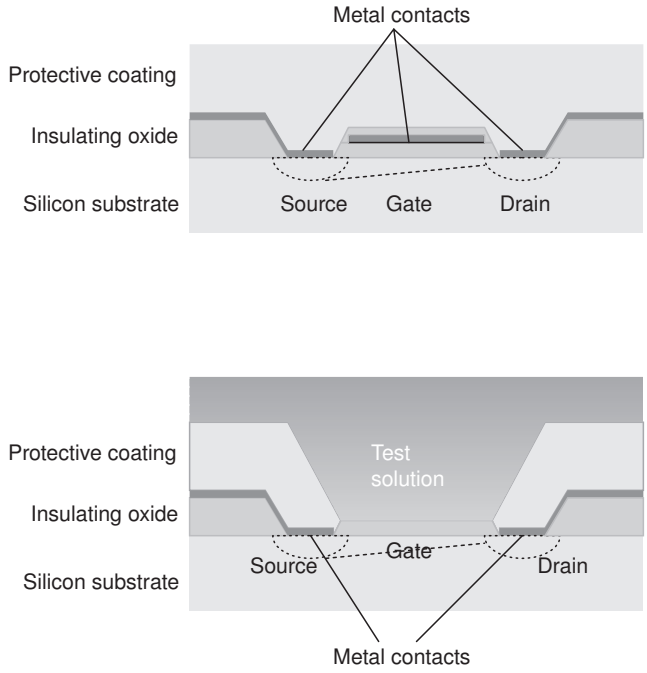


Figure 2.4 Conventional FET (upper) and ISFET (lower)

is the absence of the glass membrane. This makes it possible to make a complete no-glass structure which can be safely used also in food industry.

2.4.3 Optical indicators

The oldest way to measure pH is based on the use of optical indicators. There is a large variety of substances acting as optical pH indicators. Different indicators have a different region of color change, and by choosing a suitable indicator, the measurement can be made rather accurate even with the unaided eye as the only measurement device.

The color in an indicator molecule arises from different vibrations in the molecule [14]. Usually, the color of organic molecules is explained by *chromophores*. Chromophores are molecular structures which are essential in the resonance which pro-

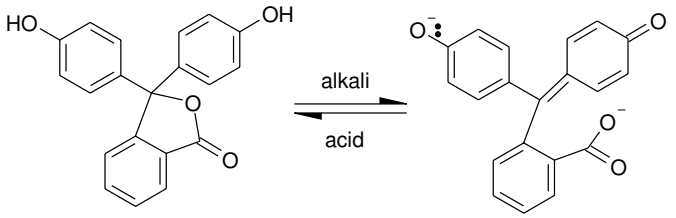


Figure 2.5 Acid and base forms of an indicator dye (phenolphthalein).

duces absorption. Chromophores alone do not necessarily produce color, but when auxiliary groups (auxochromes) are attached on the molecule, it exhibits absorption at certain wavelengths.

An indicator molecule changes its structure (usually by losing or gaining some groups) in response to external conditions. This change of structure changes the chromophore action, and so the absorption spectrum (color) of the molecule changes.

Simple optical indicator molecules have two different states, acid and base. They are usually weak acids or bases which have different optical resonances depending on their state. There is a dynamic equilibrium between these states, so that the proportion of acid form and base form indicator molecules (ions) depends on the external H_3O^+ and OH^- concentration. Figure 2.5 shows an example of this (phenolphthalein dye).

The concentration of indicator molecules mixed in the process should be low so that they do not significantly affect the process itself. This is usually rather easy to arrange as the indicator molecules have very high absorption of visual wavelengths, even a small concentration is clearly visible.

If an indicator molecule has two possible states, its spectrum has some interesting properties. Let us take a molecule with two forms with absorptivity coefficients $a_1(\lambda)$ and $a_2(\lambda)$. If the concentrations of the molecules in the two states are c_1 and c_2, then the total absorptivity a_{tot} is:

$$a_{tot} = \frac{c_1 a_1(\lambda) + c_2 a_2(\lambda)}{c_1 + c_2} \tag{2.21}$$

If there exists λ where the absorption of the two states is the same, i.e. if the absorptivity curves cross each other at any point, the absorption at this wavelength is constant ($a_{tot}(\lambda) = a_1(\lambda) = a_2(\lambda)$) regardless of the pH. This point is called the *isosbestic* point [3], and it is useful as a reference in some applications. Figure 2.6 shows

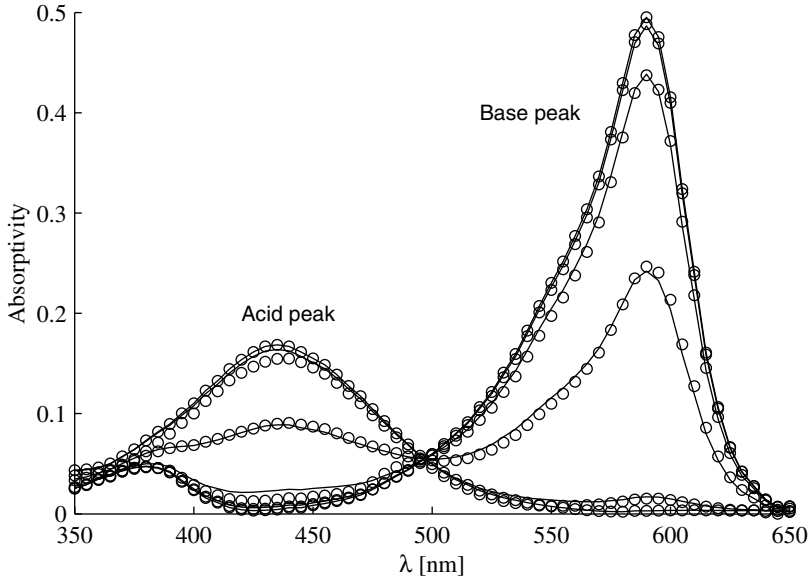


Figure 2.6 Measured absorption of BPB indicator in aqueous solution in different pH (dots) and interpolated data from the binary dye model (solid line). [15]

the absorption spectrum of bromophenol blue (BPB) in different pH environments and the interpolated values calculated from (2.21). The agreement between measurement results and calculated data is good while not exact.

Optical indicators are not perfectly selective, either. The effect of other ions or other dissolved material depends a lot on the indicator, there is no general rule on what is harmful and what is not.

Optical indicators may indicate other properties than pH, as well. For example, there are indicators for certain metal ions, such as aluminon ($C_{22}H_{23}N_3O_9$) which is sensitive to Al^{3+} ions. However, pH indicators are the best known, and they have the largest commercial value.

The most important advantage of optical indicators is their repeatability. A molecule does not wear in use. An indicator molecule is in one of a finite number of discrete states, and unless it disintegrates, it reacts always the same way under the same conditions. Usually, when the molecules disintegrate, the chromophore action becomes impossible and the absorption bands are outside of the visual spectrum.

Liquid phase

Optical indicators are widely used in liquid phase, where the indicators are dissolved into the liquid to be measured. The transmittance T of a solution is given by the Beer-Lambert law:

$$T = e^{-a(\lambda)bc} \tag{2.22}$$

where $a(\lambda)$ is the absorptivity coefficient, b optical path length, and c concentration of the indicator.

With colorless liquids the pH measurement with an optical indicator is straightforward, a known concentration of the indicator is added to the solution, and the absorption is measured at some known wavelength. If the exact concentration of the indicator cannot be controlled, the absorption can be measured at the isosbestic point and at another wavelength to compensate for the concentration variation.

This method is especially useful in titration, where one of the reagents is slowly added until the pH reaches the equivalence point indicated by color change of the indicator.

The measurement becomes more difficult if the liquid itself has some color or scattering particles in it. This can be compensated for by comparing the absorption before adding the indicator to that after the indicator has dissolved. However, this may be difficult in titration if the titration process itself produces solid particles or color into the liquid.

The largest advantage of dissolving the indicator into the liquid is the bulk nature of the measurement. There are no measurement surfaces to be contaminated, and even the spatial distribution of the pH changes may be seen (figure 2.7).

The disadvantage of this method is its irreversible nature. There are no reasonable ways to collect the indicator molecules for reuse from the liquid. This way liquid phase use of color indicators is limited to on-line or laboratory use where the product is not returned to the process.

The required liquid amounts are small, even microanalytics has been carried out with this method [16]. In the industrial use, all on-line and laboratory methods have the problem of getting a representative sample, and also the time delay may be unbearable.

Fiberoptic (MIR) sensors

An interesting way around the limitations of the optical indicators is to trap them in a porous substrate. The pore size of the substrate must be large enough to let the H_3O^+ and OH^- ions in and small enough to not let the indicator molecules out. This way

Figure 2.7 Spatial distribution of pH changes is clearly visible in titration with an indicator dye

the indicator molecules will change their state according to the external pH but will not be dissolved into the liquid around them.

One commonly used way to arrange the measurement is to coat an optical fiber with the porous substance (figure 2.8) [17]. The porous substance should have a lower refractive index than the fibre, so that light coming along the fiber will undergo total internal reflection in the border of the fiber core and the indicator layer. In each reflection the evanescent field is slightly attenuated, and as there is a large number of reflections, the total attenuation is easily measurable. These sensors are usually called either Multiple Internal Reflections (MIR) sensors or evanescent field sensors.

The porous matrix may be made of several substances. There are polymer-based plastic matrices (e.g., [18]) as well as sol-gel glass matrices (e.g. [19]). The most important issues are the pore size and chemical and physical durability of the matrix.

A large number of papers have been published on fiberoptic pH sensors [20] but

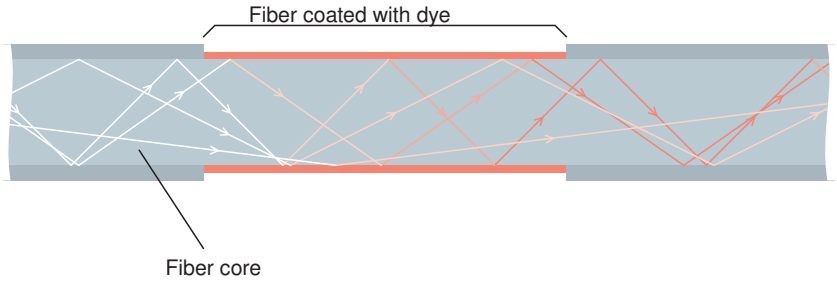


Figure 2.8 Principle of the MIR pH sensor

the actual number of industrial applications is small. There are numerous patents in the field (e.g. [21, 22]) but the applications are not in the mainstream process industry. There are some commercially available fiberoptic pH sensors, but most of these sensors (such as [23]) use a separate sensing element, and the fiber is only used for carrying light.

The largest challenges of the fiber probes relate to their fragility in process environment and to the rather complicated optics associated with the fibers. On the other hand, some rather interesting sensor systems produce information on the spatial pH distribution by using time-domain reflectometry in a fiber. Unfortunately, the repeatability is not very good in these systems [18].

Again, the main applications of fiber optic probes are in the biomedical field, as small fiber probes may be inserted inside the human body.

Thin film indicators

Another way to use optical indicators is to implant the molecules on a thin film applied on a glass substrate. This approach is more rugged than the fiber approach, and it retains the advantages of optical measurement. Substrates coated with a thin film indicator layer act as a reusable piece of lithmus paper, as shown in figure 2.9.

Thin film indicators have been used previously [24], but as it is difficult to measure light passing through the layer without having two separate windows (see 4.1), the measurement has not been applied to process instruments previously.

This thesis describes a way to circumvent this problem, and the instrument utilizing this method is believed to have the advantages of MIR measurement but fewer disadvantages.

Figure 2.9 A pH sensitive thin film undergoes reversible color changes. (Acid solution is on the left, base on the right).

Reflective indicator-based measurement is not a perfect measurement method, either. It is not bulk-based, and it has all the inselective behavior of the respective indicator.

The main advantages of the novel device are repeatability, insensitivity to pressure, and the relatively low manufacturing cost. These advantages open up an application field which is potentially very large.

3 Sol-Gel technology

Sol-gel technology is a process where ceramic materials (including glasses) are prepared from liquids through synthesis and polymerization. Brinker and Scherer [25, p.xi] define sol-gel as: "the preparation of ceramic materials by preparation of a sol, gelation of the sol, and removal of the solvent".

The concept of the sol-gel process is illustrated in figure 3.1. The process starts by making a solution (sol) with suitable reagents. One of the basic components of the sol is the precursor which carries the metal atom desired in the end product. This solution will gelate under certain conditions to form one large polymeric molecule with liquid captured inside (gel). This gel may behave as a viscous liquid or even be a jelly-like semi-solid structure (alcogel). The final product (sol-gel glass) is obtained after drying the liquid away from the structures.

In this thesis sol-gel technology is utilized to create porous thin film structures. Sol-gel can be used to numerous other applications, as well, and new applications are found all the time. The following paragraphs give a brief overview of the basics and applications of the sol-gel technology.

3.1 BRIEF HISTORY OF SOL-GEL

The French chemist Jacques-Joseph Ebelmen (1814 – 1852) was the first to synthesize a metal alkoxide (a molecule with a metal atom in the center and alkoxy groups surrounding it) in 1840s. He prepared tetraethoxysilane (TEOS) by treating $SiCl_4$ with ethanol. The resulting clear liquid had the tendency of becoming a gel under atmospheric conditions. This gel then gradually became solid.

Ebelmen was known to be interested in ceramics and mineralogy. However, he did not further develop the idea of forming ceramics in low temperatures through the sol-gel process, and the invention of the sol-gel process faded into oblivion for the following decades.

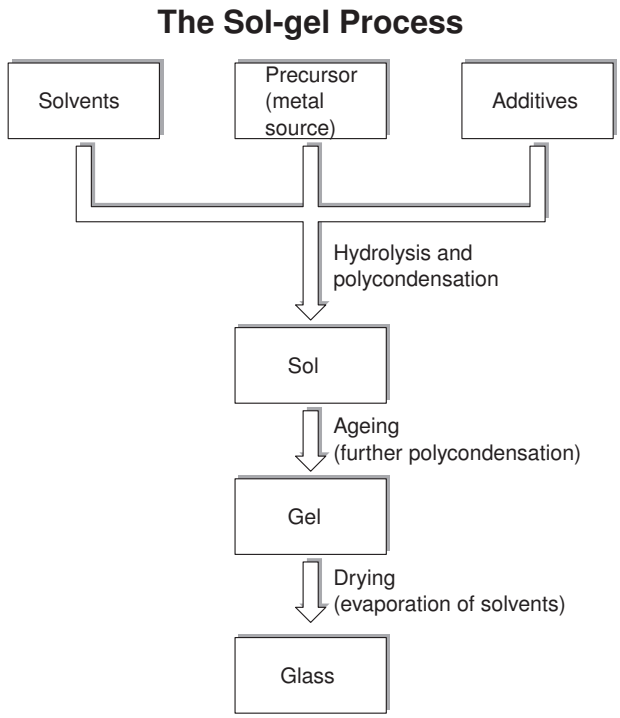


Figure 3.1 Steps of the sol-gel process

The first commercially interesting applications were found in the 1930s when the German glass manufacturer Schott & Genossen became interested in creating optical oxide films by the sol-gel process. At that time thin film coatings were new in optics, and the modern methods (such as vacuum deposition) did not exist in large scale. So, the invention [26] might have become very important, but for some reason the technology did not advance very fast. Schott's process did not use the precursors later commonly used (such as TEOS), and it is possible that the process was too slow to be practical.

After the World War II sol-gel technology found some niches, such as preparation of homogeneous powders and spheres in ceramic and even nuclear industry

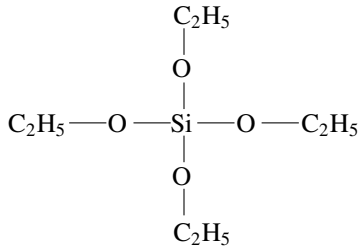


Figure 3.2 Structure of tetraethyl ortosilicate

[27]. However, the technology did not gain much more interest until late 1970s, when Yoldas [28] and some other researchers published results on the possibility of producing glass monoliths by the sol-gel process.

The development of the sol-gel technology has been rapid during the last decade. Numerous new applications have emerged, and emerge continuously. While the preparation of monoliths seems to have triggered the fast development, it was soon found to be a difficult process to control and did not find any commercial applications. For instance, the review given by Brinker and Scherer in the beginning of 1990s [25, p.12] states:

> This [the start of sol-gel explosion by monoliths] is a bit ironic in retrospect, as it is evident that monoliths are the least important of the potential application of gels.

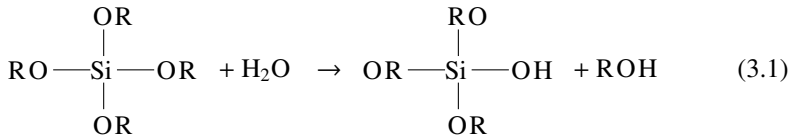
In a later retrospect even this statement seems a bit ironic, as one of the most important manufacturers of non-spherical custom optical components uses sol-gel process to mold high-precision optical components [29, 30].

3.2 BASIC CHEMISTRY OF THE SOL-GEL PROCESS

The sol-gel process can be used to produce very different results, inorganic and inorganic/organic films, powder, fibers or bulk material. This introduction concentrates on making metal oxides from metal alkoxide precursors, inorganic/organic and other more complicated systems are not discussed.

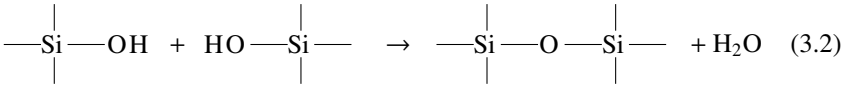
The overall reaction starts with a metal alkoxide, i.e. a molecule which has a metal atom in the middle and alkoxy groups around it. An example of this is tetraethoxysilane (or tetraethyl ortosilicate, TEOS) depicted in figure 3.2.

The first part of the process is hydrolysis, a reaction between the metal alkoxide and water. In this reaction water molecules replace the alkoxy groups with hydroxide groups. The result is a metal hydroxide and alcohol, in this case silicon hydroxide and ethanol:



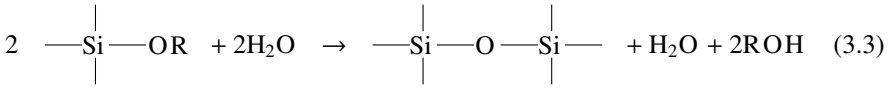
(3.1)

The metal hydroxides will react with each other and with free alkoxide groups. In this reaction water is released, the result is an oxide link between the two metal molecules plus the released water molecule. This part of the process is called polycondensation.



(3.2)

Thus, for each link the overall reaction is the following:



(3.3)

The net result of the process is one consumed water molecule and two released alcohol molecules. The original reaction reported by Ebelmen in 1847 obtained its water from the humidity in the ambient atmosphere. The reaction is then very slow, but polycondensation will occur within the course of several months.

If there is an excess of water, all alkoxy groups will be removed in the process. The result is then simply a metal oxide network. In this case each silicon atom has four siloxane bonds, i.e. it is bond to four oxygens. As all these oxygens are shared by two silicon atoms, the resulting polymer is a SiO_2 polymer.

While the stoichiometric result is clear, it does not dictate the macroscopic form of the resulting substance. The inorganic polymer may be in the form of powder, small spheres, films, or even bulk material. Also, the porosity of these materials may vary in wide limits.

It is important to understand that these two processes, hydrolysis and polycondensation, are accompanied with a physical process, drying. When the polymer has been formed, it is in the form of a gel, a large single molecule with a continuous liquid phase in the molecule. In the macroscopic world this could be analogous to a sponge. During the drying process the solvents evaporate from the gel, and a solid polymer is formed.

The process thus has three important phases whose rates determine the final product: hydrolysis, polycondensation, and drying. The rate of the chemical processes is adjusted primarily by changing the ratio of the precursors and by using catalysts. The drying process may be adjusted by controlling the physical environment in which drying takes place, its temperature and possible solvent atmosphere. Drying rate may also be affected by introducing special drying control additives (DCCA).

While there is a very large number of possible variables in the process, there are some basic rules which apply to the process.

3.2.1 Solvents

Even though the hydrolysis or polycondensation reactions do not necessarily need any solvents (except for water), usually a solvent is added to the initial solution. While the solvents do not participate in the reactions, they do change the physical properties of the sol. Adding solvents homogenizes the solution and adjusts its viscosity. The increase of volume decreases the concentration of the reagents and thus affects the reaction rates. And, finally, the relative change in solvent concentration is smaller during the reaction, if there is already a significant amount of the solvent in the solution before any hydrolysis takes place.

Most often the solvent is the same alcohol as in the alkoxy group. This is not a necessary requirement, and in some cases the solvent may be chosen to give the desired porosity of the final product. Larger solvent molecules tend to give lower porosities, probably due to the smaller number of pores.

The choice of solvent affects also the drying time of the gels. Larger molecules tend to evaporate more slowly, and thus give longer drying times. [30]

In thin film applications it is usually desirable to have high solvent-to-alkoxide ratios to give low-viscosity sols which produce thin films and gel in a rather slow pace. Typically, the ratios may be around 10 mols of solvent per one mol of alkoxide. [25, pp. 97–228]

3.2.2 Water-to-alkoxide ratio

One of the most important parameters in a sol-gel synthesis is the water-to-alkoxide ratio (usually denoted by r). If there is an excess of water ($r > 2$ in the case of TEOS), the hydrolysis process is fast. This way almost all alkoxy groups are replaced by hydroxides before polycondensation. This does not necessarily make the overall reaction any faster, as the polycondensation reaction becomes slower. Especially if there is a large excess of water, a reverse reaction to polycondensation may occur, the metal-oxide bonds start to hydrolyse.

The amount of water may also be too small for the hydrolysis to occur completely if there is less water than the stoichiometric amount for the reaction to be complete ($r < 2$). This will affect the physical properties of the resulting gel and, for example, sol-gel fibres are drawn from this type of sols.

The amount of excess water naturally changes the viscosity of the sol, as does the excess solvent. One benefit of using excess water in thin film applications is the longer storage time of the sol before gelation, as the polycondensation is very slow (months). This benefit does not apply to monoliths, as they should preferably gel quickly (minutes or hours).

3.2.3 Catalysts

In most cases the polycondensation reactions have rather high threshold energies and are thus slow. Even though the TEOS prepared by Ebelmen may have gelated accidentally due to the humidity in the air, the process is impractically slow (months). To make the polycondensation faster, catalysts are used.

The actual catalysis reactions are complicated and not completely known, but usually the main idea is to adjust the pH of the solution so that the desired polycondensation and hydrolysis rate is achieved. Both acid and base catalysts may be used, and the optimum pH depends on the metal alkoxide.

In silica thin film production the reaction is acid catalyzed so that the pH is around 2. In this range the polycondensation is very slow compared to the hydrolysis, so that practically all alkoxides are hydrolysed before the polymer starts to form. Also, the sol produced with this method is easy to store even for extended periods of time. The sols used in this research project have remained liquid for over a year.

Catalysts may also be required to avoid precipitation. The titanate-borosilicate sols used in the structures introduced in this thesis require strong acid catalysts and prehydrolysis of TEOS to remain soluble.

3.2.4 Drying

To obtain the final product the solvent has to be removed from the alcogel. The rate of removing the solvents affects the physical properties of the end product very significantly.

When the solvent is removed from the alcogel, the gel tends to shrink significantly. As the solvent molecules depart from the pores of the gel, the pore walls tend to get nearer. The mechanisms which make this change are related to the surface tension and capillary pressure effects, so the physical properties of the liquid in the alcogel change this behavior. The shrinkage may vary in large limits; ordinary xerogels tend to shrink very much, often in a ratio of 10:1. Aerogels (very porous structures), on the other hand, may shrink only some percents.

One of the obstacles in sol-gel technology is cracking during drying. While the structure shrinks, non-uniform shrinking tends to break it. Bulk material dries first at the surface, so that the surface shrinks before the inner parts.

Films shrink in a rather more uniform manner, as their small thickness offers easier evaporation of the solvents. However, films tend to shrink in all three dimensions, whereas the substrate allows shrinkage only in one direction (thickness). This causes stress both on the film-substrate interface and into the film itself, and thick films tend to crack or tear off the surface.

Cracking may be prevented by a careful control of the drying conditions. These methods, however, are more successful with bulk gels than with films, as they do not make the substrate any more flexible.

The simplest drying process is to let the gel dry in room temperature and ambient atmosphere. This approach, however, produces glasses with high porosity and water and solvent content. The physical properties of these glasses are not suitable for most uses, so the drying process usually takes place in elevated temperatures.

The elevated temperatures may vary from mildly elevated above room temperature to over a thousand degrees. The very high temperatures produce glass which is virtually indistiguishable from glasses produced by melting, as the bonds are rearranged in the amorphic substance [25, p.745].

Higher drying temperatures produce physically more durable results with smaller porosity. Also the chemical durability increases as the surface area per unit volume decreases. While higher durability is usually desirable in all applications, the decrease in porosity and increase in density may not be wanted. This way the choice of drying temperatures is usually a compromise.

It should be borne in mind that the solvent is usually a mixture of different solvents, most often an alcohol and water. The drying process may be changed by drying the system under solvent atmosphere, either under an atmosphere saturated with one

of the solvents or with all solvents. As the physical properties of the solvents are different, the resulting structure properties depend on the evaporation order of the solvents.

As a summary, it is possible to affect the resulting material properties by a suitable choice of drying conditions. Unfortunately, the number of parameters is very high, and finding the optimum drying procedure for each application requires a significant amount of work.

Aerogels

It is possible to dry the gel so that the polymeric matrix does not shrink. This process gives an aerogel, which is a very porous structure. In order to prevent shrinkage and cracking it is important to minimize the forces in the liquid-solid interface.

Aerogel drying is a challenging process which involves heating the alcogel at a high pressure so that the environment is beyond the critical point for the solvent. As the temperature and pressure exceed the critical point, the two different phases merge into supercritical fluid. This fluid has the density of the liquid but gas-like viscosity and diffusivity. So, the supercritical solvent occupies the same volume as the liquid solvent but can be extracted without exerting the forces associated with liquid-solid interfaces.

The challenges of this process are due to the high pressures and temperatures associated. The critical point of water is approximately 375 °C and 22 MPa (or 220 bar), and the critical point of ethanol is equally high. The common way to work around these extreme conditions is to use liquid carbon dioxide as the solvent.

First the alcogel is placed in the autoclave which is pressurized to a high pressure so that liquid carbon dioxide may exist in room temperature (> 3 MPa). In this pressure the aerogel is rinsed with liquid CO_2 so that the solvent in the structure is changed into CO_2.

The structure is then further pressurized and heated above the critical point (approximately 31 °C and 7.3 MPa). These temperatures and pressures are much more benign and safer to use. After this the pressure is slowly released so that the system will come to the gas phase of the carbon dioxide which is then slowly released from the structure. The actual duration of the process depends on the size of the structure but is from several hours to days.

New aerogel preparation processes have been developed since the mid 1990s. These methods involve sol structures which are able to ”spring back” after the gel has dried. So, the structure is allowed to collapse, but it will return to its original form after the pores are empty. This method has recently found its way to commercial use in film and bulk aerogel production. [31]

3.3 PROPERTIES OF SOL-GEL GLASS

The sol-gel (metal oxide) glass can be observed as a solid mixture of the glass material and air. Its physical and chemical properties are dictated by those of the specific oxide and the structure of the material. As a simplification, the structure of the material can be expressed by a single parameter, porosity. While the pore size and shape affects some properties, it much more difficult to take into account.

The simplest physical property is the density of the material. If the porosity (the volume percentage of pores) is p, then the density of the glass is:

$$\varrho = (1 - p)\varrho_{bulk} + p\varrho_{air} \tag{3.4}$$

where ϱ_{bulk} is the density of the bulk material.

If the porosity differs significantly from unity, then the density of the air (1.29 kg/m^3 at normal temperature and pressure) can be neglected. Even though this would appear to be the case always, there are reports of some extremely light silica aerogels. The lightest of these seems to be the gel presented by Hrubesh et al. [32], which has the density of approximately 3 kg/m^3. This is probably the lightest solid material ever produced. As the density of silica is approximately 2 500 kg/m^3, this means that the porosity is in the order of 99.9 %.

Another important physical material constant is the dielectric constant. The dielectric constant cannot be calculated directly, molecular polarizability has to be used. Molecular polarizability is a molecular property, and the relation between dielectric constant and molecular polarizability is given by the Clausius–Mossotti equation[1]:

$$\alpha = \frac{3}{4\pi N}\frac{\epsilon - 1}{\epsilon + 2} \tag{3.5}$$

or

$$\frac{\epsilon - 1}{\epsilon + 2} = N\frac{4\pi\alpha}{3} \tag{3.6}$$

where ϵ is the dielectric constant and N the number of molecules per volume. This equation is only an approximation, and especially with high dielectric constants the actual values may be different.

If the pores of a material are empty (i.e. vacuum), then the number of molecules per unit volume is smaller. The dielectric constant ϵ_p of a porous material and that of the bulk material (ϵ_b) are linked by:

$$\frac{\epsilon_p - 1}{\epsilon_p + 2} = (1 - p)N\frac{4\pi\alpha}{3} = (1 - p)\frac{\epsilon_b - 1}{\epsilon_b + 2} \tag{3.7}$$

[1] A similar equation is also known as the Lorentz–Lorenz formula. The difference between these two is the way how they have been discovered.

This approximation is valid if the pores are filled by gas, as the polarizability of gases is practically zero. If the pores are filled with a liquid or a solid, the approximation is not valid. In that case the average polarity weighted by the volume fractions gives an estimate of the dielectric constant of the porous material:

$$\frac{\epsilon_p - 1}{\epsilon_p + 2} = (1 - p)\frac{\epsilon_b - 1}{\epsilon_b + 2} + p\frac{\epsilon_l - 1}{\epsilon_l + 2} \tag{3.8}$$

where ϵ_l is the dielectric constant of the pore-filling material.

This approximation is known as the *ion polarizability additivity rule* and its main applications are in crystalline materials. Its validity in microporous amorphous solid-liquid structures has not been widely demonstrated.[2]

One of the advantages of sol-gel materials is the large extent to which their optical properties can be tailored. The intrinsic refractive index (square root of the relative dielectric constant) of silica is approximately 1.46, and the most porous aerogels have a refractive index less than 1.01. On the other hand, the refractive index of titania, another popular sol-gel material, is over 2.2 in its amorphous (glass) form.

The optical properties of sol-gel glasses are generally good. However, even the small pores may scatter light in very short wavelengths, and very porous aerogels tend to have yellowish transmission. Also, organic substances in the material may turn yellow during drying at elevated temperatures.

Other physical properties of the material are harder to predict. It seems obvious that as the porosity increases, the material becomes more brittle. However, there is no simple and straightforward relation between the mechanical properties and porosity, as the actual structure is very important in rigidity.

The same applies to chemical durability. As the porosity increases, there is a significant increase in the surface area. The increase, however, depends on the pore size. Also, the pore size may be so small that the attacking chemical may not enter the pores, or at least there is very little flow around the inner surfaces compared to the outer surface of the object.

The chemical durability of sol-gel glass can be improved by adding other metal oxides to form a binary or multi-component system. For example, ordinary silica glass is vulnerable to a base attack, but if a small amount (a few percents) of boron is added, the resulting borosilicate structure becomes much more durable [35]. This is equivalent to the borosilicate (e.g. Pyrex) glass prepared with traditional methods. Also zirconium can be used in this purpose.

[2] Other formulae have been used, as well. For example, Yoldas [33, 34] uses formula $(\epsilon_p - 1)/(\epsilon_b - 1) = 1 - p$ for porous films in vacuum (air).

An interesting property of porous sol-gel glass is the possibility of implanting molecules into the structure. As the manufacturing process does not require high temperatures, even rather large organic molecules can be trapped into the structure. This property is essential to the sensor structures presented in this thesis.

3.4 THIN FILM MANUFACTURING METHODS

There are several ways to produce thin films from the sol. The aim is to produce a uniform layer of sol which then gels and forms a uniform thin film after drying. The uniformity requirement depends on the use of the film. Optical coatings have to be very uniform especially in multi-layer applications, whereas coatings improving scratch resistance just have to be thick enough.

The most popular coating methods are spinning and dipping. In spinning the substrate is rotated quickly so that the sol is spread around the substrate. Dipping means withdrawing the substrate from the sol at a predefined rate.

In addition to these methods there are several less frequently used methods. The sol may be simply painted on the substrate. If the gelation and evaporation rates are slow, this may give a sufficiently even surface. However, the thickness of a painted surface is difficult to control. The sol may also be sprayed on the surface, in which case the thickness control is easier. Small droplets have large surface area compared to their volume, so the evaporation and other processes associated to it may change the gelation behavior. Also, even though the local thickness variations can be kept are small, the global thickness variations tend to be large.

According to [36] a new continuous process has been developed to produce highly repeatable (5 – 10 % thickness variation) results with very high throughput. This method uses high-pressure spraying nozzles.

The spraying result depends on the size distribution of the droplets. Experiments with ultrasonic pulverization to produce aerosol with highly controllable monodisperse droplets have been performed. This technique is still rather complicated, and the results obtained this far have not been as good as those with the traditional methods [37].

A simple way of coating large substrates is to simply pour the sol on the substrate. The thickness of the resulting film depends on the viscosity of the sol and the inclination angle of the substrate to be coated. Unfortunately, the film thus produced is thicker in the bottom of the substrate, so the method is not suitable for optical coatings.

For large flat surfaces so called meniscus method can be utilized. In meniscus coating a wave wipes the surface of the glass. This method is promising in coating

large surfaces in a continuous process. The actual commercial usage of the meniscus method (or any other large-volume method) is difficult to estimate as the coating methods are usually trade secrets of the manufacturers.

One of the major questions in coating technology is how to coat irregular substrates. One possibility is to use electrophoresis, where electrically charged particles go towards anode or cathode. This method requires the substrate to be conductive and also sets some requirements to the particles, so that it has not advanced to be an industrial sol-gel coating method.

The choice of the coating method does affect the final properties of the film. For example, the evaporation rate during spinning is higher than that during dipping, so the spinned films are thinner and less porous than equivalent dipped films.

3.4.1 Spinning

Basic spinning is probably the simplest coating method to arrange and to control. The substrate is spinned at a controlled speed and the sol is then spread on the substrate.

Usually, the spinning process is divided into four phases: deposition, spin-up, spin-off, and evaporation [25, p.795]. During the deposition phase an excess of liquid is poured over the stationary substrate. The substrate is then spun up to its final speed, and most of the liquid is thrown over the edge of the surface due to rotational forces.

The rotational velocity is kept constant during the spin-off phase, and the liquid film on the substrate becomes uniform. In the evaporation phase the solvents evaporate from the gel to give the final film.

The spinning process has the property of smoothing out all thickness variations over time. However, it should be noted that this behavior requires the coating liquid to be Newtonian, i.e. its viscosity does not depend on the shear rate.

Sols are not necessarily Newtonian in the shear rate range used in the process. Also, the viscosity of the sols is a function of the time. During the spinning the sol polymerizes into a gel film and a large part of the solvents evaporate. Both processes contribute to the increase of the viscosity during the process.

The final thickness of a film (h) is given by the semi-empirical formula proposed by Meyerhofer to be used in resist spinning [38]:

$$h = \left(1 - \frac{\varrho'_a}{\varrho_a}\right)\left(\frac{3\eta m}{2\varrho'_a\omega^2}\right)^{1/3} \tag{3.9}$$

where ϱ'_a is the initial mass per volume of volatile solvents and ϱ_a that after spinning. η is the viscosity, m evaporation rate of the solvents, and ω the angular velocity.

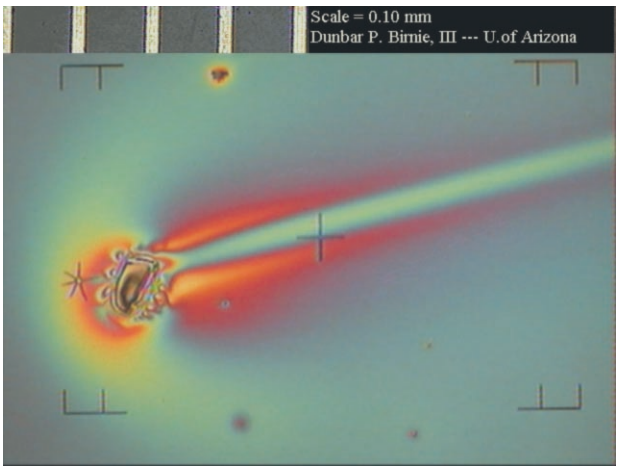


Figure 3.3 Damage due to dust during spinning (image from [39]).

However, as there are already some empirical parameters in the equation, it is often easier to use a fully empirical two-parameter model: [36]

$$h = A\omega^{-B} \tag{3.10}$$

where A and B are empirical constants.

The main advantage of spinning is the simplicity of controlling the rotational velocity accurately. Even a simple electric motor arrangement will give accurate and smooth rotation. Also, commercial spinning equipment is readily available and widely used in semiconductor industry.

Probably the most serious limitation of spinning is that the substrate must have rotational symmetry. In principle, the fluid flow is radial during the spinning, but during the spin-up there are significant tangential forces, as well. On the other hand, the radial forces associated with rotation are very large compared to the gravitational forces, so the substrate does not have to be flat.

Another practical difficulty with the spin coating method is its sensitivity to contamination. Particle impurities in the sol or from the environment tend to produce a radially outwards directed damage behind them (figure 3.3). Thus a single particle may produce a large damage area.

Spinning is naturally not suitable when working with very large substrates, as the radial forces grow proportional to the radius.

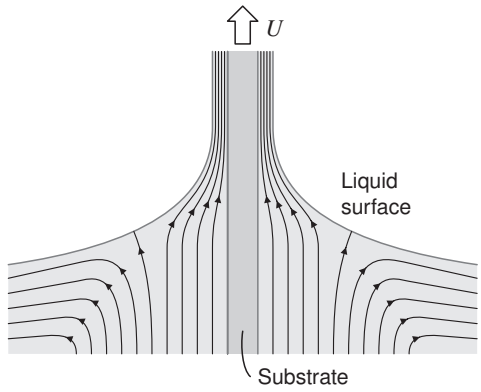


Figure 3.4 Flow of liquid during withdrawal

3.4.2 Dipping

Dipping is another important film deposition method. In its simplest form the substrate is immersed in the sol and withdrawn from it in the direction of the normal of the liquid.

The physics of dipping is illustrated in figure 3.4. As the substrate is drawn away from the liquid, it captures some of the liquid with it. The liquid surface is bent upwards to join the surface of the substrate tangentially. A part of the liquid drawn with the substrate returns to the bath. The thickness of the dipped layer is mainly a function of the viscosity of the sol and the withdrawal speed.

There are several forces associated with dipping [25, p.789]. The liquid is dragged upwards by the viscous drag induced by the moving substrate. The force of gravity limits the thickness of the layer by pulling the liquid back to the bath. The moving liquid has some inertia which tends to pull it with the substrate. The surface tension shapes the curved liquid surface, and the surface tension gradient in different parts of the region do also exert forces. Along with these forces the disjoining (or conjoining) pressure resulting from molecular interactions on the substrate-liquid interface becomes an important factor when the film is of submicrometer thickness.

A commonly cited equation for the film thickness is the Landau-Levich equation [25, p.790], [36]:

$$h = 0.94 \frac{(\eta U)^{2/3}}{\gamma_{LV}^{1/6} (\varrho g)^{1/2}} \tag{3.11}$$

where η is the viscosity and γ_{LV} the ratio between viscous drag and liquid-vapor surface tension. U is the withdrawal speed of the substrate from the sol, ϱ density of the sol, and g the gravitational acceleration.

These equations assume Newtonian viscosity, and thus they are not necessarily accurate in sol-gel applications. Also, the number of experimental parameters is large, and thus their use in predicting the film thickness is limited. In practice, the most important thing to note from the equation is that increasing the withdrawal speed or sol viscosity both increase film thickness.

There are exceptions even to these rules of thumb. If the sol is a particulate sol (composed of nanoparticles), the thickness behavior may even be reversed; higher withdrawal speed produce thinner layers.

The actual evaporation rate—which in turn affects the film thickness—depends on the ambient atmosphere immediately around the substrate. If the air is very still, the surrounding air saturates with the solvent, as the diffusion is very slow. In practice, however, the large scale movement in the air is far more important than diffusion, so the rate of evaporation has to be determined experimentally.

Dipping is probably the most popular sol-gel thin film deposition method when large surfaces are coated with optical coatings. Unlike spinning, it can be scaled up rather easily. Actually, dipping large panels may be easier than dipping small substrates as the mass makes the substrate movement smoother. Dipping is also less sensitive to particles. In the first phase of the dipping process the substrate is immersed into the sol, which prevents the deposition of dust particles onto the substrate. If a particle hits the surface during the evaporation phase, it may stick there. This produces a pointlike damage, opposed to the long streaks produced by impurities in spinning.

A new method of using skew dipping (i.e. the substrate does move in an angle to the vertical) appears to be promising in dielectric filter manufacturing. The film formed on the top side of the substrate is thicker than that on the bottom side. In this way two layers with different thicknesses are formed simultaenously. By varying the angle and the dip speed the layer thicknesses can be chosen independently. [36]

Dipping is a relatively economical manufacturing method. In principle, all material removed from the bath ends up on the substrate, whereas spinning and spraying tend to lose a significant part of the coating material. Unfortunately, the pot life of the sol may not be very long, and thus it is estimated [36] that only 20 % of the sol is really used as a coating.

Dip coated film is uniform even with non-Newtonian liquids. The movement during the deposition process is the same for each part of the surface, so dipped films are inherently more uniform than spinned films. This factor becomes more important with multi-layer coatings where the uniformity requirements are stricter.

Dip coating is difficult to use on curved surfaces. There have been some trials with bottle coatings and even with eyeglass coatings to improve the impact resistance. However, with curved surfaces the contact angle between the bath and the object is different in different parts of the object, so that the coating will become too non-uniform.

Dip coating equipment is also more complicated than spin coating equipment, especially with small substrates. It is much more difficult to produce very smooth, controllable and slow movement than to produce smooth and controllable fast rotation. While commercial spinners are available at almost all cleanrooms, there are few small dippers.

One of the disadvantages associated with dipping is the difficulty of protecting one side of the substrate, if only one side of the substrate is to be coated. Also, dipping is difficult to realize as a continuous process. These factors make other methods (such as the meniscus method) favored in some applications.

After experimenting with both methods, dipping was chosen as the film producing method for the project described in this thesis. Dipping seems to produce better quality films in non-cleanroom conditions. Also, the square substrates used are better suited to dipping. The third reason is that the commercial production is easier to implement with dipping, and using the same method in research gives more applicable results.

3.5 APPLICATIONS

The number of sol-gel applications has grown quickly during the last few years. The first applications were optical thin film coatings. As sol-gel technology can be utilized to form almost any metal oxide film, it enables the use of some films difficult to deposit by dry deposition methods. Also, vacuum deposition techniques may be difficult to scale up, and large area (several meters across) substrates are easier to coat with a liquid phase process. [40]

The wide range of refractive indices offered by sol-gel glass films is also an important factor in thin film coatings. Dielectric mirror applications benefit directly from the large refractive index differences.

The refractive index of an ideal single layer anti-reflective coating should be the geometric mean of the refractive indices of the substrate and the surrounding medium. For a glass-air interface this is approximately 1.23, which is a very low refractive index for a solid material. On the other hand, this refractive index corresponds to approximately 50 % porosity, which is not difficult to obtain.

A very interesting technological opportunity lies in graded index structures. An anti-reflective coating with continuously changing refractive index between the refrac-

tive index of the substrate and that of the surrounding environment exhibits very small reflection over a large range of wavelengths and angles. These structures have been demonstrated already in 1984 [33] but due to the difficult manufacturing of aerogels they do not have any common industrial applications so far.

It is also possible to deposit piezoelectric PZT films with the sol-gel process. The sol-gel process is attractive in this application as the stoichiometry of the film is easier to control than with vapor deposition methods.

Bulk sol-gel material is being used in a growing number of applications. It is competing with high-temperature glass casting process in custom optical element manufacturing. The main advantages of the sol-gel process are the low temperature and—rather surprisingly—shrinkage. As the gel shrinks considerably, all manufacturing errors in the mold also shrink by the same factor. This is especially important with diffractive optical elements with very small features. Sol-gel thick films and bulk material is also used in diffractive optics replication processes. [41]

Highly porous aerogels have very low thermal conductivity and they can be used as thermal insulators. Glass aerogels are good in this application also due to their ability to withstand high temperatures. The highly porous structure lends itself to different filtration applications, as well as to gas adsorption applications. As porous materials have very low dielectric constant, their use in semiconductor insulation has been considered. An important use is in Cherenkov detectors. [42]

An interesting use of aerogel is as impact reducer in space applications. The structure is light but slows down impacting objects very effectively without getting completely crushed. The Stardust cometary dust collector uses aerogel collecting structures [43].

Bulk sol-gel processing can be used with doped glass. An interesting application is in optical memories. The organic molecules used in optical memories decompose in high temperatures, and ordinary glass casting cannot be used.

Sol-gel material is intensively researched to produce optical microchips suitable for integrated optics. There are several companies and research groups working on this. Some rather promising results have been obtained, [44] presents one of them.

Sol-gel processes complement the traditional glass process, as the latter is relatively simple and very inexpensive. For example, making sheet glass out of sol-gel would require the use of molds whereas the modern float glass process is a continuous process. On the other hand, sol-gel glass seems to enable completely new applications.

4 Reflective thin film indicator measurement

As mentioned earlier, it is possible to trap optical indicator molecules into a sol-gel glass film. The color density of the indicator will depend on the film thickness and indicator density, but will be sufficient to measurement purposes even with very thin (in the order of 100 nm) films.

The glass slide shown in figure 2.9 has been coated with a bromophenol blue doped silica film. The film undergoes a reversible color change when put in solutions with different pH.

In order to take advantage of this color change, a practical measurement method has to be developed. The method should be able to detect relatively small color changes with high accuracy.

4.1 DIRECT COLOR MEASUREMENT

In the simplest measurement setup the light is transmitted through the substrate and film, and the transmission spectrum is measured (figure 4.1).

There are essentially two different ways of arranging the measurement. Either the spectrum of the light source is adjustable (e.g. by means of a monochromator) or the spectrum of the light which has passed through the system is measured in the detector.

This measurement is sensitive to several error factors. If the intensity of the light source is not measured before the light goes through the film, the light source has to be very stable or the measurement results are inaccurate. This can usually be avoided by a simple reference arrangement.

In real process measurements a much more difficult problem is the effect of sample cell window coating and the error due to the color of the process liquid. These effects can be avoided by using a two-ray variant of the measurement principle with one

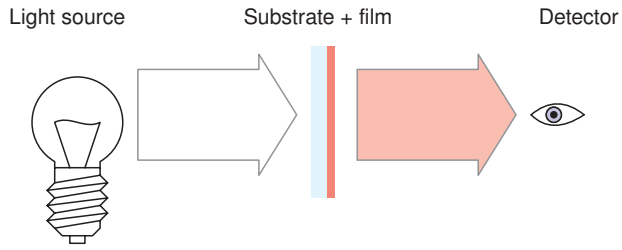


Figure 4.1 Simple transmission measurement

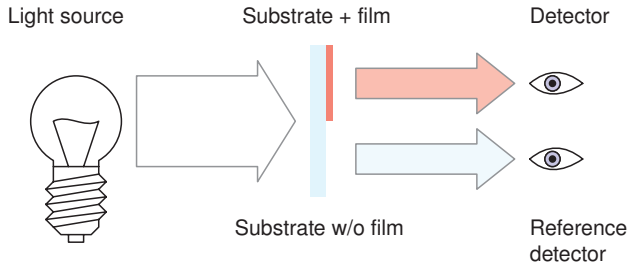


Figure 4.2 Transmission measurement with two rays

ray passing through the indicator film and another ray passing through a similar path without the film. (figure 4.2)

Even though this solution in principle compensates for the process liquid color and window scaling, it is not a viable solution in practice for three reasons:

1. It is difficult to ensure the part of the window with indicator and the part without receive the same amount of scaling on the window.
2. The process liquid has to be transparent enough. If a significant part of the light is extinguished on its path through the liquid, the measurement becomes noisy.
3. Two windows are required; one with indicator and the other on the opposite side of the process pipe.

The first problem—uneven scaling of the windows—may be possible to work around by using a striped indicator structure and suitable optics to resolve the different stripes. This way the hopefully smooth differences in scaling would be possible

to compensate for. Also, if the scaling absorbs light similarly on all wavelengths, it is possible to compensate pure intensity variations by taking advantage of the isosbestic point of the color (see 2.4.3).

The second problem cannot be worked around. Naturally, a smaller path length may be used to improve transmission, but the path length is determined by the process pipe size, and an in-line instrument must be able to accommodate to the requirements of the process. Often the pH measurement applications are with rather clear liquids, so this limitation does not apply to all possible applications.

The third limitation is the most difficult. It is straightforward to arrange one window into a process, but arranging two separate windows is challenging, especially because the optical path length or geometry must not change in vibration. Electrode pH measurement methods use the "single hole to the pipe" technology, and it is unlikely that a more complicated technology would be very attractive.

The second window can be replaced by a mirror. This, however, does not help much. It is very difficult to find mirror materials which would have specular reflection after extended periods of time in the process pipe. The mounting of the mirror would require tight tolerances (due to angular error doubling in reflection), and the path length doubles.

While the simple transmissive methods may by useful in some applications, they are not generally acceptable in in-line process instruments.

4.2 REFLECTION MEASUREMENT

A novel method of arranging the measurement was found (see appendix A). As the transmission method is not useful, the measurement may be carried out as a reflection measurement where the process window is coated with a reflective layer. In order for the measurement to work the layer has to be permeable to the ions to be measured (H_3O^+, OH^-), which rules out several simple solutions (such as vacuum deposited aluminium mirrors).

4.2.1 Diffuse surface layer

The thin film can be coated with some sufficiently diffuse layer. This diffuse surface may be obtained by coating the measurement layer with white pigment, e.g. TiO_2 or $BaSO_4$ crystals (figure 4.3).

The diffuse reflection mechanism of white pigment is based on the large number of reflections and total internal reflections in the pigment particles. If the pigment

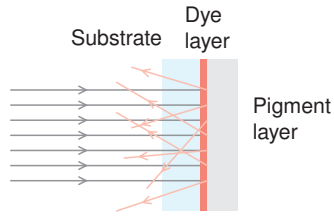


Figure 4.3 Reflective measurement with a diffuse surface

material is highly transmissive, the reflections are practically lossless, so even though the ray of light goes through a large number of reflections, the absorption is small.

Ideally, the light striking the pigmented layer is radiated in all directions according to the Lambertian distribution. This, however, requires a large number of reflections in the pigment layer. Consequently, the pigment layer has to be rather thick in order to both capture a sufficiently large part of the light and give a smooth reflection.

Because diffuse reflection spreads the light out in a wide angular distribution, imaging optics is inefficient in collecting the light. On the other hand, if the only goal is to collect the light to the detector, a system based on an integrating sphere is useful (figure 4.4). The diffuse light reflected towards the integrating sphere hits the sphere walls and is further diffusely reflected. After a number of bounces the light will hit either the detector or the window again.

If the pigment layer on the window is thick, it will pass very little light and thus the only absorbing elements in the system are the detector and the indicator film. In practice, the light from the light source has to be lead in somehow, for instance with a fiber, and that will introduce some small losses.

Despite the advantages of this measurement system, it is not generally easily realizable. If the pigment particles are to reflect light rather than scatter it, they have to be significantly larger than the wavelength of the light. As mentioned above, the thickness of the pigment layer has to be large to avoid the loss of light due to leakage into the process medium.

Constructing a thick pigment layer with sufficient ion permeability may be difficult. Diffusion in small pores is slow, and larger pores would enable contaminants from the process to to enter the layer. The pigment particles may be suspended into a sol but in general it is difficult to produce non-cracking sol-gel layers thicker than a micrometer.

The idea of the diffusing layer on top of an indicator layer has been exploited

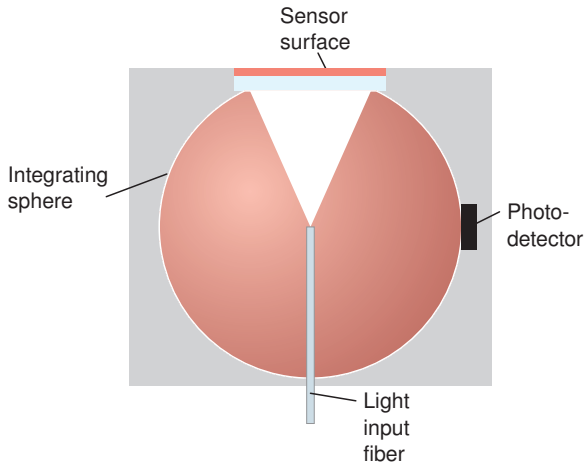


Figure 4.4 Reflective measurement from a diffuse surface with an integrating sphere

commercially by Kodak [45]. Their system is not, however, sol-gel based and it incorporates some very thick (on the sol-gel scale) films.

4.2.2 Diffuse interface reflection

A way to avoid the thick diffusing layers is to create diffuse reflection by using rough refractive index interface. The indicator layer is made rough and then coated with a high-index coating (figure 4.5). There a significant amount of reflection (even though no total internal reflection) from the interface between the low-index and high-index layers. As the interface is rough, the reflection goes to random directions.

This method will not produce Lambertian intensity distribution, as there is generally only one reflection. Also, the losses due to light leaking into the process are significant.

Again, the surface profile roughness has to be significant compared to the wavelength. This limits the maximum angle of the high and low index interface to the substrate surface. On the other hand, as the surface does not have to be flat, a thick and cracking film might be usable in this context because the cracks would form the required surface roughness.

The outer surface of the sensing element should preferably be flat as any rough surfaces tend to collect contamination more efficiently then smooth surfaces. The

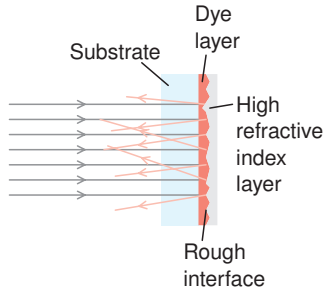


Figure 4.5 Diffuse reflection from a rough interface

diffuse interface has to be relatively thick, even though not as thick as in the case of pigments.

In practice, the controllably rough surface could possibly be manufactured by the methods described in [46]. However, as the reflection coefficient is only a few percents in a single interface, this method is more of a curiosity than a useful answer to the problem of reflective measurement.

4.2.3 Porous mirrors

From the optics point of view it is generally more efficient to use imaging optics in light collection. Diffuse surfaces spread the angular displacement without making the beam more collimated, and thus make light collection with imaging optics more difficult. A specular reflection, i.e. reflection for which the incident and reflection angles are the same, is usually more efficient from the optical point of view.

Most specular mirrors are metal mirrors. Metals have high reflectance and they are easy to deposit with vacuum deposition methods. The reflectance of a metal mirror would be sufficient in the reflective indicator film measurement. The problem lies in making the metal layer permeable to ions.

A solution would be perforating the mirror with small holes. However, if the holes are spaced far from each other, the ions have to travel long lateral distances in the indicator layer, which will take a lot of time. If the holes are close to each other, their combined area will deteriorate the reflection. Also, as the number of holes and their size increase, the exposed surface area increases and the mirror becomes vulnerable to chemical attacks.

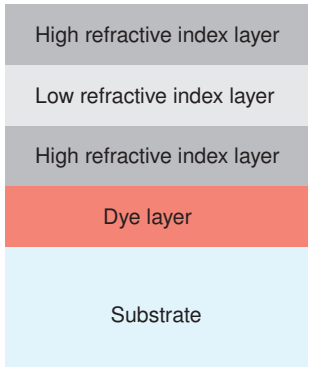


Figure 4.6 Structure of a porous dielectric mirror on top of an indicator layer

Aside from these theoretical issues there is the question of manufacturability. So far, there are no easy and inexpensive ways of producing suitably perforated mirrors in larger quantities. The holes can be made by optical lithography, but that would introduce a large number of new process steps. A suitable hole pattern could possibly be manufactured with laser drilling, but as each hole has to be drilled separately, this would take a lot of time and hence be expensive.

The reflectivity of a metal mirror is based on the conductivity of the metal surface. Another possible method of making mirrors is to use dielectric mirrors, i.e. form the mirror by using several non-conductive (and hence transparent) mirror layers. A mirror with this structure is depicted in figure 4.6.

The sol-gel process lends itself well to making dielectric mirrors, as all mirror layers can be made porous. Also, as mentioned earlier, the obtainable refractive index range is large, which makes it easier to have usable mirrors with fewer layers.

The dielectric mirror approach requires several layers of dielectric, which makes the layer stack thicker than a single layer would be. Fortunately, the layer thicknesses are in the order of a few dozens of nanometers to a few hundred nanometers. This way the total thickness of the stack remains reasonably low, and diffusion through the stack is possible.

One of the advantages of the dielectric mirror approach is the possibility to make the mirror using the same technology as the indicator layer. Once the correct parameters are found, the material costs of a sensor element are almost negligible, as the surface area is only a few square centimeters.

5 Dielectric mirrors

The theory of dielectric mirrors is based on the wave nature of light. The reflectivity of a dielectric mirror is a function of the refractive index and thickness of the mirror layers, the incident angle, wavelength, and polarization of the light. While the basic theory behind dielectric mirrors is fairly straightforward, the analytical equations associated with multi-layer mirrors become very complicated even with a small number of layers.

Dielectric mirrors and filters have been in use for over half a century [47]. Highly sophisticated structures may have over a hundred layers, and sol-gel dielectric filters have been manufactured with 30 – 40 layers [48]. The art of designing a dielectric mirror or filter is well developed.

The theory developed in this chapter is used in analyzing the thin film mirrors required in the sensing elements. This theory is essential for understanding the requirements set to the thin film stack.

While dielectric mirrors are well-known, there are some important differences between the ordinary dielectric mirrors and the mirrors used in the reflective indicator measurement application.

Usually, dielectric mirrors and filters are in constant refractive index environment, most often in air, whereas in this application the refractive index of the environment may change. In the case of porous films even the film refractive index may change as the environment refractive index changes. Also, the number of layers used in this application is limited as ions have to diffuse through the layers.

The following discussion will assume all materials to be linear, nonferroelectric and nonferromagnetic, at rest, and isotropic. These assumptions should not be limiting in this case, as the intensities used in the measurement application will not trigger significant non-linear behavior even in common non-linear materials, and dielectrics are non-magnetic at optical frequencies (i.e. their magnetic permeability is approximately the same as that of vacuum). The films are amorphous and thus isotropic.

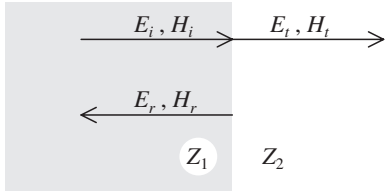


Figure 5.1 Wave arriving at an interface between two materials with different impedances.

5.1 REFLECTION BETWEEN TWO DIELECTRIC LAYERS

An interface between two media with different conductivity or dielectric constant will reflect part of the incident light. For a plane wave advancing in the direction of the normal of the surface, the reflection coefficients can be calculated, for instance, by using the impedances of the material. The following discussion is limited to normal incidence. It can be extended to non-normal incidence, but especially with absorbing media the formulae become unnecessarily complicated.

In non-absorbing media the plane wave is a transverse wave with the magnetic field vector, electric field vector, and the direction of propagation all perpendicular to each other. The ratio between the magnitude of the electric field vector ($\vec{E}$) and the magnetic vector ($\vec{H}$) is constant, the impedance of the medium (Z). The impedance can be calculated from the material parameters:

$$Z = \sqrt{\frac{\mu}{\epsilon}} = E/H \tag{5.1}$$

where E and H are scalar field amplitudes, ϵ dielectric constant of the material, and μ magnetic permeability of the material.

Figure 5.1 depicts a wave arriving at a boundary. Part of it is reflected and part of it is transmitted. The boundary conditions of the Maxwell equations state that the tangential (to the surface) components of E and H have to be continuous. In the case of normal incidence, the fields are completely tangential, so they can be expressed as scalars.

$$E_t = E_i + E_r \tag{5.2}$$

$$H_t = H_i - H_r \tag{5.3}$$

where E_i and H_i are the field amplitudes of the incident wave, E_r and H_r those of the reflected wave, and E_t and H_t of the transmitted wave.

It should be noted that the sign of H_r is negative. This sign change indicates the direction change of the wave (E, H, and the direction of propagation form a right-handed orthogonal triad). As the ratio between E and H for each wave is fixed, H can be eliminated from the equations (5.2):

$$E_i + E_r = E_t \tag{5.4}$$

$$\frac{E_i}{Z_1} - \frac{E_r}{Z_1} = \frac{E_t}{Z_2} \tag{5.5}$$

From these equations E_r and E_t can be solved in terms of E_i. By dividing each amplitude (reflection and transmission) by the incident wave amplitude, the reflection and transmission coefficients are obtained:

$$\rho = \frac{E_r}{E_i} = \frac{Z_2 - Z_1}{Z_2 + Z_1} \tag{5.6}$$

$$\tau = \frac{E_t}{E_i} = \frac{2Z_2}{Z_2 + Z_1} \tag{5.7}$$

where ρ is reflection and τ transmission coefficient.

These formulae represent a special case of the well-known Fresnel formulae.

In order to calculate the intensity (I) carried by a plane wave, the impedance is required:

$$I = \left|\frac{1}{2}E^* H\right| = \frac{1}{2}\left|\frac{E^2}{Z}\right| \tag{5.8}$$

The intensity reflection and transmission coefficients can then be calculated from the amplitude coefficients (5.6) and (5.7):

$$R = \left|\frac{E_r^2}{E_i^2}\right| = |\rho|^2 \tag{5.9}$$

$$T = \left|\frac{E_t^2/Z_2}{E_i^2/Z_1}\right| = \frac{Z_1}{Z_2}|\tau|^2 \tag{5.10}$$

Usually, only the intensity reflection and transmission coefficients are measurable, whereas the phase information in the amplitude coefficients cannot be measured directly.

5.2 CALCULATION OF A MULTI-LAYER FILM STACK

From the formulae (5.6) and (5.7) it can be seen that there is some amount of reflection whenever there is a change in the refractive index (impedance is reciprocal to the

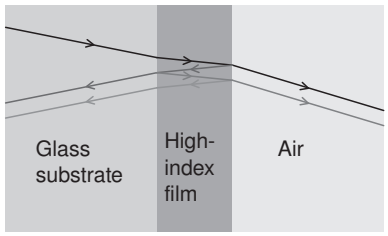


Figure 5.2 Reflections in a thin film.

refractive index, as will be shown later in this section). This phenomenon can be used to form a dielectric mirror. The simplest possible mirror is a single thin film over a substrate (figure 5.2).

There are two interfaces from which the light can be reflected. The first interface is a low-to-high interface between glass substrate and the high-index film. Second reflection is created at the interface between the film and the surrounding medium (vacuum or air in this case).

As the light propagates through any medium, its phase changes. The phase may also change in reflection. In this case the first reflection changes the phase by 180°, the second reflection does not change the phase. So, radiation reflected directly from the first surface has phase angle 180° to the incident radiation on the same interface. The radiation reflected from the second interface has the phase 2ϕ, where ϕ is the phase change the wave experiences during its travel through the layer. The second reflection does not introduce any extra phase changes as it is a high-to-low interface.

If the film is very thin, the phase difference of the two reflected waves is almost 180°, and their interference is destructive. If the film thickness is $\lambda/4$, i.e. the wave experiences a 90° phase shift in the film, both reflections are in the same phase (180° shifted), and their interference is constructive. In the latter case the film acts as a mirror, even though the actual percentage of reflected radiation is low.

The reflection and transmission coefficients of a thin film stack can be calculated from the reflection and transmission coefficients of each interface. However, even the simple case described above requires considerable effort, as the wave may undergo several reflections between the interfaces and thus an infinite series is formed.

In practice, the thin film stack calculations are performed with a matrix method [49, 50]. The method shown here is especially suitable for normal angle of incidence. It can be adapted to oblique angles but the calculation of matrix elements becomes more complicated, and the more often used methods which use both E and H fields

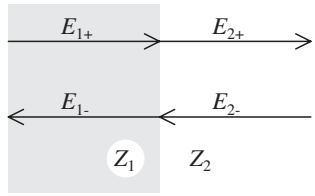


Figure 5.3 Waves at an interface.

are simpler in that case.

A single interface is depicted in figure 5.3. There are two waves propagating in both media, one to each direction. From the reflection and transmission coefficients (ρ, τ) the wave amplitudes on either side can be solved as a function of the amplitudes on the other side:

$$E_{2+} = \tau_1 E_{1+} + \rho_2 E_{2-} \tag{5.11}$$

$$E_{1-} = \tau_2 E_{2-} + \rho_1 E_{1+} \tag{5.12}$$

From these equations the amplitudes on one side of the interface can be solved as a function of the amplitudes on the other side of the interface:

$$\begin{pmatrix} E_{1+} \\ E_{1-} \end{pmatrix} = \begin{pmatrix} \frac{1}{\tau_1} & -\frac{\rho_2}{\tau_1} \\ \frac{\rho_1}{\tau_1} & \frac{\tau_1\tau_2 - \rho_1\rho_2}{\tau_1} \end{pmatrix} \begin{pmatrix} E_{2+} \\ E_{2-} \end{pmatrix} \tag{5.13}$$

By using the reflection coefficient equations ((5.6) and (5.7)) the matrix can be written as:

$$\begin{pmatrix} E_{1+} \\ E_{1-} \end{pmatrix} = \frac{1}{2Z_2} \begin{pmatrix} Z_1 + Z_2 & Z_2 - Z_1 \\ Z_2 - Z_1 & Z_1 + Z_2 \end{pmatrix} \begin{pmatrix} E_{2+} \\ E_{2-} \end{pmatrix} \tag{5.14}$$

These matrices can be chained, i.e. several consecutive interfaces can be expressed with one matrix. For a two-interface system there are six different waves in three domains with one wave to each direction in each domain (figure 5.4). Each of the two interfaces has its own characteristic matrix, which can naturally be chained as shown in equation (5.15).

$$\begin{pmatrix} E_{1+} \\ E_{1-} \end{pmatrix} = M_{12} \begin{pmatrix} E_{2+} \\ E_{2-} \end{pmatrix} = M_{12} M_{23} \begin{pmatrix} E_{3+} \\ E_{3-} \end{pmatrix} \tag{5.15}$$

where M_{12} and M_{23} are the matrices which describe interfaces from region 1 to 2 and 2 to 3, respectively.

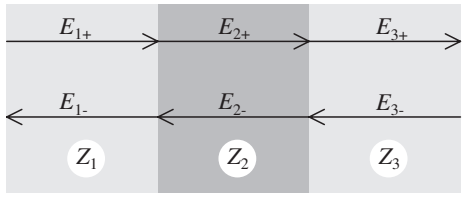


Figure 5.4 Reflections from two interfaces.

The equations shown above, however, do not take the phase change into account and are thus useless *per se*. Fortunately, the phase change can be expressed with a similar matrix. As the waves propagating to opposite directions do not interact with each other in linear medium, there are only two non-zero elements in the matrix:

$$\begin{pmatrix} E_{1+} \\ E_{1-} \end{pmatrix} = \begin{pmatrix} e^{ikx} & 0 \\ 0 & e^{-ikx} \end{pmatrix} \begin{pmatrix} E_{2+} \\ E_{2-} \end{pmatrix} \tag{5.16}$$

where x is the distance the light travels between two interfaces, and k the wave number of the radiation.

It should be noted that the signs in the exponential are different, as for the wave propagating to positive direction E_{2+} is at a later position (i.e. lags in phase) than E_{1+}, whereas for the negative direction E_{2-} is at an earlier position (leads in phase) than E_{1-}. For real wave numbers k the sign in the exponent can be chosen arbitrarily, as the real parts of e^{ix} and e^{-ix} are the same for real x. However, the wave numbers may be complex, and the sign of the imaginary part of the wave number is important. The convention of phase advancing towards the positive imaginary axis is maintained throughout this discussion.

The wave number k is defined as:

$$k = \frac{\omega}{v} \tag{5.17}$$

where ω is angular velocity of the radiation and v the phase velocity of the radiation. For electromagnetic radiation it is often useful to define the angular velocity using the wavelength *in vacuo*:

$$\omega = \frac{2\pi c/\lambda_0}{v} \tag{5.18}$$

where c is the velocity of electromagnetic radiation in vacuum and λ_0 wavelength of the radiation in vacuum.

The phase velocity v of electromagnetic radiation in a homogeneous medium can be calculated from the permittivity and the permeability of the material:

$$v = \frac{1}{\sqrt{\epsilon\mu}} \tag{5.19}$$

In the latter form the velocity is expressed by the relative permittivity and relative permeability and the velocity of electromagnetic radiation in vacuum. The ratio between the velocity in the intermediate medium and that in vacuum is called the refractive index, n:

$$n = \frac{c}{v} = \frac{\sqrt{\mu\epsilon}}{\sqrt{\mu_0\epsilon_0}} = \sqrt{\epsilon_r\mu_r} \tag{5.20}$$

where the material parameters have been expressed in terms of material parameters in vacuum (ϵ_0, μ_0) and relative material parameters (ϵ_r, μ_r).

Now, the wave numbers in the transfer matrix (5.16) can be replaced by more commonly used units of vacuum wavelength and refractive index:

$$k = \frac{2\pi n}{\lambda_0} \tag{5.21}$$

It should be noted that the refractive index may be complex, indicating losses in the material.

Similarly, the impedance of a medium may also be expressed by using the relative permittivity and permeability and the impedance of vacuum (Z_0):

$$Z = \sqrt{\frac{\mu}{\epsilon}} = \sqrt{\frac{\mu_r}{\epsilon_r}} Z_0 \tag{5.22}$$

Further simplifications are possible, as dielectric materials have $\mu_r \approx 1$ for practical purposes. This applies to all materials except for ferromagnetic materials. As no ferromagnetic materials are used in the indicator thin film systems, all relative permeabilities can be assumed to be unity. Anyway, at optical frequencies, magnetic effects are negligible.

This assumption simplifies the formulae somewhat. The impedance and the refractive index become (from (5.20) and (5.22)):

$$n \approx \sqrt{\epsilon_r} \tag{5.23}$$

$$Z \approx Z_0/\sqrt{\epsilon_r} = Z_0/n \tag{5.24}$$

The characteristic matrix of an interface (5.14) can then be expressed by using only refractive indices:

$$\begin{pmatrix} E_{1+} \\ E_{1-} \end{pmatrix} = \frac{1}{2n_1} \begin{pmatrix} n_2 + n_1 & n_1 - n_2 \\ n_1 - n_2 & n_2 + n_1 \end{pmatrix} \begin{pmatrix} E_{2+} \\ E_{2-} \end{pmatrix} \tag{5.25}$$

On a single wavelength a thin film stack can be expressed as a matrix with constant elements by multiplying the interface matrices (5.25) and transfer matrices (5.16) in the path to a single matrix (M). The actual reflection and refraction coefficients are calculated by considering the simple situation of radiation arriving from only one direction ($E_{1+} = E_i$). Some of the radiation is reflected as $E_{1-} = E_r$ and some transmitted as $E_{2+} = E_t$. These can be calculated from the matrix elements:

$$\begin{pmatrix} E_i \\ E_r \end{pmatrix} = M \begin{pmatrix} E_t \\ 0 \end{pmatrix} \tag{5.26}$$

where E_t and E_r can be solved:

$$E_t = \frac{1}{M_{11}} E_i \tag{5.27}$$

$$E_r = M_{21} E_t = \frac{M_{21}}{M_{11}} E_i \tag{5.28}$$

The overall reflection and transmission coefficients are then:

$$\rho = \frac{M_{21}}{M_{11}} \tag{5.29}$$

$$\tau = \frac{1}{M_{11}} \tag{5.30}$$

and

$$R = \left| \frac{M_{21}}{M_{11}} \right|^2 \tag{5.31}$$

$$T = \left| \frac{n_2}{n_1 M_{11}^2} \right|^2 \tag{5.32}$$

5.3 MIRROR STACKS

For a stack of m anisotropic dielectric films there are $2m+2$ parameters; thickness and refractive index for each film and refractive indices of the substrate and environment

around the film stack. Even in the case of real and wavelength independent refractive indices the wavelength response of the film stack may vary to a large extent.

A common way of producing dielectric mirrors is the use of quarter-wave stack. A quarter-wave stack has alternating high and low refractive index layers which all have optical thickness of $1/4\lambda$ (i.e. each layer introduces 90° phase lag to the propagating wave). This way all reflections from the interfaces are in phase with each other.

Naturally, as the number of layers increases, the reflection coefficient approaches unity because there are more and more reflecting surfaces. However, the mirror stack reflection is a function of the wavelength, and if the wavelength is one half of the design wavelength, the mirror stack reflection is zero as reflections from the interfaces interfere in a destructive manner.

Figure 5.5 shows the reflectance curves for different quarter-wave stacks. The substrate has $n_S = 1.52$, high-index layer $n_H = 2.0$, low-index layer $n_L = 1.45$ and environment $n_E = 1.33$. These values are chosen so that they represent the actual values in a BST/BS sol-gel thin film mirror deposited on crown glass and immersed in water. The label HLH refers to a stack with glass/high-index/low-index/high-index/water structure.

From the curves it is clear that for a given number of layers the $(HL)^nH$ structure is the most effective. An extra low-index layer on the outer surface will actually behave as an anti-reflection layer matching the impedances of the high-index layer and that of the environment.

There are well-known methods of improving the useful wavelength range of a dielectric mirror. For instance, the stack can be made so that the layer thickness changes gradually. This way the incident radiation will eventually encounter a mirror which reflects it. However, the number of layers required in these mirrors is rather large, and leads to practical difficulties with the porous sol-gel layers.

5.4 ABSORBING FILMS

Usually, dielectric mirrors and film stacks are made of purely dielectric films, i.e. films without significant absorption in the optical region. However, in the optical indicator application absorption is the quality to be measured.

In the Maxwell equations the absorption of a material results from the electrical conductivity of the material. This conductivity σ can be taken as the imaginary part of the dielectric constant ϵ:

$$\epsilon = \epsilon' - i\sigma/\omega \tag{5.33}$$

where ϵ' is the real part of the dielectric constant.

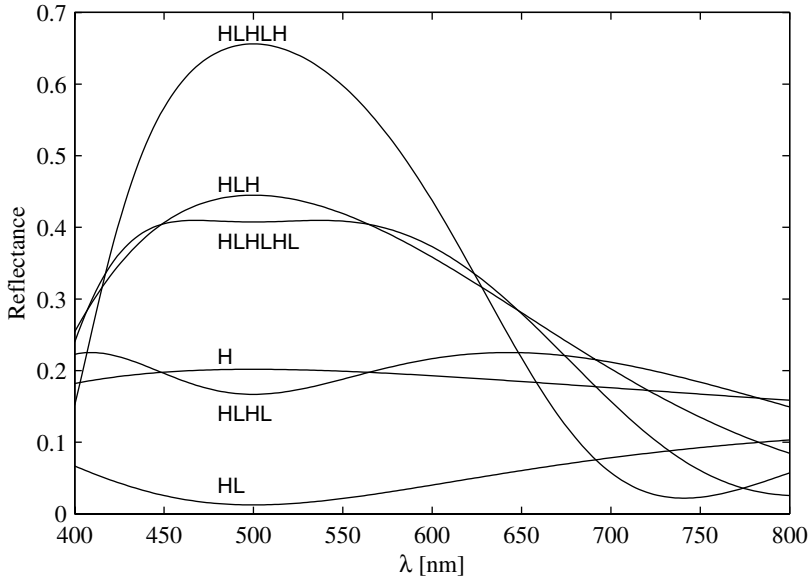


Figure 5.5 Reflections from different quarter-wave stacks ($n_H = 2.0$, $n_L = 1.45$, $n_S = 1.52$, $n_E = 1.33$)

This substitution does not invalidate the formulae presented before. When complex dielectric constants are used, also the refractive indices become complex. Complex refractive indices are often written as:

$$n = n_R - in_I \tag{5.34}$$

It should be noted that the imaginary part of the refractive index is negative, whereas n_I is always positive for all materials. Again, the sign convention is ambiguous, sometimes the sign of the imaginary part of the refractive index is chosen to be positive [49, p.113][50, p.613].

There are several actual absorption mechanisms. The conductance of a material may arise from the large number of unbound electrons, as in the case of metals, or it may be due to some molecular resonances. All these phenomena may be treated as damped resonances, where part of the resonant energy is lost.

In the case of metals the real part of ϵ is often taken as negligible compared to the imaginary part. With dyes having their absorption on the optical region, the imaginary

part of the dielectric constant is usually negligible compared to the real part. However, the thin film indicator dyes used in pH sensing films are very concentrated, and so some attention has to be paid to the region where n_I is not very small or very large.

The phase change in any harmonic resonator is large near the resonance frequency. In the optical domain this means that the real part of the dielectric constant changes near the resonant frequency because the resonance makes the propagation either faster or slower. Thus the conduction model presented above (5.33) is not an accurate physical model, and σ is always a complex function of wavelength.

It is important to note that the real and imaginary parts of the dielectric constant are not independent. The imaginary part increases towards the resonant frequency, and also the real part changes rapidly around the resonant frequency. It should also be noted that as the refractive index is the square root of the relative dielectric constant (5.23), its real and imaginary parts would be linked together even if there were no link between the imaginary and real parts of the dielectric constant.

Any real material can be thought of as a combination of large number of resonators. In a system like this the real and imaginary parts of the dielectric constant are linked by a dispersion relation (Hilbert transform), also known as the Kramers-Kronig relation in this specific context. Unfortunately, the Hilbert transform requires the knowledge of the dielectric constant behavior at all frequencies from 0 to ∞. This makes the practical use of the K-K relation rather difficult.

For qualitative analysis, the optical indicator material can be thought of having a finite number of resonance frequencies. For a dense material the dielectric constant depends on the location of the resonant frequencies (ω_k), their strength (ϱ_k) and on the damping factors (γ_k) associated with on these resonances [50, p.93]:

$$\frac{\epsilon - 1}{\epsilon + 2} = \sum_k \frac{\varrho_k}{\omega_k^2 - \omega^2 + i\gamma_k\omega} \tag{5.35}$$

The quotient on the left hand side of the equation is the molecular polarizability and it comes from the Lorentz-Lorenz (or Clausius-Mossotti) formula.

To obtain quantitative ideas of the refractive index changes in this system the number of resonance frequencies can be reduced to one. In the simplest case an indicator molecule can be modeled with only one resonance whose parameters change when the molecule changes its state.

If all other resonances of the material are assumed to remote in wavelength, the dielectric constant would be essentially constant over the optical wavelengths without the resonance. With this assumption, (5.35) can be written as:

$$\frac{\epsilon - 1}{\epsilon + 2} = a + \frac{\varrho}{\omega_0^2 - \omega^2 + i\gamma\omega} \tag{5.36}$$

where a is a constant and ω_0 the resonant frequency. If the dielectric constant is ϵ_B far away from ω_0, a can be written as:

$$a = \frac{\epsilon_B - 1}{\epsilon_B + 2} \tag{5.37}$$

This may also be written as function of wavelengths (λ, λ_0 correspond to ω, ω_0):

$$\frac{\epsilon - 1}{\epsilon + 2} = a + \frac{\varrho}{\left(\frac{2\pi c}{\lambda_0}\right)^2 - \left(\frac{2\pi c}{\lambda}\right)^2 + i\gamma\frac{2\pi c}{\lambda}} \tag{5.38}$$

or

$$\frac{\epsilon - 1}{\epsilon + 2} = a + \frac{b\lambda_r^2}{\lambda_r^2 - 1 + id\lambda_r} \tag{5.39}$$

where

$$b = \left(\frac{\lambda_0}{2\pi c}\right)^2 \varrho \tag{5.40}$$

$$d = \frac{\gamma\lambda_0}{2\pi c} \tag{5.41}$$

$$\lambda_r = \frac{\lambda}{\lambda_0} \tag{5.42}$$

ϵ can be now solved:

$$\epsilon = \frac{(2a + 2b + 1)\lambda_r^2 - 2a - 1 + id(2a + 1)\lambda_r}{(-a - b + 1)\lambda_r^2 + a - 1 + id(1 - a)\lambda_r} \tag{5.43}$$

Figure (5.6) depicts the real and imaginary parts of the dielectric constant when $\epsilon_B = 2.25$ (corresponds to $n = 1.5$) with different damping factors and resonance strengths.

It should be noted that the change of the real part of the dielectric constant crosses zero near the resonance peak. Large damping factors shift the resonance peak to the longer wavelengths (smaller resonant frequencies), as expected. Also, large damping factors make the resonance peak lower and broader. Broader peaks with similar imaginary part (extinction) give smaller changes to the real part of the refractive index than narrow peaks.

As a wave propagates in an absorbing layer, its amplitude as a function of position is:

$$E = E_0 e^{-ikx} = E_0 e^{-ixn\omega/c} = E_0 e^{-i2\pi n/\lambda} \tag{5.44}$$

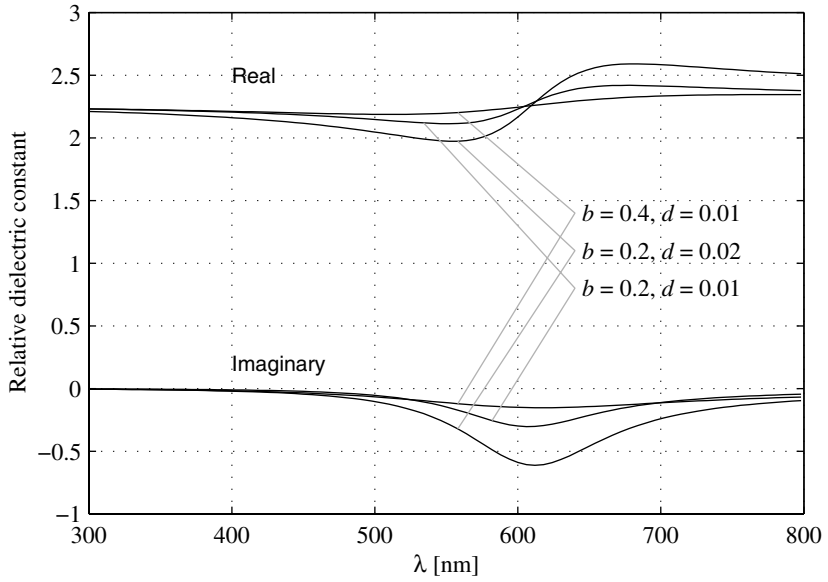


Figure 5.6 Real and imaginary parts of the relative dielectric constant for different resonance strengths (b) and damping factors (d) when the bulk $\epsilon_B = 2.25$.

Substituting n with its components (5.34) the imaginary refractive index can separated so that the transmission coefficient is:

$$E = E_0 e^{-i2\pi n_R x/\lambda} e^{-2\pi n_I x/\lambda} \tag{5.45}$$

The corresponding intensity is:

$$I = I_0 e^{-4\pi n_I x/\lambda} \tag{5.46}$$

Typical pH sensitive dyes have spectral peak widths of the order of a hundred nanometers. Their extinction coefficients can be rather high in thin film applications; a layer of a few hundred nanometers may absorb several tens of percents of the incident radiation.

If the wavelength $\lambda = 600$ nm, layer thickness $x = 200$ nm and transmission approximately 80 %, the imaginary part of the refractive index can be calculated by

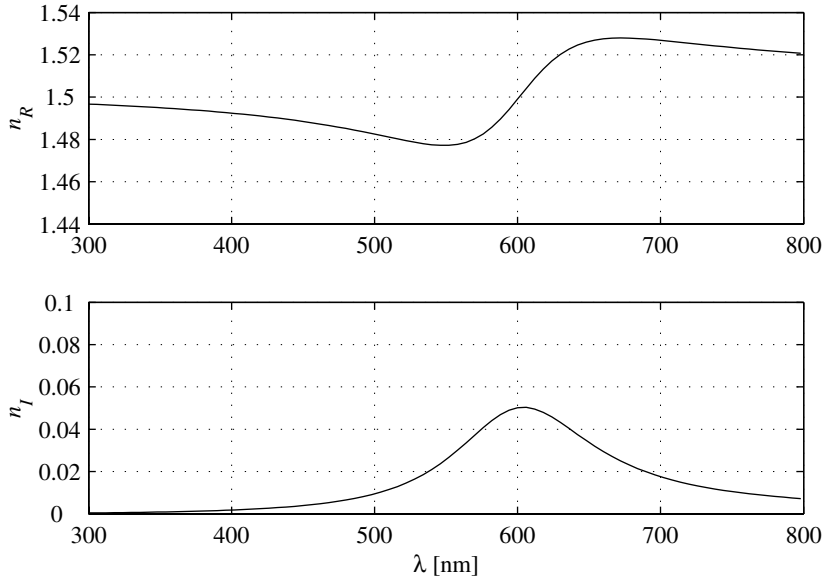


Figure 5.7 n_R and n_I for material having approximately 20 % absorption at 600 nm for a 200 nm thick dye layer.

solving (5.46):

$$n_I = -\frac{\lambda}{4\pi x} \ln \frac{I}{I_0} \tag{5.47}$$

This calculation gives $n_I \approx 0.05$ which cannot be taken as practically zero. To find the implications of this to the real refractive index, the model of single spectral peak absorption introduced above can be applied. The resulting refractive indices are shown in figure 5.7.

If this film is deposited on a substrate with $n = 1.5$, its reflection coefficient is non-zero but it remains under one thousandth (in intensity) and is thus insignificant in most applications. However, if a high-index film $n = 2$ is deposited on the dye film, the reflection coefficient between the two films changes notably as a function of wavelength. The two cases are depicted in figure 5.8.

In practice, the dye film is deposited between the glass substrate and a high-index film. So, these two cases are approximations of the two interfaces of the dye film. It can be seen that the glass / dye interface reflects very little and reflection changes at

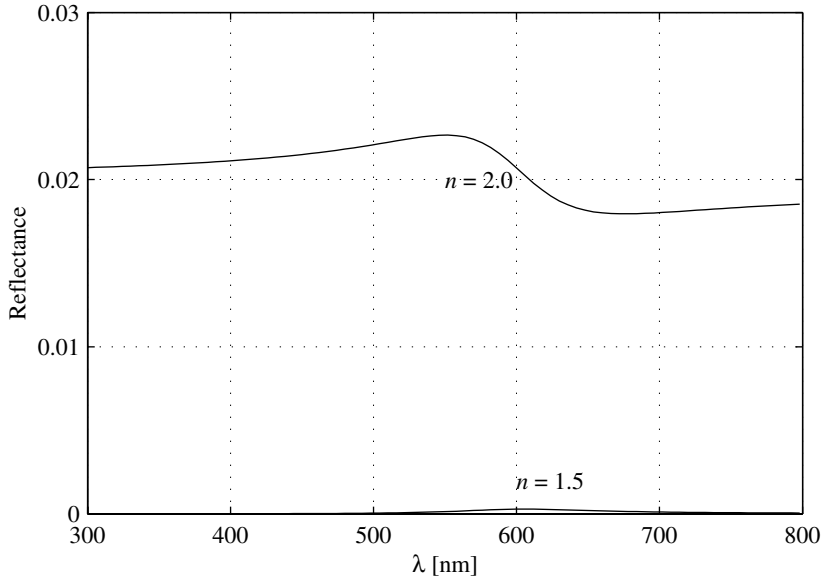


Figure 5.8 Reflection from the interface of an absorptive film (see 5.7) and a non-absorptive film ($n = 1.5, 2$).

that interface are small. However, the changes on the dye / ,high-index interface may be significant.

On the short wavelength side of a absorption peak the first interface will reflect more than if the dye effect were not taken into account. On the long wavelength side the reflection is smaller. As a net effect this tends to shift the perceived absorption peak to longer wavelengths.

There is a phase shift associated with the reflection from the absorbing / non-absorbing interface. This may introduce further changes to the reflection spectrum, and the effect depends on the mirror stack design.

The commonly used method of treating absorbing films is to treat the absorption and refractive index separately. The examples above illustrate that this cannot necessarily be done in the case of strongly dyed indicator films. On the other hand, the changes do not seem to be very large, and the simple approach is used in the mirror design below. This approximation does not change the validity of the measurement interpretation methods introduced later, as the change in reflection is deterministic;

only the perceived spectrum of the dye changes. In particular, the effect on the final result is predictable and can be corrected.

5.5 EFFECT OF CHANGING EXTERNAL REFRACTIVE INDEX

One of the unusual challenges in this dielectric mirror design is the changing external refractive index. The indicator mirror stack may be used to measure different liquids which may have different refractive indices. The range of refractive indices in the water-based liquids (pH loses its practical significance in non-aqueous solutions) is usually between 1.33 and 1.5, the latter corresponding to saturated sucrose solution near boiling point of water.

For a mirror structure having 100 % reflection the external refractive index is not significant as no light reaches the interface of the outmost film layer and the medium outside. This would suggest that a mirror structure with as high reflection as possible would be most tolerant to external refractive index changes.

Unfortunately, this is not a realistic goal as the number of layers should be limited to well under ten layers to ensure sufficient ion permeability and reasonable diffusion delays. This way the reflectance of the mirror will be a function of the external refractive index. Figure 5.9 depicts the reflectance of one, three, and five-layer mirrors ($n_S = 1.52$, $n_L = 1.45$, $n_H = 2.0$, $n_E = 1.33$).

As the measurement itself is a color measurement, the absolute reflectance is not very important. However, it is important that the mirror does not change its reflection spectrum as the external refractive index changes. This effect can be observed by normalizing the mirror reflections to the maximum reflection and varying the external refractive index. Figure 5.10 depicts the results for the three different mirror structures.

From these curves it seems that the three layer mirror is significantly better than the simple one layer mirror. The five layer mirror exhibits still higher absolute reflectances, but its spectral response is narrower, and it is more difficult to manufacture. Increasing the number of layers does improve the absolute reflection and makes its variations smaller but does not necessarily improve the relative reflection changes.

Due to these results the three-layer mirror has been chosen to be used in the pH sensor design.

5.6 EFFECT OF CHANGING FILM REFRACTIVE INDEX

As the thin film material is porous, it changes its refractive index depending on the medium it is immersed into. As discussed in section 3.3 the dielectric constant of the

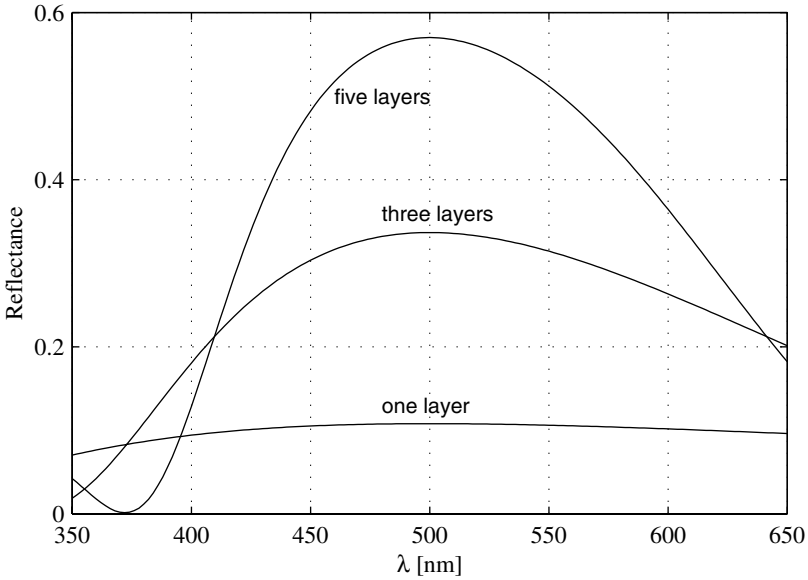


Figure 5.9 Reflectance of one, three, and five layer mirrors ($n_S = 1.52$, $n_L = 1.45$, $n_H = 2.0$, $n_E = 1.33$).

thin film layer can be taken as the weighted average of the polarizability of the two materials (3.8).

This can also be expressed by refractive indices:

$$\frac{n_p^2 - 1}{n_p^2 + 2} = (1 - p)\frac{n_b^2 - 1}{n_b^2 + 2} + p\frac{n_E^2 - 1}{n_E^2 + 2} \tag{5.48}$$

Or, solving n_p:

$$n_p = \sqrt{\frac{2(c_E - c_b)p + 2c_b + 1}{(c_b - c_E)p - c_b + 1}} \tag{5.49}$$

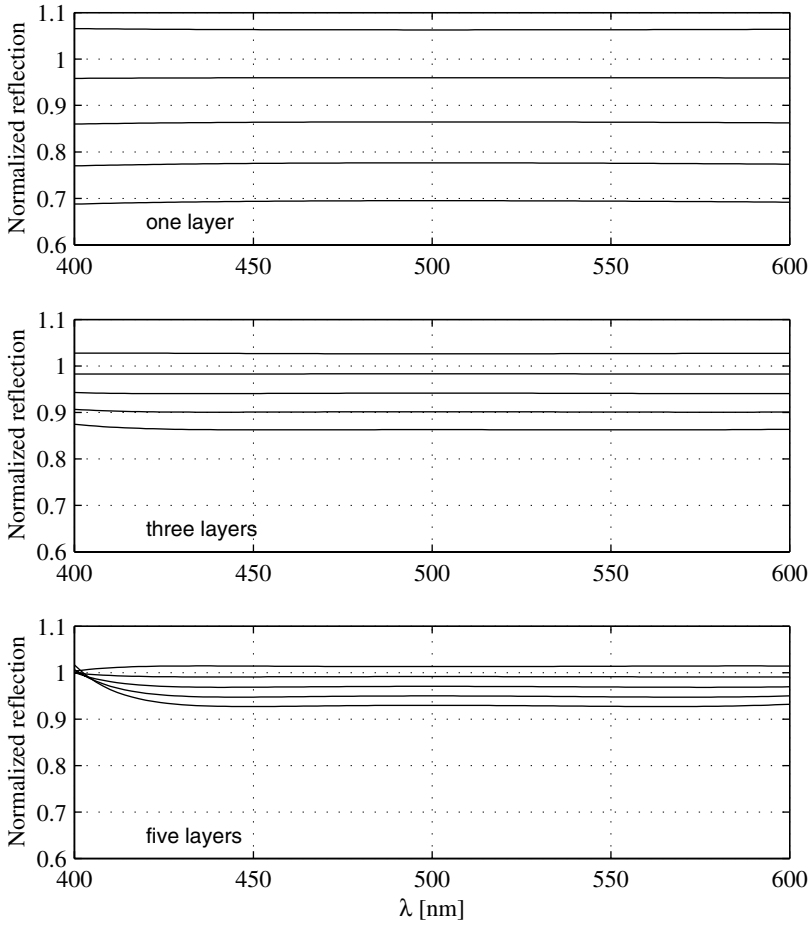


Figure 5.10 Normalized reflectance of one, three, and five layer mirrors as a function of n_E ($n_S = 1.52$, $n_L = 1.45$, $n_H = 2.0$). Lowest curve in each group represents $n_E = 1.5$, highest $n_E = 1.3$.

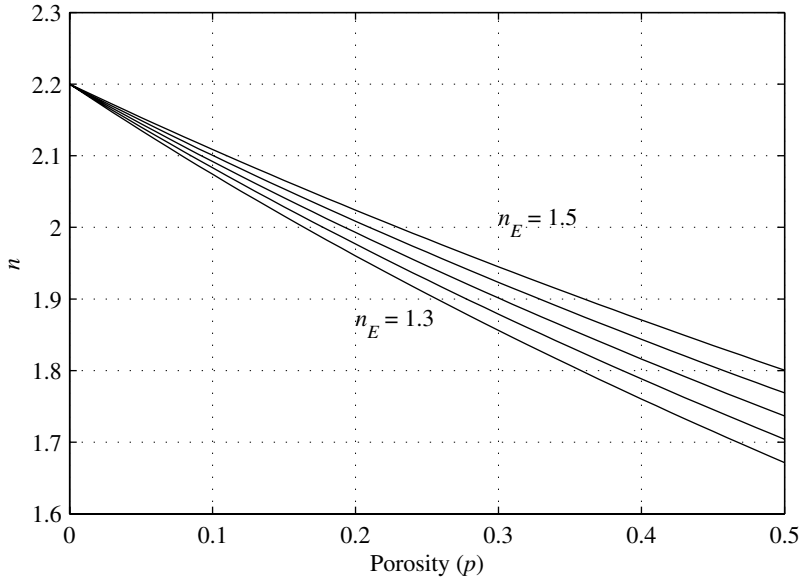


Figure 5.11 Refractive index of a porous titanate film (bulk material refractive index $n_b = 2.2$) with different pore-filling liquids ($n_E = 1.3, 1.35, 1.4, 1.45, 1.5$) as a function of porosity.

where

$$c_E = \frac{n_E^2 - 1}{n_E^2 + 2} \tag{5.50}$$

$$c_b = \frac{n_b^2 - 1}{n_b^2 + 2} \tag{5.51}$$

where n_p is the refractive index of the porous film, n_E that of the pore-filling liquid, and n_b that of the bulk material.

The change of the film refractive index is then a function of porosity (p). With $n_b = 2.2$ (corresponds to titanate films) the different refractive indices as function of porosity and pore-filling liquid refractive index are depicted in figure 5.11.

Obviously, low-porosity films are more tolerant to this change of reflection. Unfortunately, it is not practical to make films with zero porosities, as then the ions to be measured would not be able to go through the films. However, there are several other

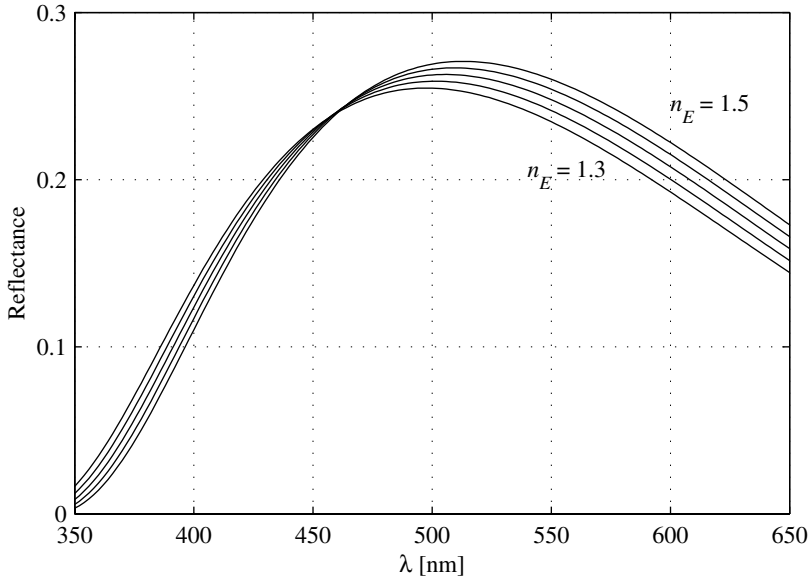


Figure 5.12 Reflectance of a three-layer porous mirror ($p = 0.2$, $n_b = 1.52/2.2$) as a function of the pore-filling liquid refractive index ($n_E = 1.3, 1.35, 1.4, 1.45, 1.5$).

advantages associated with low-porosity films (durability and high refractive index) and so it seems reasonable to aim at dense films.

The effect of the pore-filling liquid refractive index on the mirror reflection spectrum is depicted in figure 5.12. The mirror porosity is assumed to be $p = 0.2$, otherwise the mirror is similar to three-layer mirror calculated in section 5.5.

The figure shows that the change in the refractive index does not change the maximum reflection significantly. The more significant change is in the wavelength properties of the mirror. As the pores are filled, their refractive index increases and thus the optical thickness of the film increases. The low-index layers suffer from this less than the high-index layers as the refractive index of the former is already near to that of water.

It should be further noted that the liquid filling the pores does not necessarily have the same refractive index as the liquid outside. The porous structure has very small pores, and high-index aqueous solutions have a high concentration of dissolved molecules or ions. For the highest refractive index liquids the dissolved material is

usually rather large organic molecules (such as sucrose), which cannot enter the pores of the structure.

Thus, it can be reasonably assumed that the refractive index range of the liquid entering the pores is quite narrow and close that of the water. It seems that the change in the external refractive index is more significant than the change in the film refractive index.

5.7 FILM TOLERANCES

The thin film layers cannot be manufactured with zero tolerances. Even though the sol-gel process can be controlled to rather a great extent, it does not offer similar real time deposition control possibilities, e.g. real-time ellipsometry, as the continuous vapor deposition methods.

The tolerances in the refractive index are rather tight, as small relative variations in the porosity produce even smaller relative variations in the refractive index. The thickness tolerances are more significant, because even small viscosity changes may change the film thickness significantly (by equation 3.11). The viscosity change may be due to ageing of the sol or due to small variations in ambient temperature or humidity.

The film properties of the two similar high-index films can be assumed to be similar. While the drying process may make the lower high-index film denser (it is dried three times, whereas the upper film is dried only once), the bulk properties of the films should be near to each other as they are dipped within a small period of time.

It seems reasonable to use a model with four parameters, thicknesses and refractive indices of the two different layer types. While it would be possible to write the expression for the transmission of a three-layer structure in analytical form by using matrix methods described in section 5.2, the function would be too complicated to be useful in practice. Instead, the effect of manufacturing tolerances can be observed by looking at the worst case data by using the lower and upper bounds of the parameters.

In the following example the mirror central wavelength is 500 nm, high refractive index material bulk (non-porous) refractive index is 2.2, low-index bulk refractive index 1.52, and low and high-index porosities are both 20 %. These values correspond to high-index parameters $n_H = 1.97$, $d_H = 63$ nm and low-index layer parameters $n_L = 1.46$, $d_L = 86$ nm.

Figure 5.13 shows the reflectance when the layer thicknesses are varied by ±5 %. The family of curves is drawn so that for each curve the low-index and high-index thicknesses are either 95 % of nominal, nominal or 105 % of nominal, thus giving nine

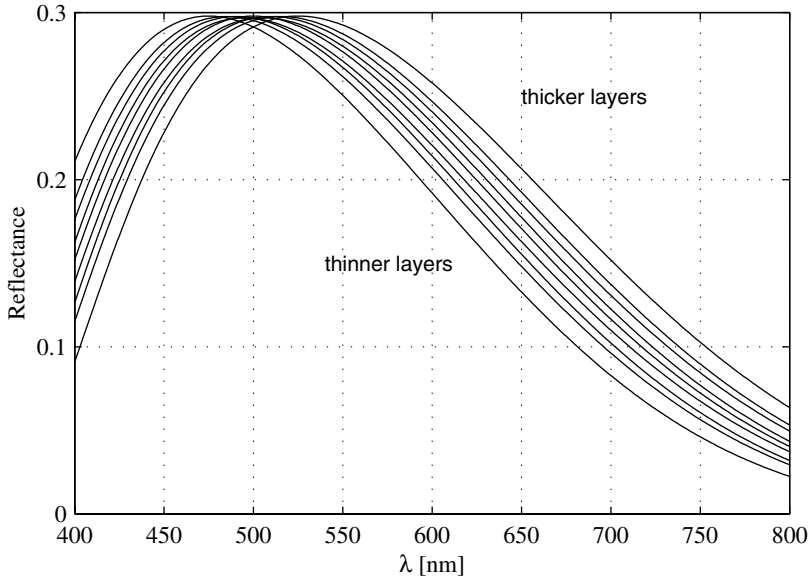


Figure 5.13 The effect of 5 % variations in the mirror layer thicknesses to the reflection spectrum.

curves. The thickness variations change the mirror reflectance maximum position but do not change the maximum reflectance or the curve shape significantly.

Figure 5.14 shows a similar family of curves when the porosity of high and low-index layers is changed ±5 % (i.e. porosity is from 0.19 to 0.21). The nine curves are in three groups of three curves. This behavior is explained by the fact that porosity changes in the high-index film produce more pronounced refractive index changes than those in the low-index film.

The reflection maximum position does not move significantly but the maximum value does change considerably as the refractive indices change. The reflection maximum position shifts slightly as the optical thicknesses (i.e. the product of the physical thickness and refractive index) change, but this change is rather small for small changes in porosity.

So, even though the thicknesses and refractive indices are not independent, the porosity tolerances are visible mainly in the reflection maximum value and thickness tolerances are visible in the reflection maximum position. It does not seem to be

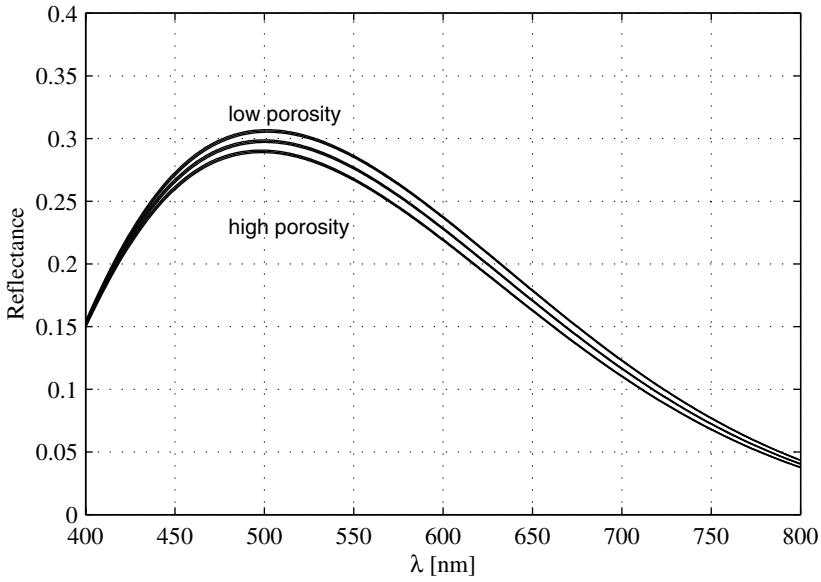


Figure 5.14 The effect of 5 % variations in porosity to the reflection spectrum.

practically possible to deduce the individual layer thickness errors from the reflection spectrum, as different errors may give the same result. On the other hand, this result makes the mirror design rather forgiving; possible reflection maximum position errors may be corrected by changing only one layer thickness value without sacrificing the reflectance.

One possible thickness error is that of differing high-index layer thicknesses. Figure 5.15 depicts the situation where the porosities and the low-index layer thickness are accurate but the high-index layer thicknesses may vary ±5 % independently. Again, the reflection maximum tends to move in wavelength but not in magnitude.

An important error source is layer densification during the manufacturing process. In a three-layer mirror the bottom layer is dried in an oven three times, whereas the top layer is dried only once. This may make the bottom layer denser than the top layer. It should be noted that there is the same amount of material in both layers, and the shrinkage is one-dimensional as the surface area of the film cannot be changed. Thus

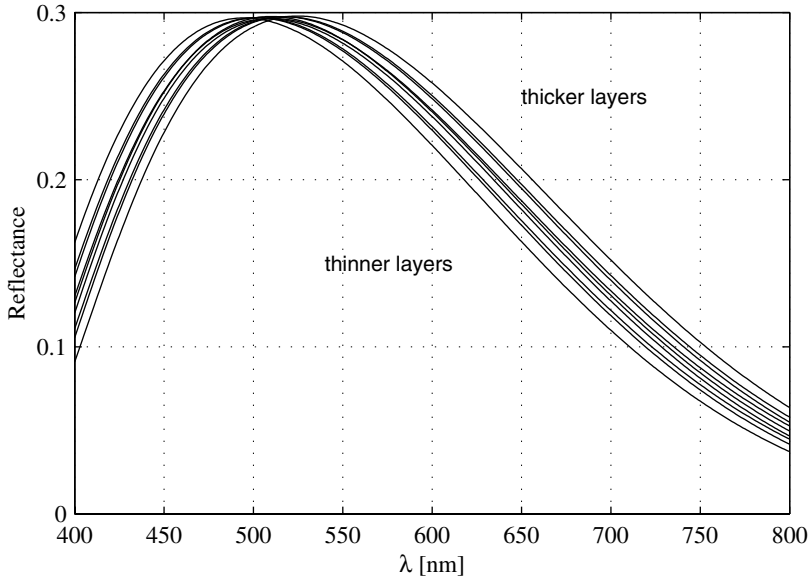


Figure 5.15 The effect of 5 % variations between the two high-index layer thicknesses.

the thickness (d) of a layer is linked to the porosity p by:

$$d = \frac{d_0}{1 - p} \tag{5.52}$$

where d_0 is the thickness of a non-porous film with same amount of material.

The most important parameter is the optical thickness of the layer:

$$d_{opt} = n_p d \tag{5.53}$$

where n_p is the refractive index of the porous film.

If the optical thickness of a layer differs significantly from its designed value, the mirror maximum reflection wavelength will be shifted. The optical thickness as a function of porosity can be calculated from (5.53) and (5.49).

Figure 5.16 depicts the situation where the optical thickness of a high-index layer ($n_b = 2.2$) is designed to be 1.00 when $p = 0.2$. Evidently, the densification of the layer will also decrease its optical thickness even though the amount of material in

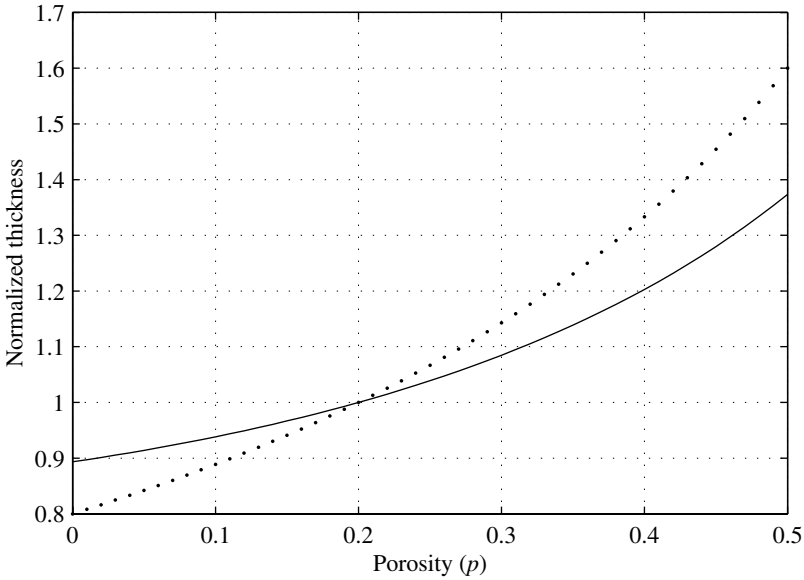


Figure 5.16 Optical thickness of a layer as a function of porosity resulting from densification of the layer ($n_E = 1.33$, $n_b = 2.2$). Dotted line represents the normalized physical thickness of the layer.

the layer remains the same. However, the change in refractive index does balance this change significantly.

6 Film dipping measurements

The final thickness of a dipped sol-gel thin film depends on the sol properties, withdrawal velocity (see section 3.4.2), and drying conditions. Sol properties have to be controlled by the synthesis and ageing processes, and drying is controlled by the drying environment. The remaining third parameter is the withdrawal velocity, which is the simplest of these parameters to control and thus lends itself well to the final thickness adjustment.

On the other hand, if the withdrawal velocity is not stable, the film thickness will vary over the substrate and horizontal stripes will be visible on the substrate.

Thus, it is essential to be able to measure the dipper movement for two reasons. In order to adjust the layer thickness accurately the velocity has to be known accurately. This knowledge is also required to identifying the velocity noise sources in the dipping equipment.

6.1 WITHDRAWAL VELOCITY

The dipping equipment used in making the films is driven by a geared DC motor which pulls the substrate carriage with a string (figure 6.1). The velocity control is realized by adjusting the motor voltage. This control gives a reasonably linear control over the velocity as the torque seen by the motor remains almost constant due to high gear ratio.

The velocity range used in sol-gel dipping applications is usually from 10 to 300 mm/min (or, to be more consistent with the SI system, 0.2 to 5 mm/s). Higher and lower velocities are also possible, but with the sols used in this work the useful dipping rates have been well within the range mentioned above.

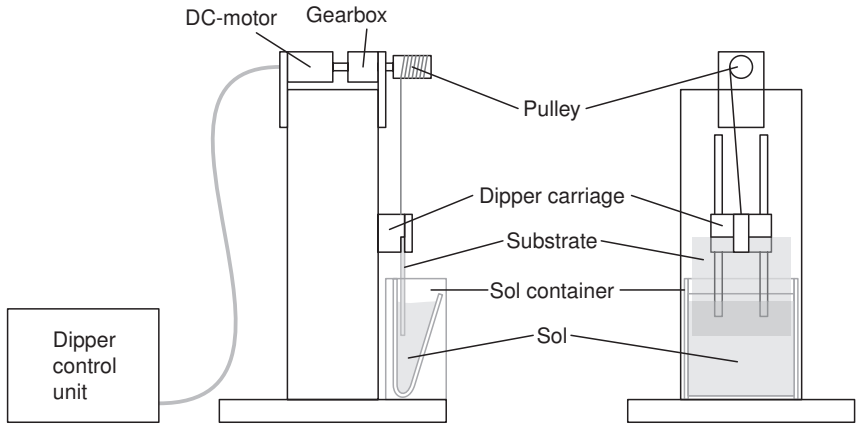


Figure 6.1 The dipper.

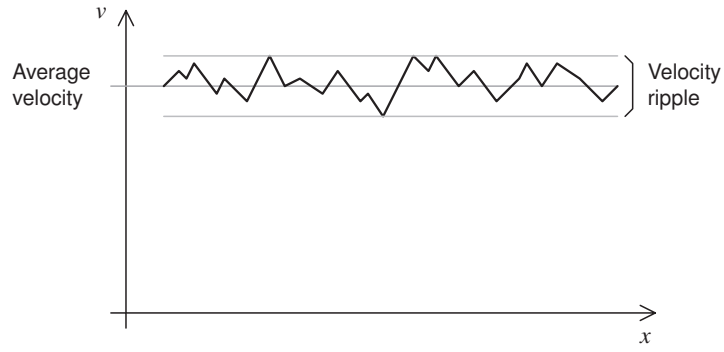


Figure 6.2 Ripple in dipping velocity.

From the practical point of view, the withdrawal velocity can be split into a constant component and ripple component (figure 6.2). As the ripple should be small compared to the average velocity, the non-linear relationship between withdrawal velocity and film thickness can be thought to be linear over the region where the ripple component fluctuates. This way film average thickness is dictated by the average velocity, whereas ripple accounts for film non-uniformity.

Naturally, the ripple component should be as small as possible. In general, high frequency noise in the velocity deteriorates the film thickness uniformity less than low frequency noise, as high frequency noise is averaged by the fluid. However, some noise frequencies may resonate with the dipping liquid container, and the waves thus formed may produce large lower frequency components at the liquid interface, and thus all noise components may deteriorate the film quality significantly.

There are several velocity noise sources in the dipping system:

- The motor does not produce even rotation; commutation in a DC motor produces small regular fluctuations in the rotation.
- The gearbox produces periodic noise with multiple frequency components corresponding to single teeth frequencies, single gear rotation frequencies and to the full period of the gearbox (dictated by the least common denominator of the numbers of teeth per gear).
- Irregularities of the pulley and especially radial errors in its mounting produce very low-frequency periodic (sinusoidal) fluctuations.
- Friction fluctuations of the substrate carriage produce noise which is partially position-dependent (repeatable) and partially random.

Of these noise sources the most important one seems to be friction of the carriage, at least at low velocities. The motor commutation noise is insignificant, as the motor rotation speed is high and the rotating mass has high inertia compared to the torque fluctuations. Pulley radial tolerances are in the order of micrometers or dozens of micrometers and as the noise thus produced has very low frequency, it represents an insignificantly small portion of the total noise.

As velocity is a differential of the position, very small position errors may produce significant velocity errors. If the velocity has a sinusoidal error component, the velocity and position can be written as:

$$v = v_0 + v_r \cos(2\pi f t) \tag{6.1}$$

$$x = x_0 + v_0 t + \frac{v_0}{2\pi f} \sin(2\pi f t) \tag{6.2}$$

where v_0 is the constant component of the velocity, v_r the ripple component, and f the ripple frequency. The position error is then:

$$x_{err} = x - (x_0 + v_0 t) = \frac{v_r}{2\pi f} \sin(2\pi f t) \tag{6.3}$$

If the desired constant dipping velocity is 5 mm/s, and the ripple is 10 % of this at 100 Hz, then the maximum position error is by (6.3) approximately 800 nm. This is rather a small figure even though the velocity error is not negligible.

The average velocity (v_0) is rather straightforward to measure over a longer period of time with four digit repeatability by using, for instance, a pair of optical gates driving a stopwatch. On the other hand, in order to detect even largish ripple, a quite fast and accurate position measurement is required.

It should be noted that the frictional forces tend to be rather independent of the dipper carriage velocity. So, the acceleration errors due to the frictional forces, and thus the absolute velocity errors, remain independent of the dipping rate. This way the slowest dipping rates suffer more severely from the noise in the dipper velocity than faster dipping rates.

Withdrawal velocity errors can be compensated for by position feedback. However, it is difficult to produce a mechanical system with wide feedback bandwidths, especially the gearbox will lower the feedback bandwidth to a few hertz or dozens of hertz. The dipper used in this work functions in an open-loop configuration because most error sources have too high a frequency to be compensated and also because the open-loop system has sufficiently stable voltage-to-frequency properties in practice.

6.2 DIPPER POSITION MEASUREMENT

As noted in the previous section, the dipper position measurement method has to be both fast and accurate. More specifically, the absolute accuracy does not need to be very high, for example 1:1000 is certainly good enough, but the differential resolution has to be high.

The most commonly cited figure of merit of position sensors is the product of repeatability and read-out rate. Slow sensors are usually more accurate than fast sensors. Other important parameters are the cost of the sensor and the mechanical tolerance requirements associated with mounting of the sensor parts.

Numerous sensor principles have been introduced for position and distance sensing. Most industrially used sensors have two parts, a track which is mounted to the stationary part of the system, and the sensor which is mounted to the moving part. Naturally, the parts can be reversed if that produces a more desirable geometry.

The sensor principle can be optical, capacitive, or inductive. Optical sensors have either a grating or a coarser intensity-modulated scale on the track. By measuring the intensity of the returning light in different positions, the position can be calculated. Optical sensors can be very accurate but they are not tolerant to contamination, and their measurement geometry may require very precise positioning of the sensor parts.

Inductive sensors often have a magnetized track with small adjacent regions having different permanent magnetization. The magnetic field of the track can then be measured with, e.g., Hall-sensor. Inductive sensors can also rely on other measurement principles, such as the measurement of eddy currents produced into the track. Inductive sensors are not sensitive to contamination but their accuracy is not as good as that of the best optical sensors.

In the capacitive configuration the track and the sensor form two plates of a capacitor. Usually, the plates are connected so that the sensor part has at least three capacitors and the track is in a zig-zag form. By driving two capacitor plates in the sensor with different phase sinusoids and measuring the phase from the pick-up plate the differential capacitance between the plates can be measured without any active electronics connected to the track. As with inductive sensors, there are other measurement principles available. Capacitive sensors tend to be somewhat slower than inductive sensors as the measurement has to made from the phase information of a weak AC signal.

There are some very accurate capacitive small-distance sensors. For example, nanometer variations in capacitor plate distance can be measured if the plate distance is in the micrometer range. However, these systems have very short operating distances.

If the resolution requirement is not high, then simple potentiometers or variable inductors may be used. These systems are fast and inexpensive but also noisy and have poor linearities in small movements.

The most accurate position measurement systems rely on interferometry. Interferometry has the advantage of being very fast, nanometer resolutions are obtainable with megahertz measurement bandwidth with standard technology. However, there are some disadvantages in using interferometry. Probably the most significant is that the coherence length of the light source (laser) has to be longer than the maximum movement of the object to be measured. For millimeter distances this can be obtained with inexpensive semiconductor lasers, but if hundreds of millimeters need to be measured, gas lasers or special diode lasers have to be used, which may not be economically viable.

The absolute accuracy of all measurement methods is lower than the differential accuracy over shorter distances. All measurement methods which use tracks suffer from possible non-linearities in the track manufacturing process and from the thermal expansion of the tracks, and interferometry suffers from changing refractive index of air.

The choice of the position measurement method is thus a compromise between practical realizability and performance.

6.2.1 Position sensor principle

The position sensor chosen to be used in dipper position measurements is a commercial optical sensor [51] which uses a stationary grating and a moving sensor which gives sinusoidal output voltage signals as a function of position. The sensor is an incremental position sensor, and the output signal undergoes a full cycle every 10 μm of movement. [52]

The sensor was chosen due to its reasonable price, fast operation, and high resolution. As the sensor produces both sine and cosine signals, the resolution can be easily quadrupled by recording the signs of the signals, and further doubled by comparing the absolute magnitudes of the signals. More advanced signal processing may increase the resolution by orders of magnitude, as shown below.

The operating principle of the sensor is shown in figure 6.3. A grating is illuminated with light coming in normal angle to the grating. The grating is designed so that it produces strong first order diffraction maxima and a small zeroth order. The first order waves advancing in opposite directions interfere with each other and form a horizontally (parallel to the grating) sinusoidal and vertically (normal to the grating surface) constant interference patterns. This pattern can then be measured at different positions to give sine and cosine signals[1].

In practice, the grating does not eliminate all zeroth order radiation. This small residual zeroth order does introduce a periodic error to the signal. The period of this signal is the same as the grating period. As the period of the desired signal is half of the grating period, the error is at half frequency.

Closer details of this and other error sources are given in appendix B.

6.2.2 Measurement interpretation

The position sensor produces four voltage signals which represent four quadratures of a sinusoid. These quadrature signals can be used to determine the position within a full cycle. In order to determine the position over a longer movement, the number of cycles has to be counted.

Due to the nature of the semiconductor detector and amplifiers used in the sensor,

[1] Naturally, the measured signals can be any combination of the quadrature signals. For instance, a measurement with three signals at 120° phase shift is possible as suggested in [53].

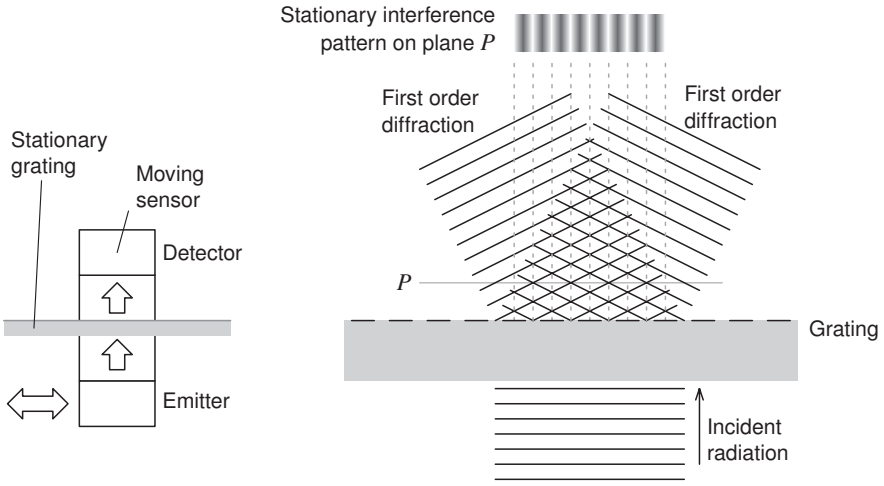


Figure 6.3 Moving sensor on a grating (left) and schematic of the interference pattern arising from the first order diffractions of the grating (right).

these signals have some bias voltage:

$$C_+ = V_0 + V_g \cos\phi \tag{6.4}$$
$$S_+ = V_0 + V_g \sin\phi \tag{6.5}$$
$$C_- = V_0 - V_g \cos\phi \tag{6.6}$$
$$S_- = V_0 - V_g \sin\phi \tag{6.7}$$

To have non-biased quadrature signals, these signals can be subtracted pairwise from each other:

$$C = C_+ - C_- = 2V_g \cos\phi \tag{6.8}$$
$$S = S_+ - S_- = 2V_g \sin\phi \tag{6.9}$$

With an ideal sensor the signal from the sensors should be a perfect circle (figure 6.4). The angular information can then be calculated from the quotient of the two difference signals:

$$\frac{S}{C} = \tan\phi \tag{6.10}$$

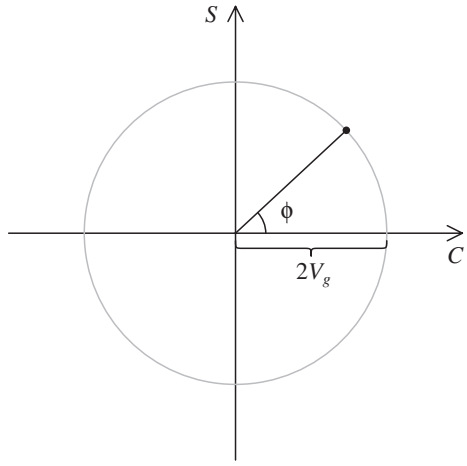


Figure 6.4 In an ideal case the voltage values draw a circle.

The measurement system developed to sample the voltages is simple (figure 6.5). It incorporates four different 12-bit successive approximation AD converters which share the same clock and reference signals. This is important as a non-synchronous measurement would invalidate the formulae presented above. There are no preamplifiers as the signal levels are sufficiently high and impedances sufficiently low for the signal to be digitized directly. This arrangement minimizes the number of error sources in the system.

6.2.3 Measurement noise

There are several error sources in the measurement. Measurement system nonlinearities and zeroth order interference will introduce a periodic error into the measurement. There will also be some noise from the photo detection and amplification. The analog-to-digital (AD) conversion introduces some quantization noise, and finite calculation resolution may produce some numerical noise

Ideally, the detector voltage noise is the limiting factor. The AD conversion resolution should be such that the voltage noise from the detector is significantly larger than the quantization noise of the converter. There is an additional advantage in this; differential non-linearities of the AD converter are less significant if the signal noise covers several digital output codes.

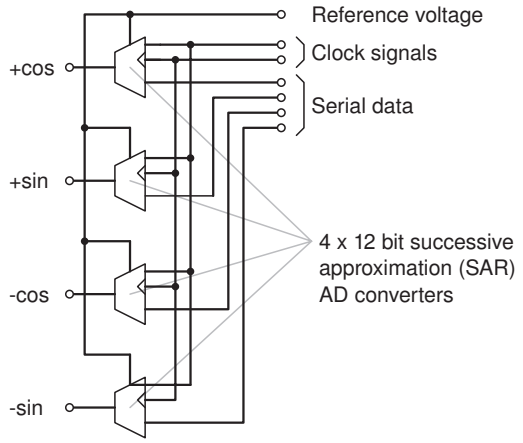


Figure 6.5 Measurement system for the quadrature signals.

The digital noise emerging from the calculations can be made very small, so the dominating noise source in the system is the voltage noise from the sensor.

If the total random noise of one voltage signal is e_n, then the voltage noise for the difference of two signals (6.8) is $\sqrt{2}e_n$, assuming the noise is uncorrelated. The two difference signals (cosine and sine) thus form a spot on the xy-plane with this noise radius (figure 6.6). The noise (ϕ_n) in the calculated angle (ϕ) is then:

$$\phi_n = \frac{e_n \sqrt{2}}{V} \tag{6.11}$$

where V is the amplitude of the difference signals S and C.

While this result is accurate only for an ideal circle and small noise, the real data is approximately circular and noise small enough for this calculation to be valid.

In the real system it has turned out that the sensor is relatively noiseless, and the AD converter output tends to keep stationary, i.e. the quantization noise becomes the dominating noise source. This is not desirable, and it seems probable that the system performance could be further improved by increasing the sampling resolution.

The quantization noise for a random signal is:

$$e_n = \sqrt{\frac{1}{12}} q \tag{6.12}$$

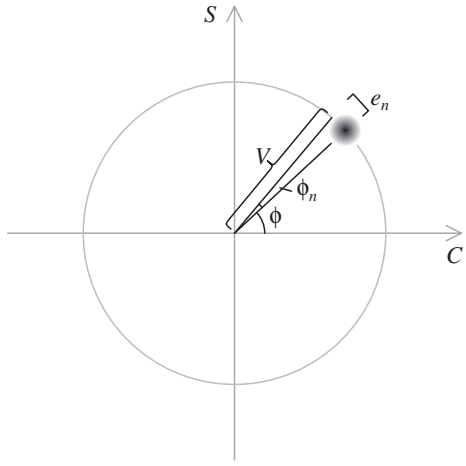


Figure 6.6 Angular error due to voltage noise.

where q is the quantization step.

The voltage signals measured from the system are approximately 500 mV peak-to-peak (i.e. 250 mV radius for the circle) and the converter quantization step is approximately 1 mV. By using (6.11) and (6.12) the approximate angular noise is slightly over 1 mrad. In terms of position this gives approximately 2 nm, as one full cycle is 10 μm for the 20 μm grating used in the system. This can be thought as the ultimate limit for the system performance.

The sampling system frequency is 10 kHz. At this frequency the 2 nm position error will give 20 μm/s velocity error in the measurement for single measurement points, if the velocity is calculated as a difference between consecutive position samples.

The actual performance of the system does not seem to be on this level. Figure (6.7) shows some real position and velocity data from the dipper velocity measurement.

While the different noise sources are difficult to distinguish from the position data, the noise seems to be much larger than that calculated above. All of this noise cannot be attributed to the real movement noise, as there are significant high-frequency components which are quite unlikely to be present in the mechanical movement.

A closer look at the voltage data (figure 6.8) collected from the sensors reveals at least one significant error source. The double-period component arising from the ze-

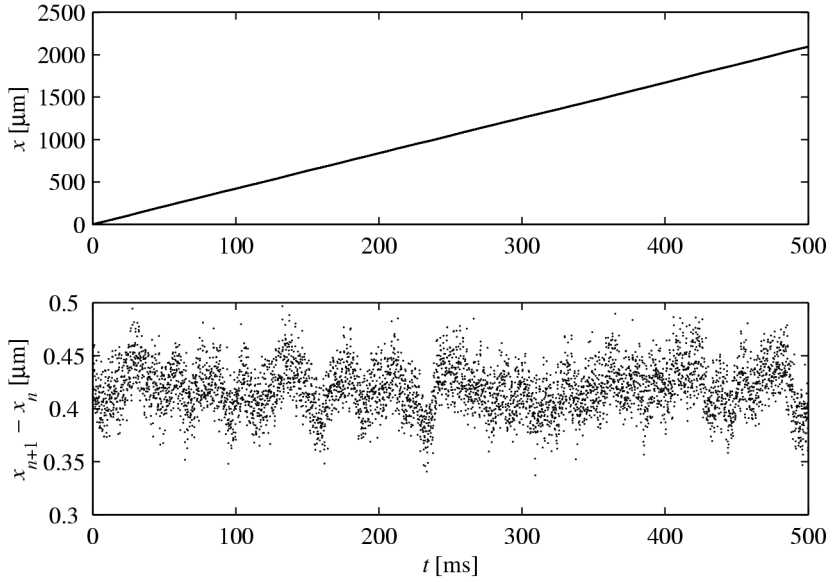


Figure 6.7 Measured position data (upper) and velocity data calculated by taking difference of consecutive samples (lower).

roth order is clearly visible, and thus the measurement has significant periodic errors.

6.2.4 Periodic error correction

In the previous section the position was directly calculated from the angle of the two voltage signals (6.10). In the ideal case this would give accurate results, but in practice, there are several error sources in the system:

- the photodiode biases may differ from channel to channel
- the photodiode gains may differ from channel to channel
- the AD converters may have gain and bias errors
- the system may be non-linear at some point (photodiodes, amplifiers, quantization)

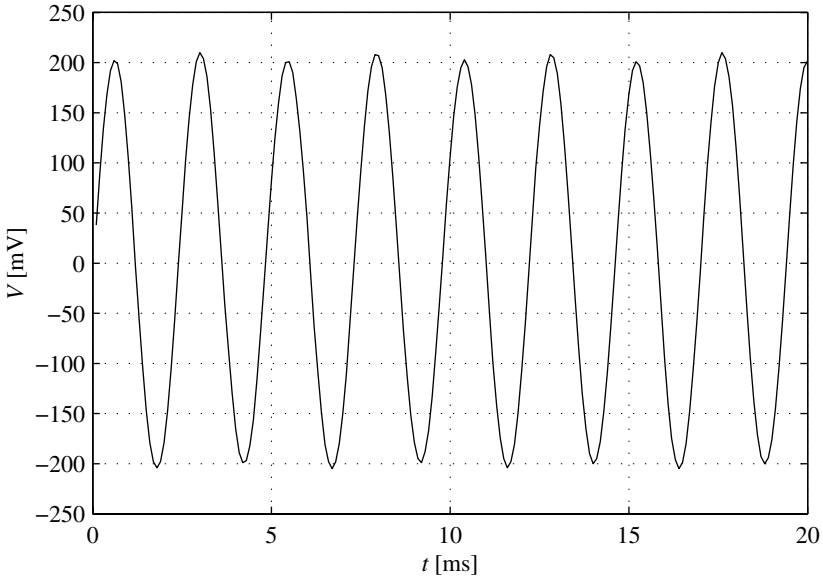


Figure 6.8 Measured voltage signal from the position sensor.

- the zeroth order interference introduces a double-period error

All these deterministic errors together make geometry errors to the circle drawn by the measurement results. Bias errors move the circle, gain errors change its aspect ratio (i.e. make it elliptic), and non-linearities distort its shape. The most interesting error is the zeroth order error whose period is double that of the measurement signal. This makes the data draw a double-loop as shown in figure (6.9).

However, as long as the errors are such that increasing position will give increasing angle values, there is a one-to-one mapping between the measured, i.e. the position calculated from (6.10), and real position. It has to be noted that as the calculated angle is between $[0, 2\pi[$, two different correction functions are required, one for odd and one for even cycles due to the double period error signal.

So, the problem of correcting the calculated results consists of two questions. First, it has to be known somehow whether a measurement result belongs to an odd or an even round. Second, the correction functions have to be determined.

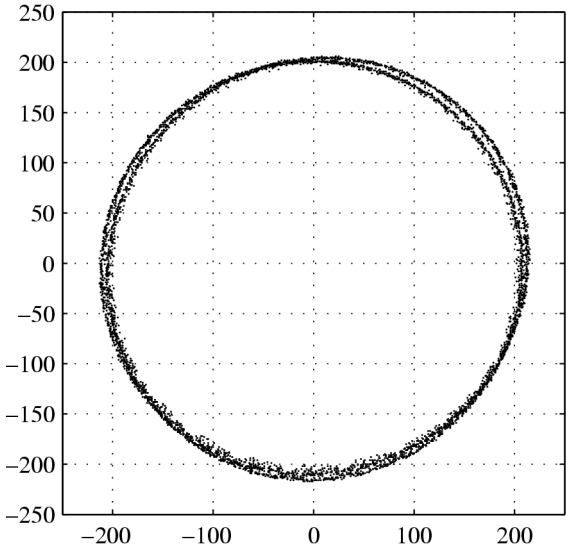


Figure 6.9 The zeroth order interference error splits the locus drawn by the data points into a double loop.

The first problem is one form of the standard problem with relative position encoders; the cycle number has to be determined from the data with some external signal processing by keeping book of the cycles. It is usually not enough to know where within one cycle the position is but also the number of cycle has to be known.

A simple method of cycle bookkeeping is:

$$\Delta\phi \leftarrow \phi_n - \phi_{n-1} \tag{6.13}$$

$$\Delta\phi \leftarrow \begin{cases} \Delta\phi + 2\pi & \text{if } \Delta\phi < -\pi \\ \Delta\phi - 2\pi & \text{if } \Delta\phi \geq \pi \\ \Delta\phi & \text{otherwise} \end{cases} \tag{6.14}$$

$$\Phi \leftarrow \Phi + \Delta\phi \tag{6.15}$$

where ϕ_n represent calculated angles from different sampling points. These steps are performed once for each sample. The algorithm assumes that the position changes

less than π to either direction between two consecutive samples. This condition is equivalent to the Nyquist sampling theorem, which requires the sampling frequency to be more than double the highest frequency to be sampled. The absolute position is accumulated into Φ.

When the absolute position is known, it is straightforward to find the position ξ between $[0, 4\pi[$ for each sample:

$$\xi = \Phi - 4\pi \text{ floor } \frac{\Phi}{4\pi} \tag{6.16}$$

where floor(x) gives the nearest integer below x.

As the error is periodic with period 4π, it is enough to find the measurement position correction function in range $[0, 4\pi[$. The corrected measurement results can then be obtained from the relation:

$$\Phi_{corr} = \Phi + f(\xi_m) \tag{6.17}$$

where $f(\xi_m)$ is the correction function and ξ_m measured positions given by (6.16).

Figure 6.10 depicts the function $g(\xi)$ which is the relation between measured (ξ_m) and real positions (ξ_r). As both functions are periodic (period $P = 4\pi$), both positions have to be lie between $[0, P[$, and as $0 \equiv P$, the endpoints of the curve have to be at $(0, 0)$ and (P, P).

If it is assumed that the movement does not have any significant noise components on the encoder spatial frequency, the cumulative probability of the real position ξ_r is even:

$$\mathrm{p}(\xi_r < a) = \frac{a}{P} \tag{6.18}$$

If a suitable function $g(\xi_m)$ exists, it will have an inverse $g^{-1}(\xi_r)$. So, the same probability can be used by replacing the real data with measured values and using the inverse function:

$$\mathrm{p}\left(g(\xi_m) < a\right) = \frac{a}{P} \tag{6.19}$$

$$\mathrm{p}\left(\xi_m < g^{-1}(a)\right) = \frac{a}{P} \tag{6.20}$$

It turns out that the inverse function can be determined from actual data collected with the sensor by simple statistical methods.

A large number (n) of separate measurement points are collected from the sensor. The angle ξ between $[0, 4\pi[$ is calculated from these points as described above. The points are then ordered in ascending order (σ_i), so that:

$$\sigma_i \leq \sigma_j \text{ iff } i \leq j \text{ for all } i, j \tag{6.21}$$

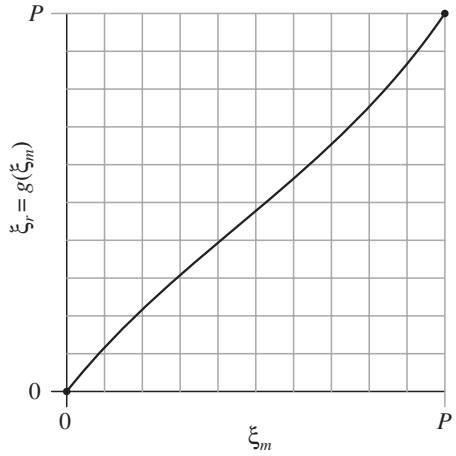


Figure 6.10 The function $g(\xi_m)$ between measured and real positions.

If the sample is representative, the probability that a measured value is below the k^{th} value in the sorted data is:

$$\mathrm{p}(\xi_m < \sigma_k) = \frac{k}{n} \tag{6.22}$$

If a in equation (6.19) is replaced by kP/n, then:

$$\mathrm{p}\left(\xi_m < g^{-1}\left(\frac{kP}{n}\right)\right) = \frac{k}{n} \tag{6.23}$$

This would suggest:

$$g^{-1}\left(\frac{kP}{n}\right) = \sigma_k \tag{6.24}$$

Naturally, this is not the definition of g^{-1}. The real data may not be representative, and the data will be noisy in any case. Also, only a finite number of discrete values is defined, not a continuous function. However, this suggests the use of measured data pairs:

$$\left(\frac{kP}{n}, \sigma_k\right) \tag{6.25}$$

which represent pairs:

$$(\xi_r, \xi_m) \tag{6.26}$$

The additive correction function $f(\xi_m)$ defined in (6.17) can be calculated from $g(\xi_m)$:

$$f(\xi_m) = \xi_r - \xi_m = g(\xi_m) - \xi_m \tag{6.27}$$

The data points representing this function can be obtained by subtraction from (6.25):

$$\left(\sigma_k, \frac{kP}{n} - \sigma_k\right) \tag{6.28}$$

To find the correction function $f(\xi_m)$, a suitable function has to be fit to this data. In practice, a Fourier series is a good candidate, as the errors are periodic in nature:

$$f(x) = a_0 + \sum_{j=i}^{\infty} a_j \cos\frac{2\pi x j}{P} + \sum_{j=i}^{\infty} b_j \sin\frac{2\pi x j}{P} \tag{6.29}$$

The simplest methods for calculating the coefficients a_j, b_j assume equal sampling intervals. In this case the sampling points (σ_k) are not evenly distributed, and thus different points have different weight. One possible method would be to form an interpolated continuous function from the data and integrate over the function. However, the real data points are noisy, and their number is large, so interpolation is not necessarily required.

A simple method of obtaining the data points to be used in calculating the coefficients is averaging consecutive data points (figure 6.11) and approximating the function with a piecewise continuous staircase function.

The cosine coefficients can then be calculated from the integral:

$$a_j = \frac{2}{P}\int_0^P y(x)\cos\frac{2\pi x j}{P}dx \tag{6.30}$$

$$= \frac{2}{P}\sum_{k=1}^{n}\int_{\sigma_{k-1}}^{\sigma_k}\frac{1}{2}\left(\frac{(k-1)P}{n} - \sigma_{k-1} + \frac{kP}{n} - \sigma_k\right)\cos\frac{2\pi x j}{P}dx \tag{6.31}$$

$$= \frac{2}{\pi j}\sum_{k=1}^{n}\left(\frac{P(k-1/2)}{n} - \frac{\sigma_k + \sigma_{k-1}}{2}\right)$$
$$\cos\frac{\pi(\sigma_k + \sigma_{k-1})j}{P}\sin\frac{\pi(\sigma_k - \sigma_{k-1})j}{P} \tag{6.32}$$

In this sum a zeroth point with zero value has been added to the sorted data set ($\sigma_0 = 0$).

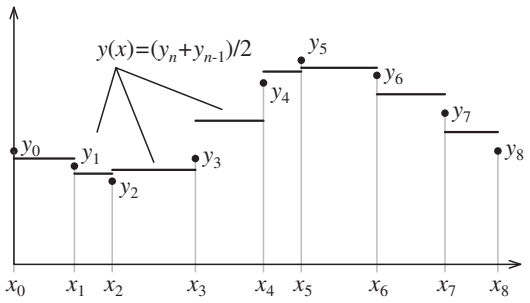


Figure 6.11 Approximating the discrete point set with a staircase function

In practice, the sine weighting term is small, as the distance between consecutive samples has to be small for the sample to be representative. Thus, the sine term can be approximated by $\sin\alpha = \alpha$, which gives:

$$a_j = \frac{2}{P}\sum_{k=1}^{n}\left(\frac{P(k-1/2)}{n} - \frac{\sigma_k + \sigma_{k-1}}{2}\right)(\sigma_k + \sigma_{k-1})\cos\frac{\pi/(\sigma_k + \sigma_{k-1})j}{P} \tag{6.33}$$

Similarly, the sine coefficients of the series are:

$$b_j = \frac{2}{P}\sum_{k=1}^{n}\left(\frac{P(k-1/2)}{n} - \frac{\sigma_k + \sigma_{k-1}}{2}\right)(\sigma_k + \sigma_{k-1})\sin\frac{\pi/(\sigma_k + \sigma_{k-1})j}{P} \tag{6.34}$$

The constant coefficient of the series (a_0) can be chosen freely, as there is no absolute zero position in the incremental sensor. However, it may be practical to preserve zero correction at $\xi_m = 0$. This will require the sum of the cosine coefficients to be zero:

$$\sum_{j=0} a_j = 0 \tag{6.35}$$

and

$$a_0 = -\sum_{j=1} a_j \tag{6.36}$$

The direct calculation of sine series coefficients in (6.33) and (6.34) is not computationally efficient if a large number of coefficients is required. For instance, making the data equally spaced by linear interpolation and using fast Fourier techniques (FFT) would be faster for a large number of coefficients. However, in this case only the few

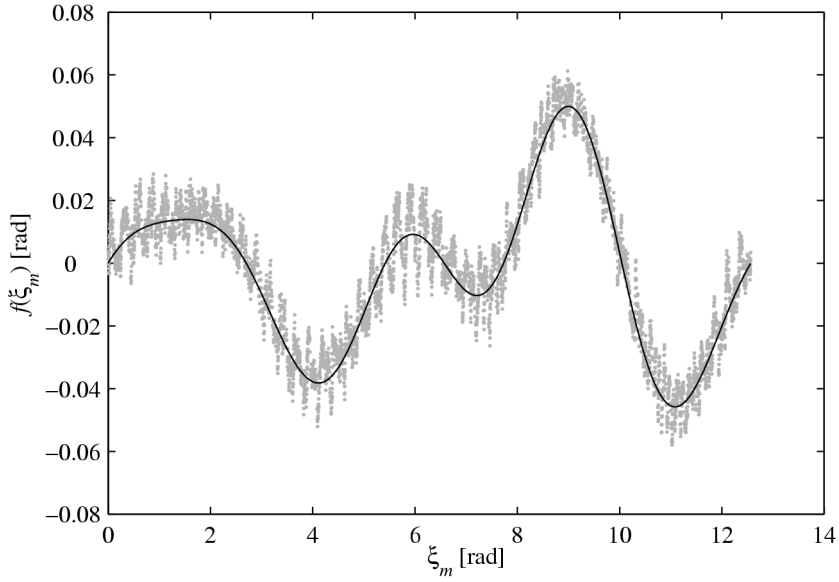


Figure 6.12 Data set used to calculate the correction function and correction function calculated by the five first sine and cosine coefficients.

first sinusoids are meaningful; high-frequency components contain more noise than signal.

Figure 6.12 shows the data pairs required to form the correction function and the corresponding correction function $f(\xi_m)$. The function utilizes the five first cosine and sine terms of the series. While the data set has certain jaggedness, the general trend is still very clear.

When this correction function is applied to the data depicted in figure 6.7, the results are shown in figure 6.13. The reduction of noise is significant, a careful estimate would be that approximately half of the noise has been removed.

However, even this correction does not bring the noise down to the theoretical levels calculated above. By looking at the data it seems that the required correction function changes significantly along the passage of the dipper carriage over the grating. This may be due to the y-direction movement of the carriage which will change the zeroth order interference significantly (see appendix B). It is also possible that some properties of the grating may vary over the x-axis.

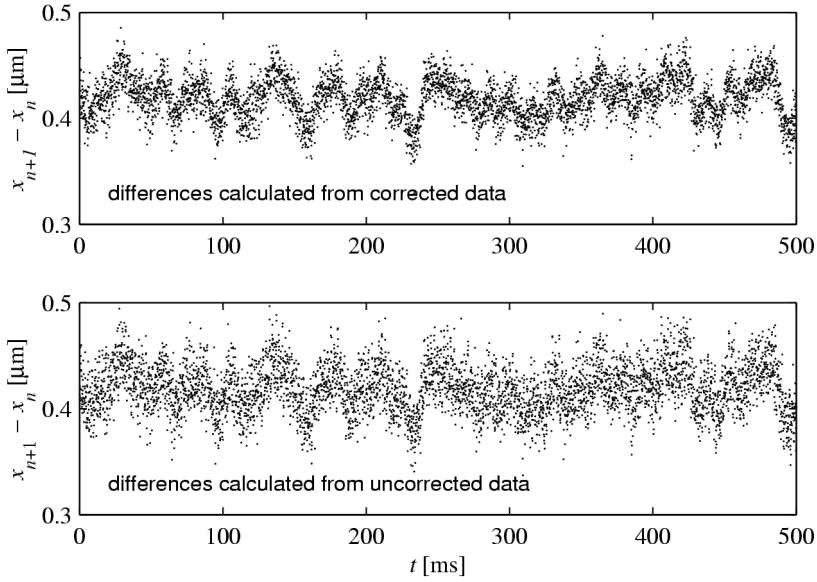


Figure 6.13 Velocity data corrected with the periodic correction function.

The correction method outlined above is not perfect. If very high precision is required the best correction method is measuring the actual position with, e.g., an interferometer and then creating a table of all position errors. However, the advantage of the method introduced here is that it can be applied *a posteriori* without any reference measurements.

Typical commercial interferometry devices have differential resolutions in order of 5 nm. The inexpensive position sensor used in this work seems to achieve differential error in the order of 20 nm when the novel correction method is used.

6.3 DIPPER MOVEMENT MEASUREMENTS

The dipper carriage movement was measured with several withdrawal velocities to determine the dynamic properties of the dipper. The data thus obtained has some clear patterns, most notably, slower movements suffer from large relative velocity noise.

Figure 6.14 depicts the dipper velocity profiles at the highest and lowest with-

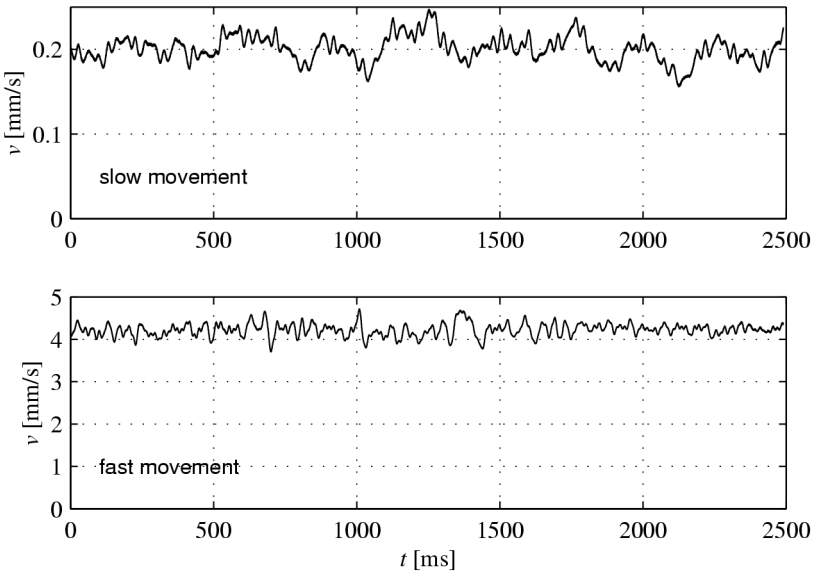


Figure 6.14 Dipping velocity as a function of time for a slow (v_0 = 0.200 mm/s) and a fast withdrawal velocity (v_0 = 4.22 mm/s).

drawal velocity settings. The position data has been smoothed out by running a 100 point (in this case a 10 ms) running average filter over the data. As the measurement noise in one point seems to be in the order of a few dozens of nanometers, this averaging practically removes the position measurement noise.

The averaging also cuts the frequency response significantly. It is assumed that the > 100 Hz components of the velocity noise are small and negligible from the film quality point of view.[2]

For the high withdrawal velocity the standard deviation of the velocity is approximately 4 % of the average velocity. For the low speed this ratio is doubled to 8 %. However, a much more significant change is in the noise spectrum. While the noise in the slow dipping contains a large amount of low-frequency components, faster movement is dominated by high-frequency noise in the fast dipping profile. It seems that

[2]This view is supported by some preliminary scanning electron microscope images which do not show any thickness variations on the micrometer scale. All perceived variations are in the millimeter scale, i.e. hertz or sub-hertz in frequency.

the frequency content of the noise scales with the dipping speed which would suggest that a large part of the noise is from the gearbox or the motor itself.

If all noise were deterministic and arising from the gearbox, then the percentage of noise should remain independent of the dipping velocity. This is not the case in the data. However, there are at least three phenomena which limit the high-frequency noise. First, as the drive system is slightly flexible (the string is bent), some of the gear noise may be absorbed. Also, as the velocity increases, forces required to produce similar noise percentage are larger, i.e. the inertia of the carriage becomes more significant.

The third possible explanation is that as the measurement has been carried out by using a moving average filter, some high frequency components may be filtered out. This explanation, however, does not seem to be likely as using a shorter (down to 10 point, i.e. 1 ms) running average does not increase the standard deviation significantly.

As the film thickness is approximately linearly proportional to the dipping velocity, the velocity variations measured in this case are large, in the case of the slow movement the maximum peak-to-peak variation is almost one half of the average velocity. This variation does produce some visible non-uniformity to the films. The non-uniformity produced by this equipment in the commonly used withdrawal velocity range is acceptable for the sensor element manufacturing. However, if tighter control over the optical parameters of the film is required, the dipper movement has to be made smoother.

7 Film parameter determination

Sol-gel thin film manufacturing methods are relatively simple compared to vacuum deposition methods. However, the wet process has one significant disadvantage; the film thickness cannot be measured during the deposition in real time, as the final thickness is formed during the drying phase.

The film thickness in the dipping process is repeatable over a short period of time. Over longer periods of time (weeks or months) the properties of the sol may change due to continuing chemical processes so that the layers become thinner or thicker. So, the film thickness and refractive index parameters have to be found for each dipping session.

It should be noted that the refractive index changes in the film are relatively small, and usually it is enough to control the changes in the optical thickness of the film (i.e. the product of the real thickness and refractive index). This simplification reduces the number of variables to one per film layer.

The optical thickness of a single non-absorbing layer can be determined from the reflection or transmission spectrum of the layer. Unfortunately, this method requires high accuracy if the refractive indices of the substrate and the film are close to each other, and it is thus impractical in many cases.

7.1 SPECTROPHOTOMETER CONSTRUCTION

A simple spectrometer was constructed to measure the transmission of a glass substrate with sol-gel thin film(s) (figure 7.1). An intensity-controlled incandescent bulb is used as the light source. Its light is condensed to a small (500 μm) aperture with a pair of condensing lenses and a mirror. The light is then passed through the glass and finally to the end of a fiber carrying the light to a spectrometer.

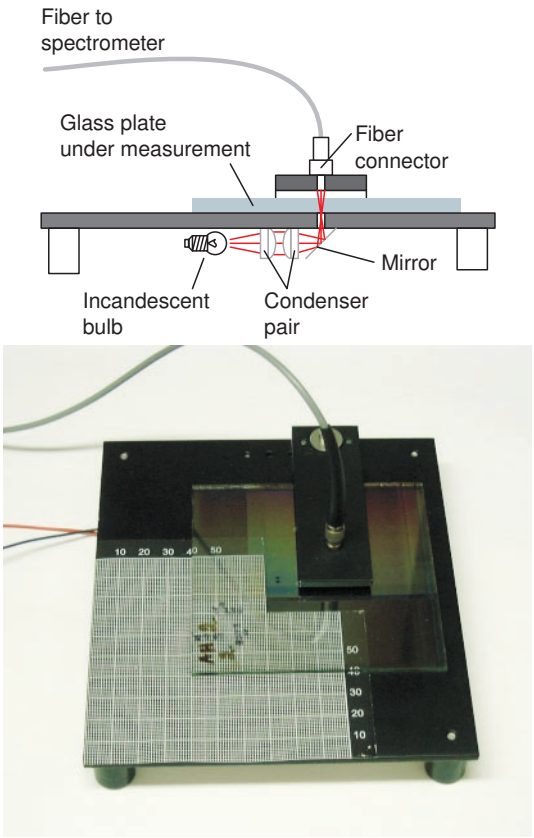


Figure 7.1 Spectrophotometer constructed for transmittance measurements.

An ordinary miniature incandescent bulb was chosen as light source due to its very smooth (black body) spectral characteristics. The output of the lamp is kept constant by using a pulse-width modulation driving signal with feedback from a photodiode. Otherwise the lamp output would fluctuate rather significantly even over short periods of time.

The condensing optics projects the image of the incandescent filament to the spectrometer aperture. The purpose of the aperture is to remove stray light. Stray light would both change the measurement geometry by introducing non-normal rays and produce measurement error in form of rays which have undergone several reflections in the glass plate.

Another aperture is placed in front of the fiber entrance to reduce stray reflections from the fiber connector. To reduce the reflections further, coated glass substrates are measured coating down (i.e. facing the light source). As the coated surface reflects significantly more than the non-coated surface, this arrangement directs most of the reflected light back to the light source side of the system.

The spectrometer element used in the instrument is a microspectrometer manufactured by MicroParts GmbH [54]. Light is coupled to the spectrometer through a 105/125 μm multimode glass fiber. Light emerging from the end of the fiber is decomposed to its spectral elements with a monolithic molded optical element which is attached on top of a linear photodiode array detector. The part is specified to have approximately 3 nm pixel resolution and 12 nm spectral peak half width. While this resolution is not sufficient for chemical spectral analysis applications, it is good enough for the relatively smoothly changing spectra of the multi-layer coatings with only a few layers.

The dispersion of the array has been calibrated with a two-point calibration at 633.0 nm[1] and 546.0 nm produced by a HeNe laser and by the mercury emission peak from a fluorescent lamp. The calibration results are well in accordance with the manufacturer specification of 2.93 nm/pixel.

No intensity calibration has been performed to the device. The final spectral measurement result is determined from three measurements, a dark measurement D (with black plate in place of the glass substrate), a reference measurement R (with a glass substrate without coatings) and the spectral measurement S. The measurement results are calculated from the formula:

$$T(\lambda) = \frac{S(\lambda) - D(\lambda)}{R(\lambda) - D(\lambda)} \tag{7.1}$$

[1] All wavelengths discussed here are vacuum wavelengths, hence the difference to the more familiar 632.8 nm.

The dark measurement cancels dark currents and amplifier biases in the photodiode array. The reference measurement cancels the spectral non-uniformity of the light source and gain variations in the pixels.

The errors resulting from these slow variations could be decreased by using interpolated reference measurement values instead of a single measurement. Another possibility would be to actively stabilize the temperature of the photodiode array. On the other hand, the practical usability of the system is not limited by this behavior, and thus additional corrections have not been incorporated.

There is very little noise in the external read-out electronics as the photodiode array output signal is directly digitized with a 12 bit AD converter. However, there is a lot of noise in the output signal, especially if long integration times or small integration capacitances are used. This noise seems to have white noise characteristics, and averaging several readings (up to 64) reduces the noise in accordance with the square root law (noise is proportional to the inverse square root of the number of samples).

Figure 7.2 gives an example of a typical measurement result obtained with the spectrophotometer to illustrate the noise behavior of the instrument. The increased noise levels in the short wave end of the spectrum are due to the low spectral content of the incandescent light source in that region and also due to the decreasing sensitivity of the diode array.

The most important limitations of the measurement system come from the performance of the microspectrometer. The stray light sensitivity of the component is specified to be below one percent at 470 nm when illumination is at 510 nm. This level of stray light is still quite significant as the light source is not spectrally flat; the mid-wavelength peak of the light source will produce significant errors in the short wavelength end of the spectrum. Also, significant near-infrared emissions from the light source may produce errors in the visible part of the spectrum.

In order to remove or diminish these effects it is possible to use a spectrally flattening filter in front of the light source. Also, an infrared blocking filter would improve the situation. As the stray light is a deterministic phenomenon, it can be reduced algorithmically, as well. However, while the accuracy of the simple spectrometer is not very high, it has proven to be a very useful tool in determining dipped film thicknesses even without any correction methods.

A reflective measurement would have been better in some situations, primarily because the instrument under development (optical pH measurement instrument) uses reflective principles. However, a reflective measurement at normal inclidence is geometrically difficult to realize.

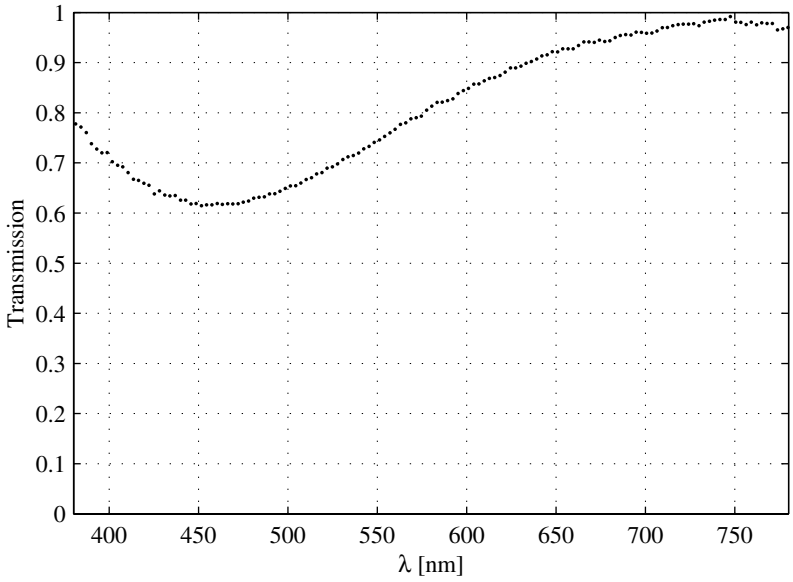


Figure 7.2 A transmission spectrum measured with the spectrophotometer.

7.2 CHECKERED FILMS

Accurate measurement of a single low-index film requires high accuracy from the spectrophotometer used to measure the film thickness as the reflection coefficients are small. Also, single layer measurements may not be very accurate, as the densification in lower layers during later drying stages may still change the layer thickness and refractive index (see section 5.7).

Fortunately, one adjustable parameter per layer, i.e. the withdrawal velocity, is sufficient to control the optical thickness of the layer. Drying conditions do affect the film thickness and porosity, but it is relatively simple to control the drying temperature and time, so there is little thickness variation from the drying.

A three-layer mirror with one parameter per layer gives a three-dimensional space for all possible combinations. One way of finding the dipping parameters would be the hard-work approach of choosing points in this space and using them in dipping. Unfortunately, choosing only ten different withdrawal velocities per layer would give one thousand combinations. Making a thousand different film stacks just to find the

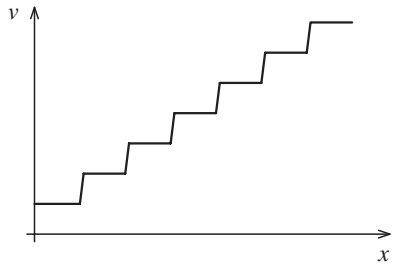


Figure 7.3 A withdrawal velocity profile as a function of position used in dipping staircase profile films.

best parameters is practically impossible.

The number of parameters may be reduced to two for a three-layer HLH mirror if the dipping velocities of the bottom and top layer are taken to be the same. This will probably produce slightly thinner bottom layer than top layer but as the difference is small, its impact on the mirror bandwidth and reflection is small (see section 5.7).

For two parameters there is a simple method of producing a large number of thickness combinations on a single substrate. As the dipping process is computer controlled, and there is an accurate position measurement device in the system, it is possible to vary the withdrawal velocity as a function of distance. This makes it possible to produce a staircase-type thickness profile (figure 7.3).

The staircase profile makes it possible to vary the film thickness in the dipping direction. The second direction can be used if the substrate is rotated 90° between dippings. Starting from the bare substrate the staircase profile is first dipped with the high index BST sol. This layer is then dried, and the second—low index—layer is dipped on the 90° rotated substrate. The last high-index layer is dipped in the original orientation. This way a checkerboard thickness variation is produced (figure 7.4).

The same method can be extended to three parameters if the staircase profile is split into smaller strips. For example, all 100 combinations of 10 different withdrawal velocities per layer can be made in one direction if the dipping velocities for the layers are as depicted in figure 7.5.

However, splitting the substrate into very narrow strips emphasizes errors caused by abrupt changes in the dipping rate. In practice, the strips should be several millimeters wide to reduce alignment problems between consecutive dippings in the same direction.

Making a large number of different film thickness combinations on a single sub-

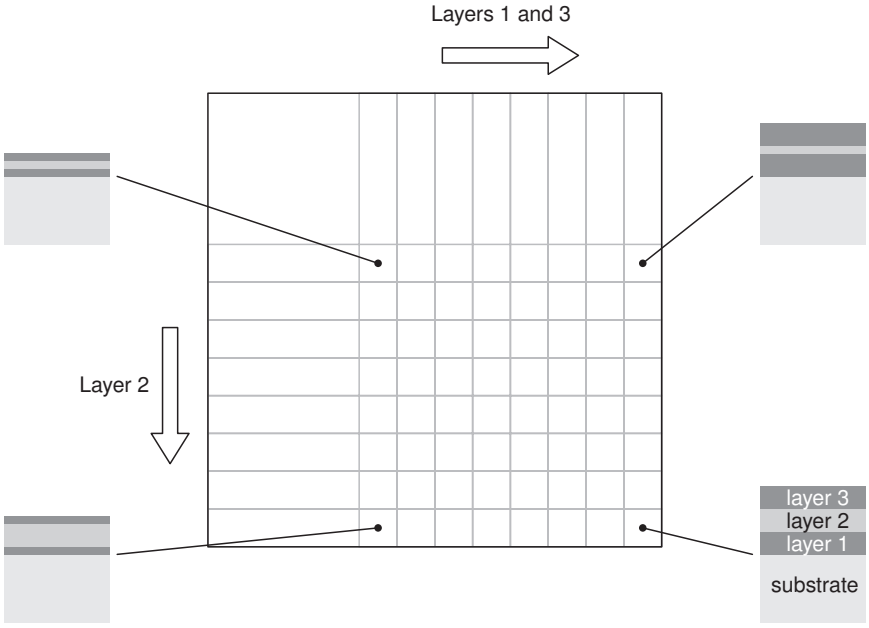


Figure 7.4 The thickness variation obtained by two-direction staircase profile dipping.

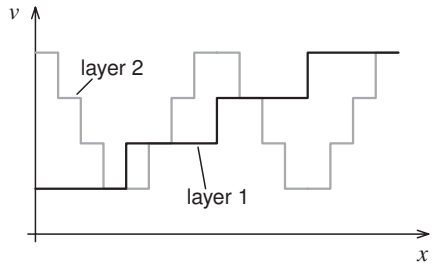


Figure 7.5 A withdrawal velocity profile which enables the variation of two parameters in the same direction.

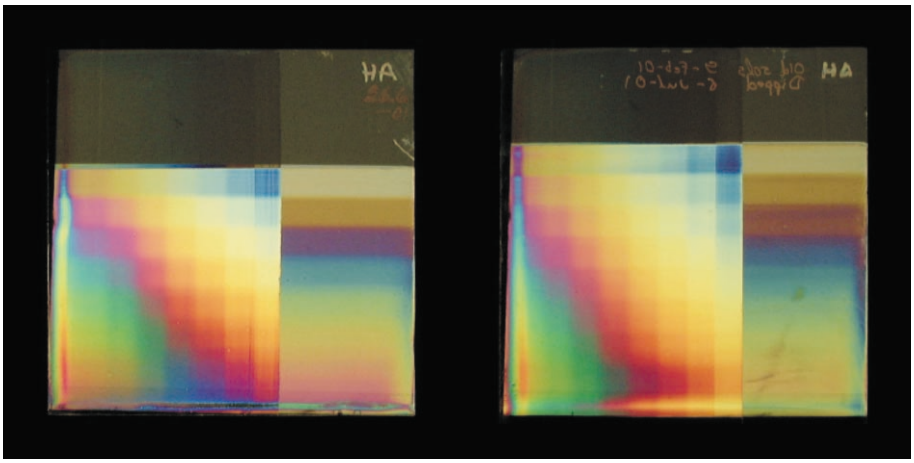


Figure 7.6 Two different checkered mirrors manufactured from similar sols with different ageing. (Older sols are used on the right substrate.)

strate does remove the need of dipping a large number of substrates during the parameter determination phase. However, a large number of measurements is still required per substrate. If the measurement is not automatized, the effort required to measure, say, a hundred spectra is considerable.

In the two-parameter case there is a simple visual method to see the overall behavior of the sol compared to some reference. Figure 7.6 shows reflection images of two substrates with checkered mirrors. The mirror on the left has been dipped with freshly prepared sols whereas the mirror on the right has been dipped with sols which have been aged for some months.

The color shift between the two mirrors is clearly visible. BST layers thicken downwards, thicker BS layers are on the left. The color shift seems to be towards lower left corner from new to old sols, which indicates thinning of both layer types[2].

The visual method gives a good overall view of the behavior of the different layers. As the thicknesses of the different types of sols (low and high index) are on different axes, different sol parameters can be tested very quickly. Also, the visual test may be used to see if a sol has changed its original behavior. The human eye is not sensitive

[2]While it would be expected that the sols become thicker during ageing, these sols seem to have considerable amount of reverse reactions to polycondensation. This also explains the surprisingly long pot life of the sols.

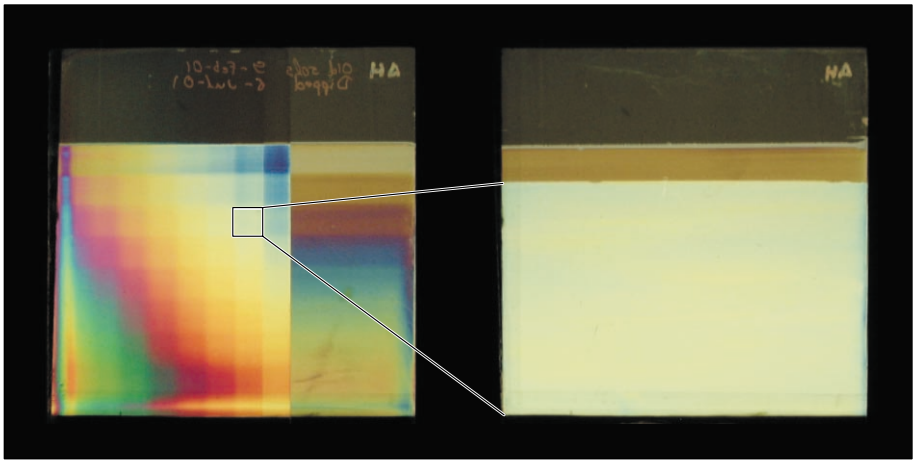


Figure 7.7 A mirror manufactured by choosing a suitable patch from a checkered mirror.

to all color changes, but as there are several colors, the overall color shift is clearly visible.

An added benefit to this method is that when the suitable mirror is found on the checkered substrate, its manufacturing parameters are directly available (figure 7.7). If the desired mirror parameters seem to lie in between of color patches, a magnification of that area can be produced by interpolating the dipping parameters.

The actual film parameters can be determined from the color patches by fitting the calculated film properties to the measurement data. The thickness data points in figure 7.8 have been obtained by using a least-square method with constant film refractive indexes and same thicknesses to the bottom and top film.

These assumptions are not very accurate. At least the high-index titania film exhibits significant dispersion, and also the film thicknesses may vary between the bottom and top of the stack. However, this simplified model does produce reasonable results and serves well in illustrating the potential of the checkered dipping.

The curve fitting gave refractive indices $n_L = 1.45$ for the borosilicate film and $n_H = 1.97$ for the borosilicate-titanate film. It should be borne in mind that these values are only approximations. The curve fitting produces almost as good results with slightly different indices, so it is not a reliable method of accurate refractive index (and thus porosity) measurement.

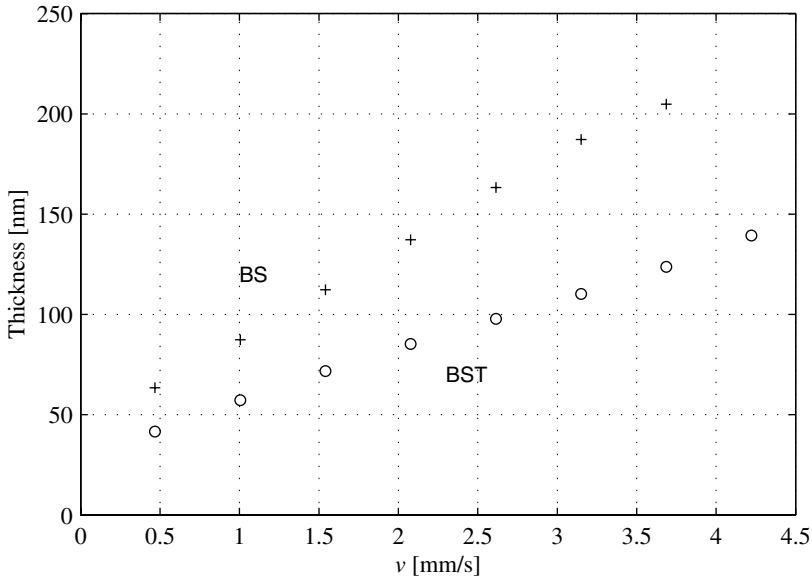


Figure 7.8 Mirror thickness parameters obtained by fitting calculated spectra to spectra measured from a checkered mirror.

8 Indicator color measurement

Previous chapters have concentrated on how to make the chemically sensitive sensor element with an indicator layer and a dielectric mirror. While the sensor element is the heart of the instrument, the realization of the optical and electronic system and signal processing methods is equally important from the accuracy, reliability, manufacturability, and cost point of view.

8.1 CALCULATION OF MEASUREMENT RESULTS

In an optimal situation a single wavelength measurement would give the measurement result if the wavelength were placed in a region of changing absorption in the indicator spectrum (e.g. wavelength λ_B in figure 8.1). In practice, this measurement would fail in several situations. If the external refractive index changes (section 5.5), the measurement reading would change without any possibility of compensation. Also, if the amount of dye in the indicator layer changes (due to escaping indicator molecules) the measurement result will change.

To compensate for the three free parameters—external pH, external refractive index, and changing dye concentration in the indicator layer—at least three different wavelengths are required.

The measurement signal for one wavelength is:

$$m = R(\lambda, n)T(\lambda, \mathrm{pH}, c)C(\lambda) \tag{8.1}$$

where R is the intensity reflection coefficient of the mirror, T transmission of the indicator layer, and C a factor which depends on the light source, optical system, and detector spectral properties.

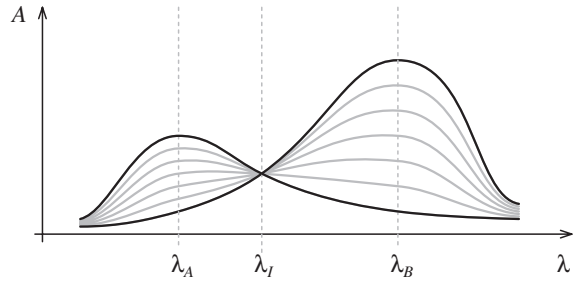


Figure 8.1 Different measurement points of a spectrum.

Reflection coefficient of the mirror depends on the wavelength and the external refractive index. The mirror should be designed so that changing external refractive index changes only the absolute reflectivity, not the shape of the reflection spectrum. If this is the case, R can be written as a product of a wavelength dependent and a refractive index dependent function:

$$R(\lambda, n) = r(\lambda)r_0(n) \tag{8.2}$$

The transmission of the indicator layer depends on the dye concentration, dye absorption, and layer thickness. The layer thickness is constant, but dye absorption changes as pH changes. Also, the dye concentration c may decrease over time. By using the Beer-Lambert law, T can be written as:

$$T(\lambda) = e^{-a(\lambda,\mathrm{pH})c} \tag{8.3}$$

where $a(\lambda, \mathrm{pH})$ is the absorption of the dye as a function of wavelength and pH.

By combining (8.1), (8.2), and (8.3), three-wavelength measurement gives:

$$m_1 = r(\lambda_1)r_0(n)e^{-a(\lambda_1,\mathrm{pH})c}C(\lambda_1) \tag{8.4}$$

$$m_2 = r(\lambda_2)r_0(n)e^{-a(\lambda_2,\mathrm{pH})c}C(\lambda_2) \tag{8.5}$$

$$m_3 = r(\lambda_3)r_0(n)e^{-a(\lambda_3,\mathrm{pH})c}C(\lambda_3) \tag{8.6}$$

The refractive index dependent term can be eliminated by dividing any of these measurement results by another of them. Two different measurement pairs can be chosen from the set of three measurements without redundancy. Taking the logarithm

of the quotients thus obtained gives the result:

$$\ln \frac{m_1}{m_3} = \ln \frac{r(\lambda_1)C(\lambda_1)}{r(\lambda_3)C(\lambda_3)} + \left(a(\lambda_3, \mathrm{pH}) - a(\lambda_1, \mathrm{pH})\right) c \tag{8.7}$$

$$\ln \frac{m_2}{m_3} = \ln \frac{r(\lambda_2)C(\lambda_2)}{r(\lambda_3)C(\lambda_3)} + \left(a(\lambda_3, \mathrm{pH}) - a(\lambda_2, \mathrm{pH})\right) c \tag{8.8}$$

The first term in each of these expressions is a constant which depends on the instrument structure and sensing element properties. For the sake of simplification, these constants are from now on referred to as k_1 and k_2.

To make the measurement interpretation easier, the two expressions in (8.4) can be joined to form the real and imaginary parts of a response function $f(\mathrm{pH}, c)$:

$$f(\mathrm{pH}, c) = \ln \frac{m_1}{m_3} + i \ln \frac{m_2}{m_3} = k_1 + ik_2 + c\left(a_1(\mathrm{pH}) + ia_2(\mathrm{pH})\right) \tag{8.9}$$

where

$$a_1(\mathrm{pH}) = a(\lambda_1, \mathrm{pH}) - a(\lambda_3, \mathrm{pH}) \tag{8.10}$$

$$a_2(\mathrm{pH}) = a(\lambda_2, \mathrm{pH}) - a(\lambda_3, \mathrm{pH}) \tag{8.11}$$

The first term $k_1 + ik_2$ is a constant which can be subtracted from the measurement result. Writing the variable term in polar form (argument and magnitude) gives:

$$\left|c\left(a_1(\mathrm{pH}) + ia_2(\mathrm{pH})\right)\right| = c\left|a_1(\mathrm{pH}) + ia_2(\mathrm{pH})\right| \tag{8.12}$$

$$\arg c\left(a_1(\mathrm{pH}) + ia_2(\mathrm{pH})\right) = \arg\left(a_1(\mathrm{pH}) + ia_2(\mathrm{pH})\right) \tag{8.13}$$

The functions a_1 and a_2 depend only on the wavelengths and the dye, and are thus known. If these functions are such that only one value of pH corresponds to one argument value, the measurement result can be calculated from the argument alone. After the pH has been determined, the concentration can be calculated from the absolute value by a simple division.

The products $r_0(n)r(\lambda)C(\lambda)$ can be calculated from the original measurements and knowledge of the absorption function $a(\lambda, \mathrm{pH})$. As $r(\lambda)C(\lambda)$ is a constant, the external refractive index can be determined this way.

The calculations above show that any three wavelengths λ_N will give the measurement results if there is a one-to-one mapping between (8.13) and pH. There is no guarantee this condition is satisfied in the general case, e.g., the dye may have exactly the same spectrum at two different pH values. Fortunately, in the simple case of a binary dye (see section 2.4.3) the situation is straightforward to analyze.

The absorptivity of a binary dye is:

$$a(\lambda) = p_{\text{acid}} a_a(\lambda) + (1 - p_{\text{acid}}) a_b(\lambda) \tag{8.14}$$

where $p_{\text{acid}}(\lambda)$ is the proportion of acid form dye molecules, and $a_a(\lambda)$ and $a_b(\lambda)$ absorptivities of the acid and base forms of the dye molecule, respectively.

In this case functions a_1 and a_2 (8.10) take the form:

$$\begin{aligned} a_1(\text{pH}) &= p_{\text{acid}}(\text{pH}) \left[a_a(\lambda_3) - a_a(\lambda_1)\right] && (8.15)\\ &\quad + (1 - p_{\text{acid}}(\text{pH})) \left[a_b(\lambda_3) - a_b(\lambda_1)\right] && (8.16) \end{aligned}$$

and

$$\begin{aligned} a_2(\text{pH}) &= p_{\text{acid}}(\text{pH}) \left[a_a(\lambda_3) - a_a(\lambda_2)\right] && (8.17)\\ &\quad + (1 - p_{\text{acid}}(\text{pH})) \left[a_b(\lambda_3) - a_b(\lambda_2)\right] && (8.18) \end{aligned}$$

These are a linear functions of $p_{\text{acid}}(\text{pH})$. In the color change region of the dye $p_{\text{acid}}(\text{pH})$ is a monotonous function of pH. As long as the following condition is *not* satisfied:

$$a_a(\lambda_3) - a_a(\lambda_1) = a_b(\lambda_3) - a_b(\lambda_1) \quad \wedge \quad a_a(\lambda_3) - a_a(\lambda_2) = a_b(\lambda_3) - a_b(\lambda_2) \tag{8.19}$$

the locus of $f(\text{pH}, c)$ is a line segment on the complex plane, and the measurement result can be resolved. Condition (8.19) is fulfilled in two cases. Either the indicator molecule does not change its absorption on any of the measurement wavelengths as a function of pH, or then the absorption changes the same amount on all wavelengths. The former case means the indicator dye is not an indicator at all (at least not on the chosen wavelengths). In the latter case the most likely explanation is that the wavelengths are chosen very near each other.

While it can be concluded that almost any choice of wavelengths will give a measurement result in theory, the choice of measurement wavelengths will affect the measurement significantly in practice.

With a binary dye the locus of $a_1(\lambda)+ia_2(\lambda)$ is shown in figure 8.2. A measurement point m is shown on the diagram with its uncertainty. It is evident from the image that the distance of the line segment from the origin should be as large as possible to give the smallest measurement errors.

It seems reasonable to choose λ_1 and λ_2 so that they correspond to the acid and base absorption maxima. It should be noted that this is not a generally optimal choice, but a good starting point with ordinary dyes.

The choice of the third (reference point) is slightly more complicated; any point which has very low absorption for both the acid and base form would be suitable.

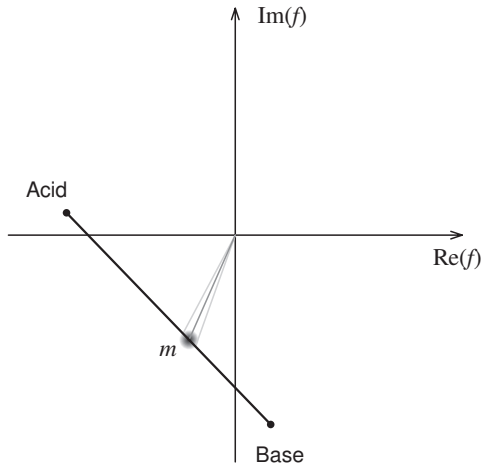


Figure 8.2 Locus of measurement points in the case of a binary dye.

Unfortunately, these points tend to be located far away from the maxima wavelengths. If the wavelengths are far away from each other, the approximation (8.2) becomes invalid for the dielectric mirror. So, the third point is preferably chosen from between the two other points.

In this case the isosbestic point (wavelength λ_I in figure 8.1) is a good choice. While it is not necessarily optimal from the mathematical point of view, it gives a fairly simple diagnostic method for the instrument, as the measured value at the isosbestic point is independent of pH.

In practice, real dyes do not exhibit exactly binary behavior. Figure 8.3 depicts the loci calculated from the measured absorption of bromophenol blue (BPB). While the loci are not exactly line segments, the error is small and the calculation methods developed above are valid.

The direct measurement results from the instrument are three intensity values. As equations (8.4) shows, there are several constants and correction factors which have to be determined in some way.

In order to find the pH value, the constants k_1 and k_2 in (8.9) have to be known. One way to accomplish this is to take two pH points such that at one point (pH_1) $a_1(\lambda) = 0$ and at another point (pH_2) $a_2(\lambda) = 0$. Then the measurement values at these

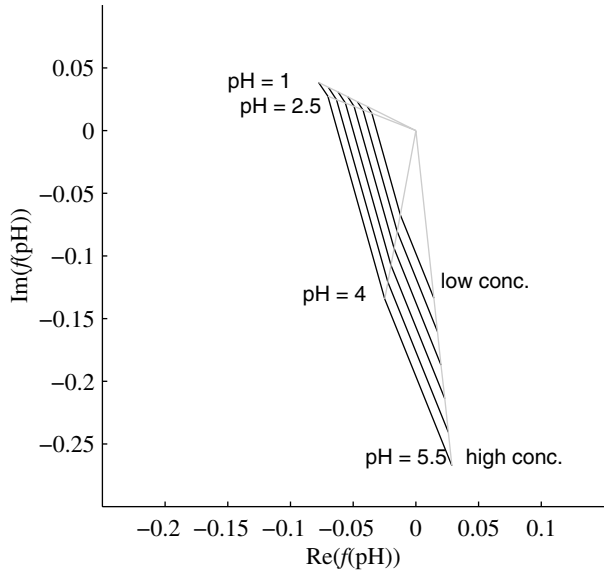


Figure 8.3 Loci calculated from BPB (aqueous solution) absorption spectra.

points are, according to (8.9):

$$f(\mathrm{pH}_1) = k_1 + i\,[k_2 + ca_2(\mathrm{pH}_1)] \tag{8.20}$$

$$f(\mathrm{pH}_2) = [k_1 + ca_1(\mathrm{pH}_2)] + ik_2 \tag{8.21}$$

The parameters k1 and k2 are thus:

$$k_1 = \mathrm{Re}(f(\mathrm{pH}_1)) \tag{8.22}$$

$$k_2 = \mathrm{Im}(f(\mathrm{pH}_2)) \tag{8.23}$$

These two points can be used in determining the initial concentration of the dye (c) . Also, the initial values of $r_0(n)r(\lambda)C(\lambda)$ for each wavelength can be determined from these measurements.

Even though the primary measurement result is a single number (pH), the two other results (c, n) can be used for diagnostic purposes. Naturally, if c falls below some preset limit, the sensor membrane has to be replaced. Also, if c rises above its initial value, something is certainly wrong in the measurement system.

The measurement uncertainty of n is quite large as the mirror is intentionally designed to be as immune against changes in n as possible; the instrument is a lousy refractometer. However, large changes in the calculated value of n indicate significant changes in the mirror total reflection. One important application to this is recognizing the empty process pipe; if there is air immediately adjacent to the sensor membrane, the refractive index drops very significantly. Also, n can be calculated from all three measurement values. If the calculated values differ from each other, that is an indication of a problem in the measurement system.

8.2 OPTICAL CONSIDERATIONS

The optical system has to produce repeatable results even in difficult conditions. Process pipes may vibrate significantly, and this requires the optical system to be rigid. All adjustments should be eliminated, as adjusting elements may loosen under vibration and thermal changes. Also, adjustments make the assembly process more difficult and prone to errors.

If adjustments are to be eliminated, the optical system has to allow certain tolerances. For instance, lens manufacturers usually give several percent tolerances to lens focal lengths [55, 56]. In practice, if the lenses are chosen from the available stock focal lengths, they have usually much tighter tolerances. In any case, these tolerances have to be taken into account in the system design.

The indicator element will be optimized for normal incidence measurement. Naturally, the element could be optimized for other measurement angles, but then the angular tolerances would be tighter. Small errors in the angle near the normal are not important, as the optical thicknesses of the layers are proportional to the cosine of the angle. The use of normal incidence makes it possible to fit the measurement into a tight space.

While it seems prudent to keep the angular divergence of the measurement beam small, the area of measurement should be as large as possible. If the measurement area is small, even a small scratch or other damage will make it impossible to obtain good measurement results. Fortunately, the combination of large area and small divergence is relatively easy to produce, as the product of beam area and divergence keeps constant (or increases) in a lossless system.

In addition to these technical requirements, the optical system has to be inexpensive to manufacture. While custom molded optics or diffractive optics may answer some questions, their use is prohibitively expensive until the manufacturing volumes are high. The system should optimally be realized with only standard stock lenses of medium size or with lenses manufactured using off-the-shelf grinding tools. Extrema

in lens size or focal length increase the lens price, as well as non-standard focal length specifications.

Reliability, manufacturability, and price considerations do all suggest the use of as few optical elements as possible.

8.3 LIGHT SOURCES AND DETECTORS

Very much the same requirements apply to optoelectronic components as to the optical components; they have to be able to survive high vibration for long periods of time, and they should not need any adjustment. The acquisition cost of the components has to be reasonable, as well.

In addition to these general requirements, there are some specific electronic requirements. As the instrument is to be used also in chemical factories, it should preferably comply with the regulations concerning instruments used in explosive environments. In practice, the electronic circuitry has to operate with a low voltage, low current, and small capacitance values. For example, typical values could be 5 V, 50 mA, and a few hundred nF total capacitance.

In practice, only semiconductor light sources fulfill these power consumption requirements. The choice has to be made between light emitting diodes (LED) and semiconductor lasers (diode lasers). Diode lasers have usually a rather high lasing threshold, and their availability is currently limited to red (> 630 nm) and blue-violet (430 nm). Also, their price is high compared to that of LEDs, so in practice LEDs offer the only viable possibility.

LEDs are almost monochromatic light sources. Depending on the LED junction structure, the wavelength peak width is a few dozens of nanometers. This improves the total system efficiency from electric energy to optical energy on the desired wavelength. Wideband emitters would require filtering either on the emitter or on the detector side of the system.

The detector choice is also rather simple. Amplified high-efficiency detectors (photomultiplier tubes, avalanche photodiodes) are expensive and require high operating voltages. Thus, only the simple photocurrent generating junction photodiodes and transistors remain possible. Phototransistors are less linear and may have largely temperature dependent gain, so the use of photodiodes is a natural choice.

Photodiodes are very inexpensive and linear over a large range of incident radiation intensities. They have significant and temperature dependent bias currents, but this current is additive and does not disturb the linearity of the sensor. Over large temperature variations also photodiodes have temperature dependent gain [57] which has to be taken into account in high-precision applications

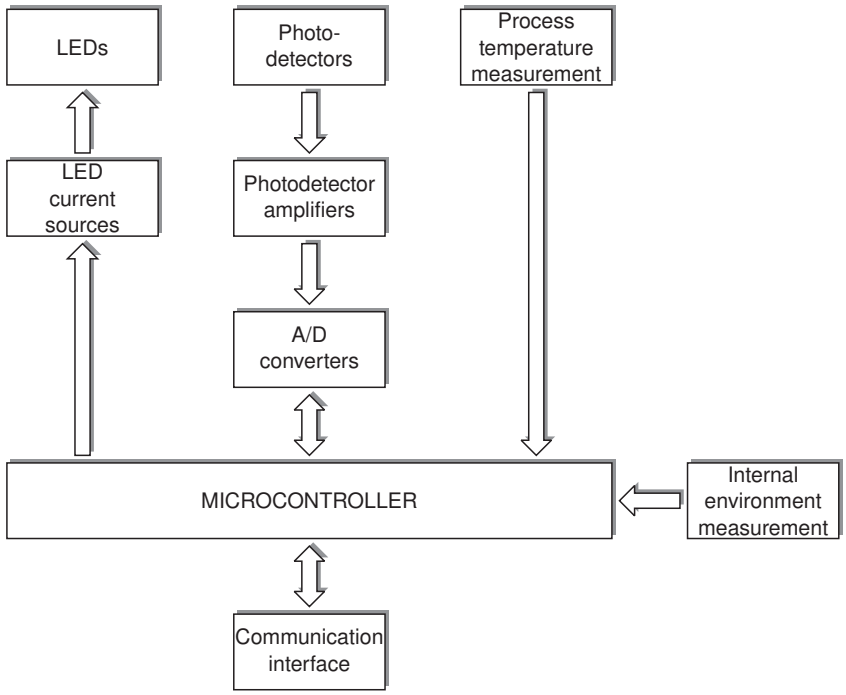


Figure 8.4 Block diagram of the pH measuring instrument.

8.4 MEASUREMENT ELECTRONICS

Essentially, the electronic part of the sensor has to control the light sources, convert the measurement photodiode signals to digital form, and to calculate the result (figure 8.4). Some interface is also required to the external world. The pH measurement is not useful if the instrument is not able to measure temperature from the process as well. Internal temperature and vibration measurements may be added for diagnostic purposes. Internal humidity measurement could warn about possible leakages before any damage occurs.

The light sources (three LEDs) are driven with constant current to make them

more stable. This does not necessarily ensure constant light output, at least not at the level required in this application, so other compensation methods are still required. However, driving the LEDs with, say, only a series resistor would make the circuit very vulnerable to small variations in the supply voltage and temperature.

The detector amplifiers are very simple single-stage transimpedance amplifiers which transform the small currents to measurable voltages. The measurement photodiodes are reverse-biased to decrease their capacitance. Reverse-biasing may increase the leakage currents, but as the DC leakage is easy to compensate for, this is not a significant problem. Two photodetectors are required; one measures the light emitted by the photodiode, and the other measures the reflected light. This way thermal and other intensity fluctuations can be compensated for.

The voltage signals from the amplifiers are directly digitized without further filtering or amplification. Digital signal processing is generally more flexible, less noisy, and consumes less power than analog processing of slow signals. Preferably, both the reflected light and the reference signal are sampled simultaneously with identical AD converters.

The LEDs and AD conversions are controlled by a microcontroller. This controller takes care of the sequencing of the measurement, data processing, result calculation, diagnostic routines, and external communications. Even though the processor has numerous tasks, the measurement is rather slow so that in practice any common microcontroller is sufficient to the task.

The sequencing of the measurement has to be such that the dark signal voltages can be compensated for. The simplest way is to light the LEDs alternatingly and keeping a dark period between the sequences. This method may produce erroneous results if the reflectance change is not slow compared to the sequence repetition frequency. If a synchronous measurement is required, then the LEDs can be cycled with some orthogonal binary sequences (such as the Hadamard–Walsh codes). The system would then be analogous to the CDMA (Code Division Multiple Access) multiplexing used in communications technology.

The external communication method depends on the application. In industrial applications the most common communication method is still the traditional $4-20\,\mathrm{mA}$ current loop. Different field bus solutions are also becoming more common, but they require much more complicated interface electronics and software. An intermediate solution would be to use a combination of the $4-20\,\mathrm{mA}$ signal and some simple digital serial interface combined (e.g., HART).

In this form the electronic design is straightforward. The circuit uses only a few components, and there are no special components. The electronics can be manufactured with moderate cost and by using well-known and reliable circuit solutions.

8.5 ERROR SOURCES

There are several possible error sources in the system. While most of these can be compensated for, it is important to recognize them.

8.5.1 Sensing element errors

As long as the sensing element functions according to equation (8.4), the calculation methods described above produce accurate results. However, there are some factors which have not been taken into account in the calculations.

First, the mirror may be somehow damaged. This will change its reflectivity, in practice make it smaller. The most important way to detect this is to compare the external refractive index values calculated from the measured intensity values at each wavelength. If the values start to differ significantly from each other, the reflectivity of the mirror has changed and the measurement results are not reliable.

The dye may creep out of the indicator layer. This can be readily compensated for, but if some external dye creeps into the indicator layer, the situation is much more complicated. In some cases this will be visible only in the external refractive index readings but if the external dye color is near to that of the acid or base form of the indicator dye, a significant measurement error is introduced.

In theory, the external refractive index does not have to be the same at all wavelengths. Dispersion will produce small errors, and strongly colored process liquids will have large differences in their extinction coefficients (imaginary part of the refractive index) at the measurement wavelengths. Fortunately, even the darkest-colored process liquids usually have very small imaginary part of the refractive index, and thus this effect does not have practical importance.

The mirror may not be perfect in its response to the refractive index. A mirror with a small number of layers will respond to the external refractive index slightly differently on different wavelengths, i.e. the decomposition (8.2) is only an approximation. This error can be compensated for by calibration with liquids having different refractive indices and known pH.

The response of the indicator dye may change as temperature changes. The absorption spectra of the indicator acid and base forms remain similar under moderate temperature changes as long as the molecule does not change its form. However, the balance reactions between the indicator dye and the solution to be measured do change, and thus similar proportions of the acid and base state indicate different pH values in different temperatures. Some experimental work is still required to give more information on this subject.

The dye may also be sensitive to other ions than hydronium and hydroxide. This error is analogous to the alkali error of the electrochemical pH sensors. As this error is dye specific, there are no general rules concerning this error.

It seems obvious that dye-related errors are difficult to detect, at least if they do not change the spectra of the acid and base states of the molecule. Optical errors seem to be easier to detect while they may not be possible to compensate for.

8.5.2 Light source errors

The LED light sources are not strictly monochromatic. The equations developed in section 8.1 assume monochromatic light sources. The signal for a light source having wider wavelength distribution has to be expressed as an integral instead of the simple product of (8.4):

$$m = \frac{1}{\lambda_2 - \lambda_1} \int_{\lambda_1}^{\lambda_2} r(\lambda) r_0(n) C(\lambda) e^{-a(\lambda,\mathrm{pH})c} d\lambda \tag{8.24}$$

Here the light source and detector spectral properties are embedded in $C(\lambda)$. If $a(\lambda, \mathrm{pH})$ changes little over the emission wavelength range $[\lambda_1, \lambda_2]$:

$$a(\lambda_a) \approx a(\lambda_b) \quad \text{for all} \quad \lambda_a, \lambda_b \in [\lambda_1, \lambda_2] \tag{8.25}$$

then the integral can be approximated by

$$m \approx r_0(n) e^{-a(\lambda,\mathrm{pH})c} \frac{1}{\lambda_2 - \lambda_1} \int_{\lambda_1}^{\lambda_2} r(\lambda) C(\lambda) d\lambda \tag{8.26}$$

In this case the integral is a constant and all formulae in section 8.1 are applicable.

The peak width of a typical LED is around 20 nm. The absorption peak width of a dye is over 100 nm, so there is an order of magnitude difference between the two. However, this does not necessarily guarantee sufficient accuracy, and at least with some dyes the emission peak width should preferably be narrower.

Another error source is the wavelength drift of the LEDs as the external temperature changes. The output intensity and wavelength of a LED are functions of external temperature and junction current. These phenomena are related to each other; high junction currents heat the junction and thus produce wavelength shift.

LED manufacturers usually give very little information on these effects. Some sources claim the peak wavelength shift due to temperature change is in the order of 0.1 nm/K [58]. This would give a few nanometers over the applicable temperature range. On the other hand, LED data sheets suggest that changing the drive current

from 0 to 50 % of the nominal current may change the output peak wavelength over 10 nm. [59]

Temperature coefficients of the LED intensity tend to be large. Again, few manufacturers give these figures in their data sheets but the values which are available seem to be around 0.1 – 1 %/K. In practice, this means there has to be some intensity feedback or other compensation system in the circuit.

One way to compensate for the wavelength shift would be to actively control the LED temperature. This would, however, require significant amounts of power, and add complexity to the circuit. Also, thermostating systems would not be compliant with the explosion proof requirements. A more practical approach would be to modulate the junction current to compensate for the junction temperature changes. Higher temperatures tend to increase the wavelength and decrease the intensity whereas higher currents increase intensity and decrease wavelength.

Another error associated with the LEDs is that the LED manufacturing process is not very even. LEDs from the same semiconductor wafer may have 20 nm variation in the peak emission wavelength [58]. Fortunately, this characteristics does not change during the life of the light source, and it is possible to pick the LEDs with suitable wavelength characteristics. Also, the calculation methods introduced in the previous chapter are not very sensitive to the absolute wavelength, wavelength shifts during use are much more difficult to cope with.

A simple way of avoiding the problems with changing wavelength is to place a narrowband filter in front of the LEDs. This way the emission spectrum is always in the range limited by the filter, and wavelength shifts in the LED are changed to intensity variations. There are, however, two disadvantages in this construction. First, the light output of the light sources is considerably lower with the filter as most of the output power is filtered away. Second, precise narrowband filters are relatively expensive interference filters, and their use will increase the system cost significantly.

An ideal solution would be to use light sources with small temperature coefficients and narrow emission bands. While there are no light sources available to fit these requirements along with the cost and power consumption requirements, it seems possible that resonant cavity LEDs (RCLED) may offer a good solution to this problem when they become commonly available.

Wavelength shift and wide wavelength range of the light source may cause measurement errors, at least in theory. The practical significance of these phenomena has not been fully analyzed as yet, and it is possible that even uncompensated LEDs perform well enough.

8.5.3 Detector drift

Silicon photodetectors are linear over a wide range of intensities. The reverse current flowing through a biased photodiode has two components, a temperature and bias dependent dark current and the photocurrent which is in linear relationship with the incident radiation falling on the detector.

The dark current is an exponential function of the temperature, and thus it can be significant in higher temperatures. Fortunately, the dark current can be compensated for by a simple subtraction (as explained in section 8.4).

There is another—more subtle—temperature dependency in silicon photodiodes; their gain (efficiency) depends on the temperature. The temperature coefficient in this case is not very large, in the order of 0.1 %/K. However, if there is a temperature difference between the two photodetectors, the measurement results will be multiplied by a constant.

Errors which affect all wavelengths equally will be interpreted as changes in the external refractive index. This is not very significant, as the external refractive index measurement is not very accurate in any case.

The easiest way to avoid errors associated with the thermal behaviour of the photodetectors is to keep the photodetectors at the same temperature. Also the transimpedance amplifiers should be kept at the same temperature to avoid errors caused by thermal coefficients of the feedback resistors. Naturally, the detection channels should be as similar to each other as possible in all respects.

9 Optical construction

As discussed in the previous chapter, the optical system has to be efficient, easy to assemble, rugged, and inexpensive. The construction described in this chapter fulfills all these requirements (see section 8.2) to a sufficient extent.

Obviously, the solution described here is a compromise between several desired properties. One of the most important aspects of this optical system is that it does not have to be an imaging system, as only transferring the light is important. This enables the use of some light conduit type optical components in the system. Optical aberrations are not important and thus the use of multi-component lens combinations (e.g., achromats) can be avoided.

9.1 LIGHT FOCUSING

As noted in section 8.2, it is desirable to have the light hit the sensor surface as a wide pencil of light normal to the surface. With a point source this can be easily done by a single lens if the light source is placed at the focal point (figure 9.1).

In the case of a transmissive measurement, the light is easily recollected with another lens with the detector in its focal point (figure 9.2).

Clearly, the left and right hand sides of figure 9.2 are mirror images of each other. If the transmissive filter is replaced by a mirror, the two lenses fold on top of each other. In this way only one lens is required to perform the focusing for both the emitter and the detector.

In practice, both the emitter and the detector have a finite surface area and they cannot thus be at the same point. Fortunately, the optical system does not require the components to be exactly on the axis. Figure 9.3 illustrates the image forming for an off-axis point. However, if the distance between the lens and the reflecting surface is long, or if the points are far away from the optical axis, a significant part of the light is lost. Figure 9.4 illustrates this situation.

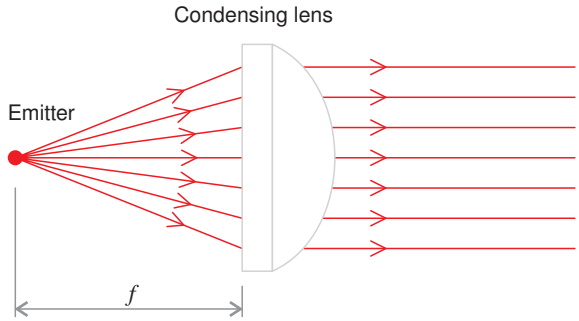


Figure 9.1 Making a wide beam from a point source with a single lens.

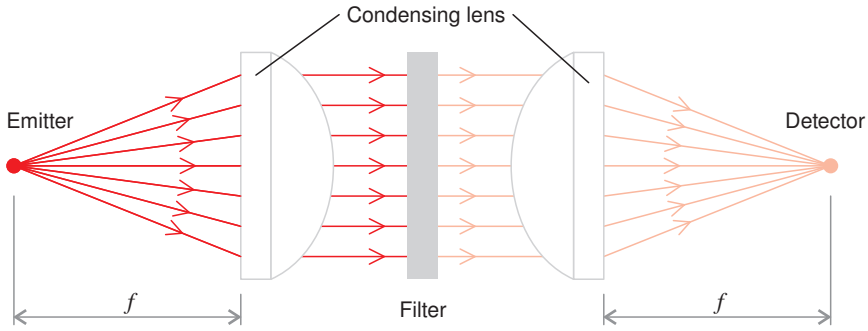


Figure 9.2 Measurement arrangement for a transmissive measurement.

Evidently, if the detector and emitter are placed symmetrically around the optical axis, an image of the emitter is projected onto the detector. Depending on the aperture of the lens and intensity distribution of the emitter this optical system is quite efficient. As the optical system does not have to be free of image aberrations, inexpensive and efficient aspheric condenser lenses may be used. Aspheric condensers often have F numbers well below unity (i.e. their focal length is smaller than their diameter).

The lens system may be considered as a two-lens system which can then be reduced into a single thick lens. Thus moving the emitter away from the lens will move the detector image nearer to the lens and make it smaller and vice versa. However, aspheric condensers are usually designed to applications where the light source is in

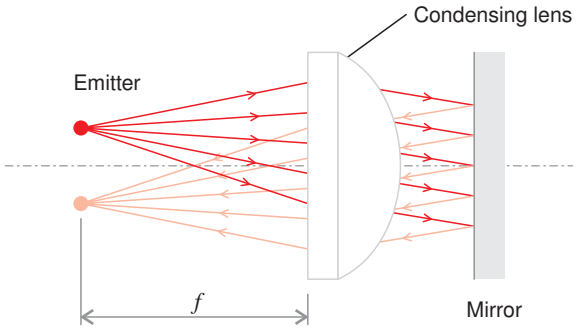


Figure 9.3 Image of a non-axial point.

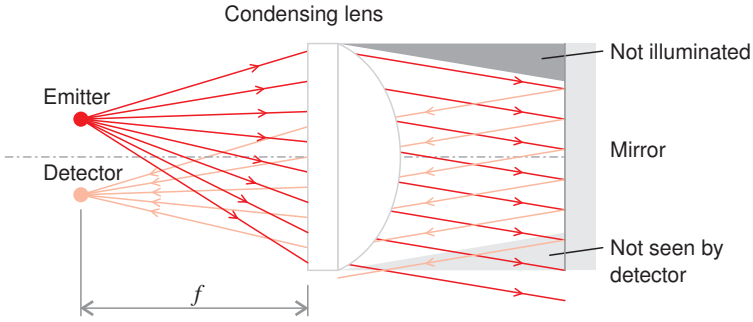


Figure 9.4 Losses due to non-axial emitters and detectors

the focal point, and thus the imaging performance of this lens system may be poor with other than 1:1 imaging applications. Also, placing the detector and the emitter at different distances would be mechanically more difficult than having them at the same distance.

There are several possible alignment errors which may occur during assembly. The most significant error due to component tolerances is the focal length tolerance of the lens. However, the system described here is not very sensitive to this error. In case either the focal length of the lens or the lens to emitter / detector distance is wrong, the image of the emitter is not focused axially precisely onto the detector, i.e. the image is not sharp. As the optics does not have to be imaging, the only measurable change

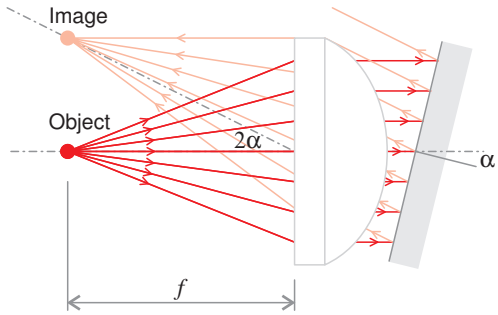


Figure 9.5 Image position error due to angular mounting error of the reflecting surface.

is a slight loss in intensity.

Radial errors in the system have more severe consequences. If either the lens is radially misaligned or the emitter and detector are not symmetric about the optical axis, a large part of the light does not hit the detector. The assembling tolerances allowed in the radial direction depend on the size of the emitter and the detector. The displacements have to be small compared to the emitter and detector diameter.

Angular mounting errors of the lens do not change the system performance significantly. An ideal lens would be completely immune to these errors as the rays going through the center of a lens do maintain their direction in geometric optics. In practice, the possible angular mounting errors are so small that even with the non-ideal aspheric lenses the resulting errors are small.

Angular mounting errors of the sensor surface are not allowed. If the mirror surface is not normal to the optical axis, the mirror action will double the angular error of the returning ray (figure 9.5). Thus, already rather small angular errors will produce considerable intensity loss.

As mentioned above, the distance between the lens and the sensor surface should be minimized. However, small variations in this distance do not alter the system performance significantly. Unfortunately, there is a disadvantage associated with this behavior; also the light reflected from the non-sensing back surface of the sensor element is focused on the detector.

There are other unwanted reflections in the system, as the lens surfaces also reflect some light. The stray light reflected from the lens surfaces, however, is scattered over a large area whereas the back reflection of the sensor element is focused on the sensing element (figure 9.6).

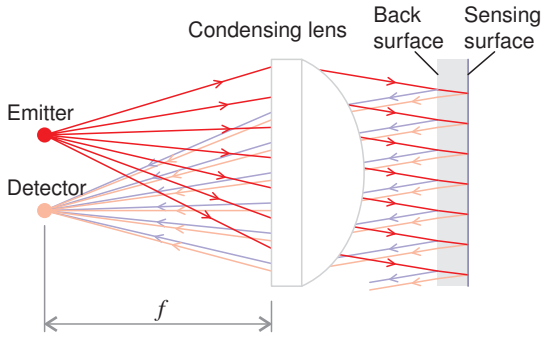


Figure 9.6 Stray reflection from the back side of the sensor element.

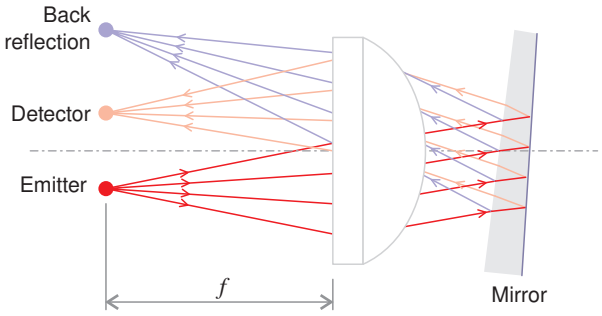


Figure 9.7 Eliminating stray reflections by tilting the back surface of the sensor element.

There are at least two possible methods to avoid the back reflection. First, the back surface of the sensor element may be made non-parallel with the sensor surface. In this way the back reflection is focused away from the detector (figure 9.7). Or, the back surface may be coated with an anti-reflective (AR) coating to reduce the reflection.

While the non-parallel approach to the sensor element has certain advantages—such as efficiency and the possibility of placing a reference photodetector into the point where the back reflection is focused—the optical construction becomes difficult to manufacture. As long as all elements have rotational symmetry and at least one plane surface normal to the optical axis, they are relatively easy to mount with high axial and angular accuracy. In this case both surfaces are non-normal to the optical

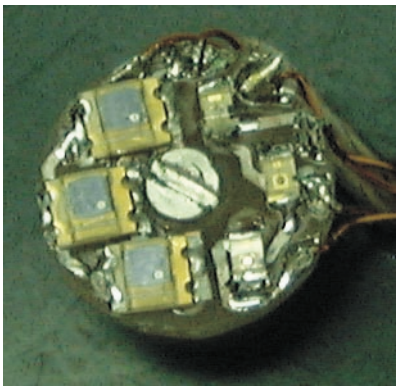


Figure 9.8 First prototype LED and photodiode arrangement. LEDs are on the right side, photodiodes on the left.

axis. Also, manufacturing the wedge-shaped sensor elements is difficult, e.g., film dipping and precise element cutting are difficult.

It should be noted that the error resulting from the back side reflection is additive and constant. It can be removed by subtracting a fixed value from the reflectances. Especially if the back side is AR coated, secondary reflections (mirror – back side – mirror) are unimportant. Also, the reflection from the coated back side is only in the order of 1 % of the total reflection, and so the error introduced into equation (8.4) is small as a first approximation.

9.2 REFERENCE MEASUREMENT

The first prototype of the instrument incorporated a very simple construction with three LEDs and three photodiodes arranged in a ring form (figure 9.8). The small circuit board was placed at the focal plane of the mirror.

While the optical construction turned out to be reasonably efficient despite the low directivity of the surface mount LEDs, initial measurement results drifted significantly even with constant current drive. The drift turned out to be mostly due to the large temperature coefficient of the LED intensity.

Obviously, a reference measurement is required in the system. There are several possible reference schemes which can be used in the system.

The most reliable way of measuring the reference would be to split the measuring thin film stack into two parts; one part with the indicator dye and one without. In this way all other effects (e.g., mirror reflection changes, even mirror damages) could be cancelled out.

However, there are two main obstacles associated with this approach. First, producing the membrane with two different regions is considerably more difficult than producing a uniform film stack. The variable speed dipping methods outlined in section 7.2 might be utilized in this context. Or, with highly precise dipping equipment, it might be possible to dip the substrate from opposite directions to have one half of the substrate coated with a color layer and another half with the same layer without color. However, either of these methods would add considerable complexity to the layer forming process.

Another problem is related to the measurement optics. With the optical construction outlined above the reflection is averaged over the whole sensing surface. Some modification is required to make it possible to measure different parts of the surface separately.

One solution would be to make the back of the sensor element slightly prismatic. This would enable the use of two different regions in the sensing surface. The principle is depicted in figure 9.9.

In practice, using only two regions may lead to large errors if either of the regions coats or gets damaged. The solution to this is to divide the surface into finer strips, which then have their own prisms.

While this reference system would undoubtedly be accurate, it adds a great deal of mechanical complexity to the system. In addition to the manufacturing challenges of the striped sensor element, making and mounting of the prisms required in this approach increase manufacturing costs significantly.

Probably the most widely used optical reference sampling system is to use a semi-transparent mirror to take a sample of the light. Figure 9.10 shows one possible configuration. Again, the use of a semi-transparent mirror requires several additional non-axial components to the system. A more compact application of this would be to use the reflection from the back surface of the sensor substrate, i.e. using a configuration similar to that depicted in figure 9.7.

All the reference systems above—and the discussion about detectors and emitters—have considered the situation with only one relatively small emitting surface. In practice, there are three different LED junctions which are in separate packages. So, along with the reference sampling problem, there remains the question of combining the light output from three separate sources to one source with relatively uniform radiation characteristics.

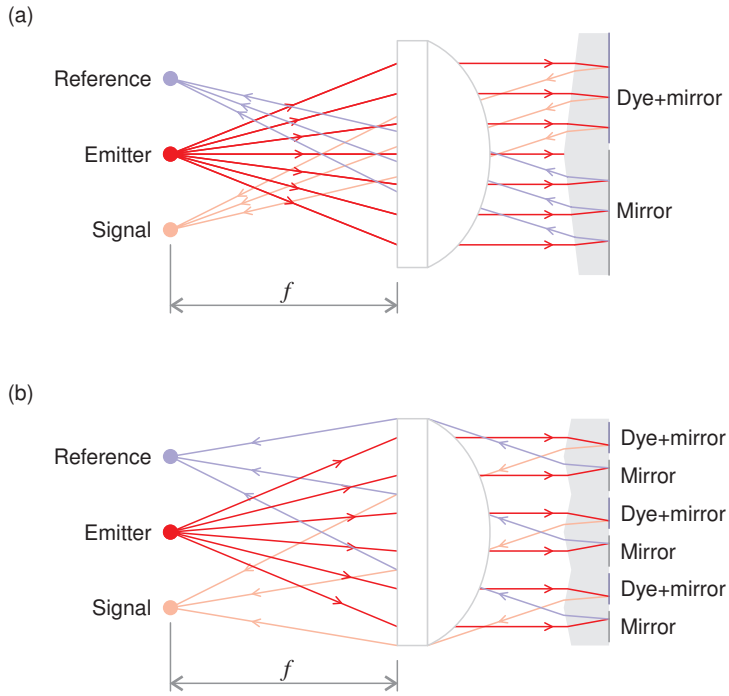


Figure 9.9 Possible configurations for two-region reference measurement (a) with two regions and (b) with a large number of regions.

Combining the light from three different light sources can be made with a long and narrow glass rod (or fiber). Such light mixing rods have been used in industrial instruments in combining light output from several LED light sources [60]. With a slight modification the same light mixing rod can be used in obtaining the reference signal (figure 9.11).

The light conducting rod is a mechanically simple element to mount. While it does not have any large plane surfaces, it is easy to align axially and radially due to its long and narrow cylindrical shape. Also, the glass rod is rather inexpensive, and there are several possible manufacturing methods in making the required groove to the rod.

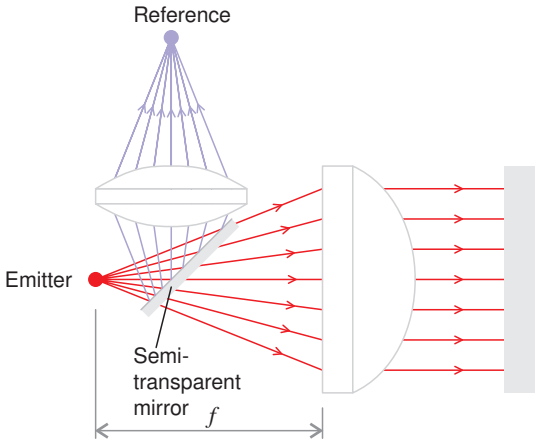


Figure 9.10 Intensity reference sampling by using a semitransparent mirror.

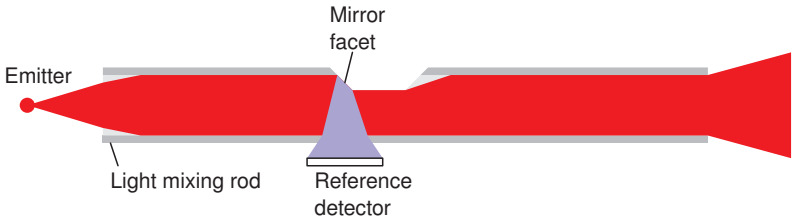


Figure 9.11 Reference measurement by using a light mixing rod

9.3 CYLINDRICAL ROD OPTICS

While the concept of the reference sampling light conduit rod (or light pipe) is intuitively clear, there are some rather interesting—and even counterintuitive—phenomena in how light advances in the rod. As an optical component the light-conducting rod is catadioptric, rays going through the system are both refracted and reflected.

Most optics textbooks (for example, [49, p.170]) do handle the theory of multimode optical fibers, which is in close relation with the light pipes. However, the emphasis is usually on the dispersion properties of the fibers resulting from the multiple reflections. In that treatment it is assumed that each ray undergoes a large number

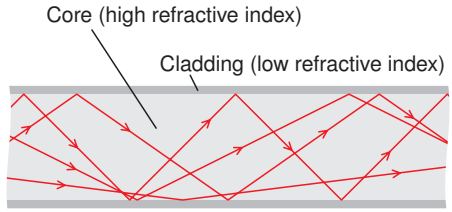


Figure 9.12 Light passing in a clad rod.

of reflections.

With short light pipes the number of reflections is small, and the resulting intensity and angle distribution at different positions of a light pipe have some interesting properties which do affect the design of structures such as the construction of figure 9.11.

9.3.1 Basic principle

The basic element in rod optics is a clad rod (figure 9.12). The rod core is made of high index glass, and the cladding has considerably lower refractive index. If a ray of light strikes the interface between the core and the cladding in a small angle (near to the tangent), it is reflected by total internal reflection. This critical angle (φ_c) can be calculated from Snell's law:

$$\sin \varphi_c = \frac{n_{\text{clad}}}{n_{\text{core}}} \tag{9.1}$$

where n_{clad} is the refractive index of the outer layer and n_{core} that of the rod core. If the angle of incidence is larger than this value, light enters the cladding and strikes the cladding/air interface. This interface has its own critical angle, and it is possible that the light undergoes total internal reflection at this interface. In many applications it may be beneficial to remove the rays which would advance in the cladding, as they are in larger angle to the axis of the cylinder. This can be accomplished by either cutting the cladding away for a short segment of the rod or by coating the clad rod by some very absorbing material.

Naturally, the rod would function without the cladding, as well. In practice, mounting the non-clad rod would introduce points and surfaces where the advancing light can escape the rod. This would change the angular and intensity distribution in the rod. The main purpose of the cladding is to form a continuous and well-defined interface for the total internal reflections.

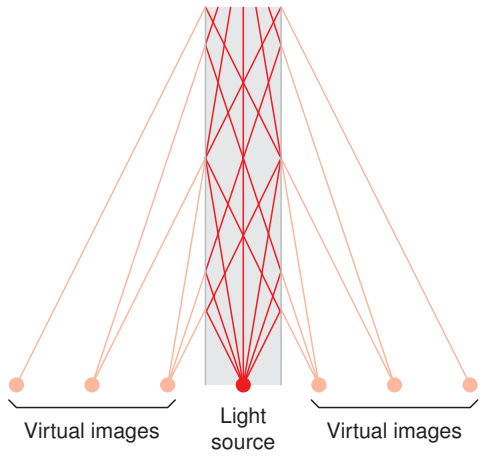


Figure 9.13 Light in a two-dimensional light conduit.

The light-mixing property of the rod can illustrated by the two-dimensional situation, where several mirror images of the light source can be used to describe the multiple reflections in the rod (figure 9.13) [60].

While this approach gives a qualitative idea of how the rod works, the mirror principle cannot easily be applied to the three-dimensional problem. Instead, the rays have to be followed through the rod with ray tracing techniques.

Calculating the intersection of a line (vector) and a cylinder is a relatively simple geometrical problem. However, intensity distribution simulation requires a very large number of rays, so the general ray tracing methods applied segment by segment in a system with a large number of reflections lead to long processing times.

Fortunately, the cylindrical symmetry of the problem provides some opportunities to simplifications in the calculation. Figure 9.14 shows the passage of one ray through the system in two different projections (radial and axial).

The passage of a ray of light can be handled separately in the z direction (along the axis) and in the plane perpendicular to the axis. As all rays should be advancing in the z direction, all ray vectors are normalized so that their z component is always unity. In this way the ray directions can be described with only two parameters:

$$\vec{v} = \begin{pmatrix} v_x \\ v_y \\ 1 \end{pmatrix} \tag{9.2}$$

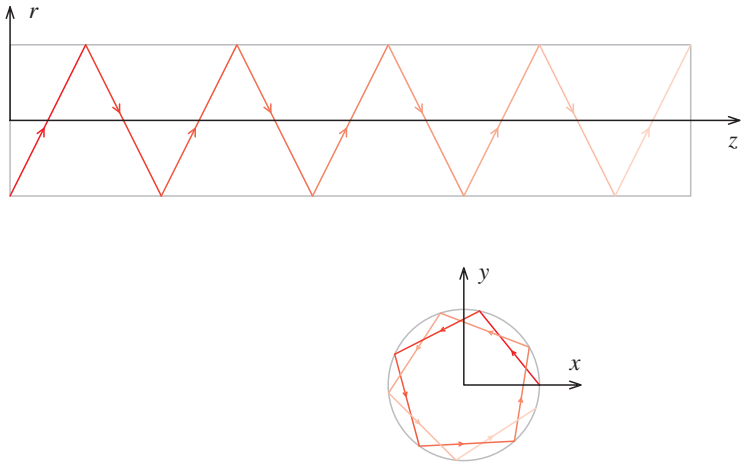


Figure 9.14 A ray of light going through a clad rod.

As the angle of incidence and angle of reflection referred to the normal are equal, the advancing ray draws secants to the circle in the xy-plane. These secants are equidistant from the center of the circle, and at each reflection the secant is rotated by a fixed angle which depends on its distance from the center.

Figure 9.15 shows a two-dimensional ray trace through two full secants. The ray starts from the reflection point $\vec{s}_0$, continues to $\vec{s}_1$, undergoes yet another reflection, and ends up at point $\vec{s}_2$. Due to the law of reflection, the angles δ at $\vec{s}_1$ are equal. As the two triangles ($\vec{s}_0 O \vec{s}_1$ and $\vec{s}_1 O \vec{s}_2$) are isosceles, the angles at $\vec{s}_0$ and $\vec{s}_2$ have to be δ, as well.

If the angle of the vector $\vec{s}_0$ is α_0, then the angle α_1 of the vector $\vec{s}_1$ can be calculated from the triangle $\vec{s}_0 O \vec{s}_1$:

$$\delta + (\alpha_0 - \alpha_1) + \delta = \pi \tag{9.3}$$

$$\alpha_1 = \alpha_0 + 2\delta - \pi \tag{9.4}$$

So, at each reflection the triangle is rotated by γ:

$$\gamma = \alpha_1 - \alpha_0 = 2\delta - \pi \tag{9.5}$$

The length of one secant is:

$$l_p = 2R \cos \delta \tag{9.6}$$

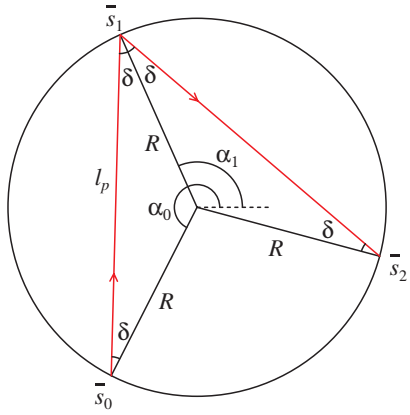


Figure 9.15 Passage of light through two secants.

Figure 9.16 depicts the situation when a ray of light starts from point $\vec{p}_0$ within the cylinder and advances to direction $\vec{v}_0$. To make the calculations simpler, it is useful to make the ray start from the perimeter of the circle. To this end, distance l_0 is defined:

$$\vec{s}_0 + l_0\vec{v}_0 = \vec{p}_0 \tag{9.7}$$

The length l_0 can be solved from this by solving $\vec{s}_0$ and squaring the result. As $\vec{s}_0 \cdot \vec{s}_0 = R^2$, l_0 can be solved from this second degree equation:

$$l_0 = \frac{\vec{p}_0 \cdot \vec{v}_0 \pm \sqrt{(\vec{p}_0 \cdot \vec{v}_0)^2 + (R^2 - p_0^2)v_0^2}}{v_0^2} \tag{9.8}$$

Only the positive solution is valid, as $l_0 \geq 0$. The negative solution represents the distance from point $\vec{s}_1$ and is thus invalid in this context.

Point $\vec{s}_1$ (and hence angle α_0) can now be solved from equation (9.7). If the angle of vector $\vec{v}_0$ is β_0, then the angles δ and γ are by geometry:

$$\delta = \beta_0 - \alpha_0 + \pi \tag{9.9}$$

$$\gamma = 2\delta - \pi = 2\,(\beta_0 - \alpha_0) + \pi \tag{9.10}$$

The next question is how many reflections a single ray encounters in its way through the cylinder. If the length (in z direction) of the cylinder is h, then the number

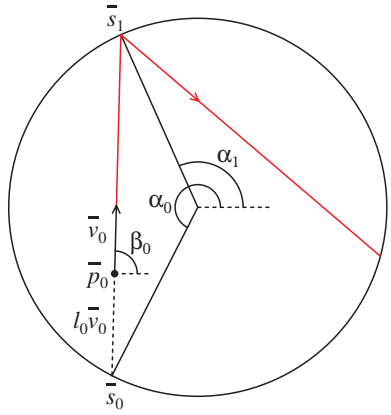


Figure 9.16 A ray of light entering the clad rod.

of reflections is:

$$n = \text{floor}\left(\frac{h + l_0}{l_p}\right) \tag{9.11}$$

This is based on the normalization of the ray vector so that its z component is unity. The cylinder has been expanded downwards by l_0 so that the ray starts from the perimeter (point $\vec{s}_0$).

The n^{th} reflection will occur at point $\vec{s}_n$ for which the angle α_n is (figure 9.17):

$$\alpha_n = \alpha_0 + n\gamma \tag{9.12}$$

Similarly, the new direction β_n of the ray ($\vec{v}_n$) is rotated by $n\gamma$:

$$\beta_n = \beta_0 + n\gamma \tag{9.13}$$

The exit point is then:

$$p_{\text{exit}} = \vec{s}_n + (h + l_0 - nl_p)\vec{v}_n \tag{9.14}$$

By this method the number of calculations does not depend on the number of reflections. The limitations come from the calculation resolution used in the angular rotation.

There are some fairly obvious rules the rays will obey. First, as the angle between the optical axis and the ray is always preserved in a reflection, the light mixing rod

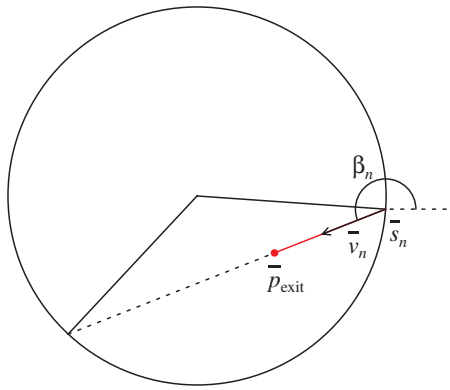


Figure 9.17 A ray of light at the exit end of the rod.

does not completely mix the angular distribution. The angle is mixed in the *xy*-plane but if there are no paraxial rays entering the rod, no paraxial rays will exit the rod. One side effect from the angular rotation about the optical axis is that the polarization state of the light is lost in the rod.

The position mixing seems to be more complete. Even if there is no light entering the rod at certain position, there may be light exiting it at any position. With the extreme case of a pointlike light source on the entrance surface, the output intensity distribution will still be even in a long rod.

It is interesting to note that this is a truly chaotic process. While the process is strictly deterministic, minute changes in the input parameters—position or angle—will cause large changes in the output position.

9.3.2 Simulations

To illustrate the passage of light in the rod, figure 9.18 shows the intensity and angular distribution at the end of 1.6 mm diameter rods with different lengths. The angular distribution graphs are equivalent to the intensity distribution projected on a screen far away from the rod end. Light entering the rod in these simulations has even intensity distribution but is directed as shown in figure 9.19.

The slight graininess of the images is a result of the simulation procedure; the rays entering the image are chosen randomly from a predefined distribution. A total of 10 million rays have been used to trace each of the 256 x 256 pixel images. Also, for

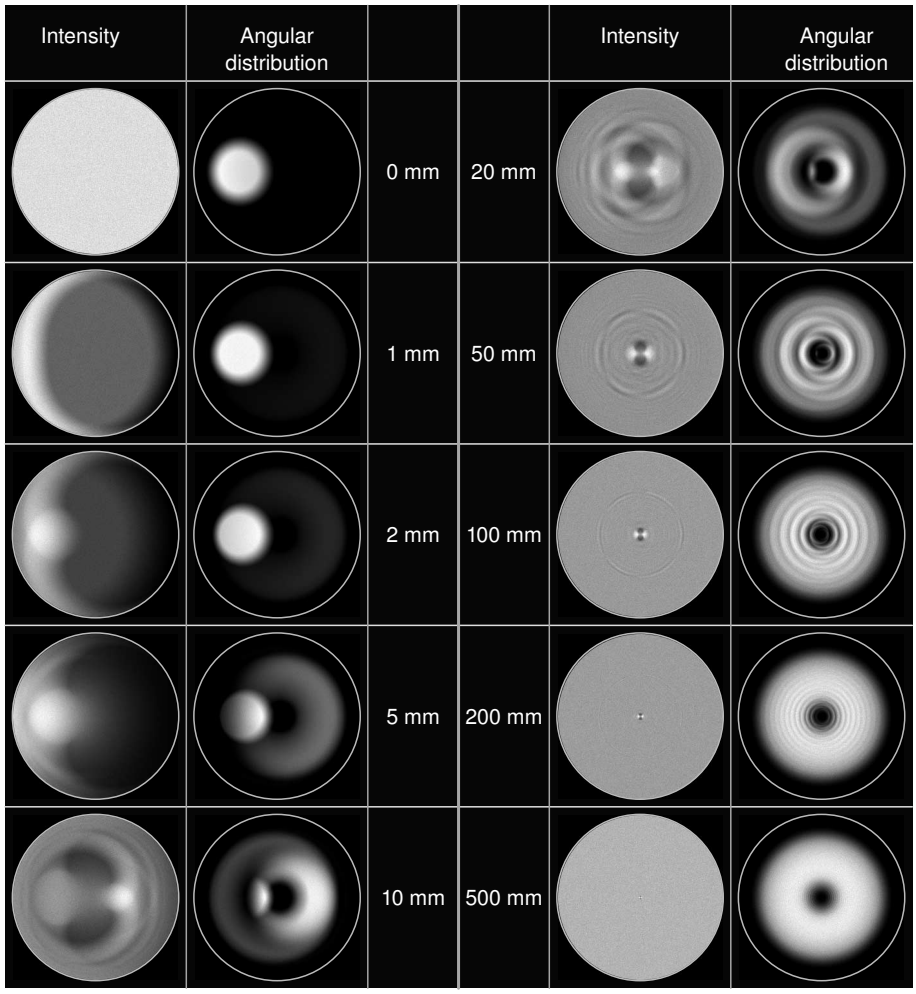


Figure 9.18 Intensity and angular distributions at the end of a 1.6 mm rod with different lengths.

Figure 9.19 Light input geometry used in figure 9.18

reproduction reasons the intensity of each image is normalized so that the maximal dynamic range is used for each image. For instance, the intensity in the center of the intensity diagram at $z = 1$ mm should be equal to that at $z = 0$.

From the simulation results it seems that the intensity mixing is much faster than the angular smoothing. While the intensity distribution at $z = 50$ mm is rather even, the angular distribution at the same position can hardly be called smooth.

The long rod intensity distribution ($z = 500$ mm) is practically even. Also the angular distribution is as predicted; the original distribution rotated around the optical axis. A closer inspection of the angular distribution will, however, still show considerable fringes. While the exact figures depend on the incident angle, it seems that as a rough approximation the angular distribution is not smooth if the length/diameter ratio is less than 100:1.

A rather interesting feature is the "eye" in the intensity distribution which forms in the middle of the rod. This phenomenon is caused by the radial rays which form a quasi-stationary pattern. As the number of reflections increases, the number of near-normal rays which have significantly rotated from the original direction grows, and the eye becomes smaller.

Figure 9.20 shows actual intensity and angular distribution photographs from a 160 mm long 1.6 mm diameter rod. The features explained above are visible, even though the central eye is not very clear, due to limitations in the photographic system (mainly due to the non-zero image plane depth).

9.3.3 Facet mirrors

A mirror surface can be embedded into the glass rod by making a groove to the side of the rod. The facet of this groove will act as a prism surface, nearly paraxial light will undergo total internal reflection from the prism surface. Figure 9.21 shows an illustration of a light conducting rod with the mirror groove.

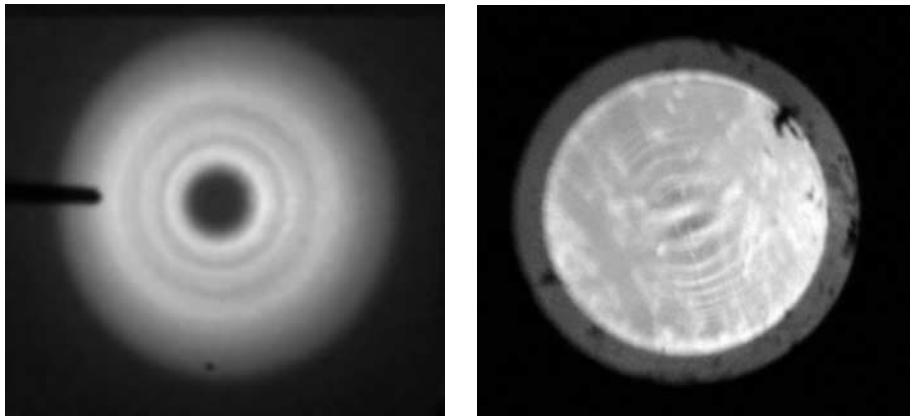


Figure 9.20 Intensity (right) and angular distribution (left) photographed from a 1.6 mm diameter 200 mm long rod. The dark line on the angular image is the end of the rod, black spots on the intensity photo are dust particles.

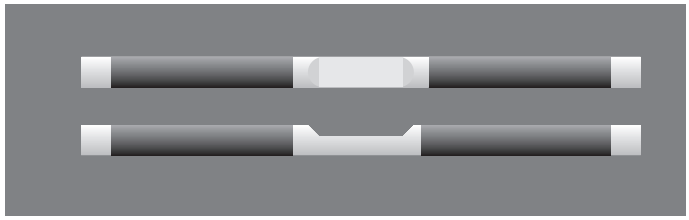


Figure 9.21 A clad rod with light-reflecting groove and absorbing coating.

It should be noted that in this surface the total internal reflection critical angle is determined by the refractive indices of the rod core material and air, as there is no cladding on the mirror surface. Part of the light striking the prism surface may be lost in refraction. Also, some light may be lost through the part with D-cross-section between the two slant facets. The walls of the rod outside the groove area are coated with black coating to remove rays propagating in the cladding.

The results obtained in the previous section suggest that any disturbances in the intensity distribution will be smoothed away in a rather short length of the rod. Simulation results confirm this is indeed the case even when the groove cuts the intensity distribution very significantly. Figure 9.22 depicts some simulation results from a shortish rod with the mirror facet.

An interesting feature in the intensity distribution projected on the screen S is that it is considerably narrower than the rod (all intensity images have side length of 1.6 mm). This is because the rod walls act as a cylindrical lens.

In these simulations the Fresnel equations have been taken into account in refractions. In each refraction the ray has randomly been either refracted or reflected so that the random probabilities represent the reflection coefficients given by the Fresnel equations to randomly polarized light.

The depth of the groove has to be chosen so that the proportion of light reflected to the sensor to that remaining in the rod is suitable. Ideally, the amount of light on the reference detector should be equal to that striking the measurement detector. If the system were lossless, then half of the light should be deflected to the detector. In practice, the optimal proportion is much smaller, as there are significant losses in the optical system and sensing element.

The efficiency of the clad rod optics is good. Actual efficiency figures depend on the rod dimensions, but typically the proportion of rays lost in the facet mirrors is below 10 % of the total rays. In this context lost rays are those that do not hit the reference detector or reach the exit end of the rod. Equal losses are encountered in the entrance and exit ends of the rods.

In theory, a simple V-groove would be sufficient to deflect the light. Also, V-shape would eliminate the D-shaped cross-section which is responsible for some losses. However, a V-groove would make the rod extremely fragile, and the groove shape shown above is more practical.

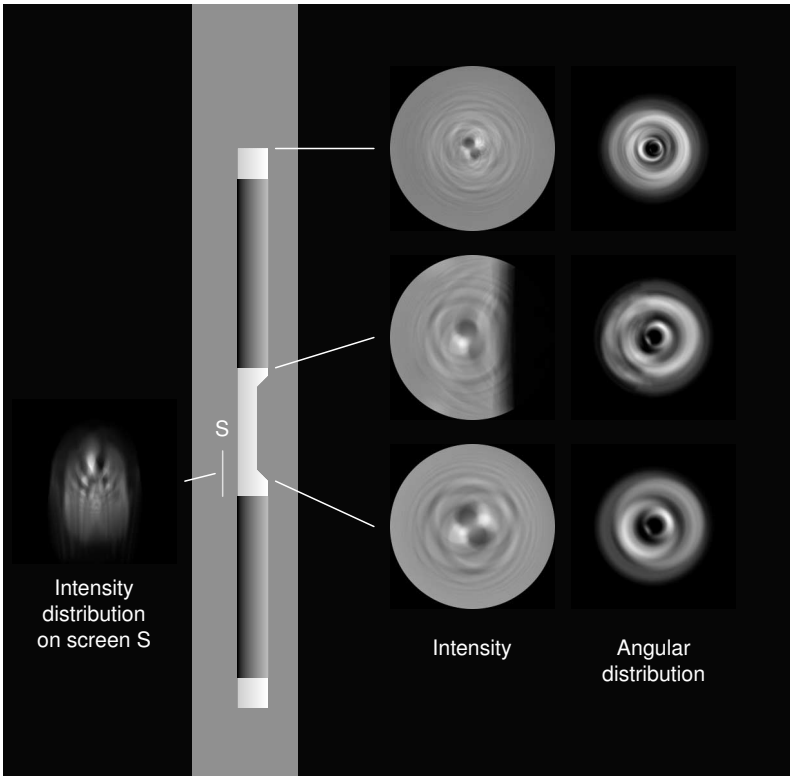


Figure 9.22 Intensity and angular distribution of light in a light conducting rod with a mirror groove. Light reflected from the one of the prism surfaces to the screen S is shown on the left.

10 Prototype construction

A prototype was built for evaluation purposes. While this prototype is primarily intended to be used in laboratory evaluations, it has been designed so that it can also be used in the first real world process evaluations.

From the industrial point of view this prototype is far too large to be manufactured economically. However, there is a lot of room for miniatyrization, and the actual optical and electronic components occupy only a small fraction of the sensor volume.

In the prototype all digital signal processing takes place outside of the sensor. In its current incarnation the sensor has no own intelligence, an external microcontroller is used in measurement control and signal processing. However, similar functions can very easily be integrated into the measurement electronics inside the sensor body.

10.1 SENSOR BODY

Figure 10.1 shows the prototype sensor from outside. The sensor body has been modified from a commercially available process instrument, K–Patents refractometer PR–03. Materials used outside the sensor are polished stainless steel (parts wetted in the process) and PTFE-coated aluminium. This instrument is originally intended for food industry and thus it has been made as easy to clean as possible—an useful feature with the pH measurement instrument prototype, as well.

The sensor surface is visible on the top of the instrument. The narrow white line surrounding the sensor surface is a PTFE gasket which seals the sensor head. Another PTFE gasket is between the stainless steel and the aluminium part. This gasket limits thermal flow from the possibly hot wetted parts to the electronics housing.

There is an o-ring seal between two body halves (the border between these halves is just visible slightly above the ruler in figure 10.1). In principle, this seal makes the

Figure 10.1 The prototype of the pH measurement instrument.

body watertight and enables cleaning with a spray of water. However, the electrical connections in the prototype are not watertight at this phase of the project.

10.2 SENSOR ELEMENT

The sensor element used in the instrument has a total of five thin film layers (figure 10.2). Two bottom layers are similar color indicator layers doped with BPB indicator dye. Two layers are used instead of one to give higher absorption, and as these layers are drawn from the same sol, there should not be any significant reflection from the interface of these layers, and thus they can be treated as one layer.

A three-layer mirror of BST/BS/BST sols is used on top of the indicator layer. The reflection spectrum of this mirror is depicted in figure 10.3.

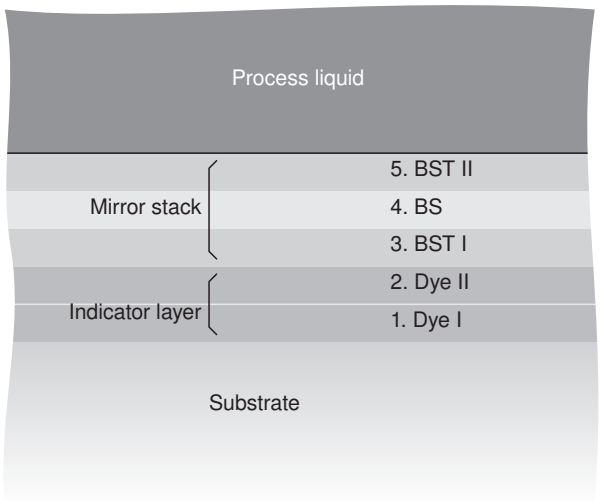


Figure 10.2 Thin film structure used in the prototype.

In addition to these sensor elements, mirror elements without dye layers were manufactured for reference purposes. The mirror reflection spectrum has been determined by transmission measurement from a non-dyed element. To ensure similarity between the elements with and without dye, the layers have been deposited simultaneously.

Figure 10.4 shows two substrates, one with and one without dye. The substrate imaged on the left has a dye layer and the letters "pH" have been written on it with drops of base to illustrate the color change. The letters are visible only from the non-coated side of the substrate because only then the reflected light goes through the indicator layer.

In order to reduce unwanted back reflections (section 9.1) the films are deposited on a glass substrate with an anti-reflection coating on one side. The glass material itself is ordinary window glass. Borosilicate glass would be more durable against thermal and mechanical shocks, but as it is more difficult to handle (especially cut), a softer glass was chosen. Thickness of the substrate is 2 mm.

During the film dipping process the back side (AR coated) is protected by a plastic film which is removed before drying.

The film layers are originally deposited on a $100 \times 100\,\mathrm{mm}^2$ glass plate. Approx-

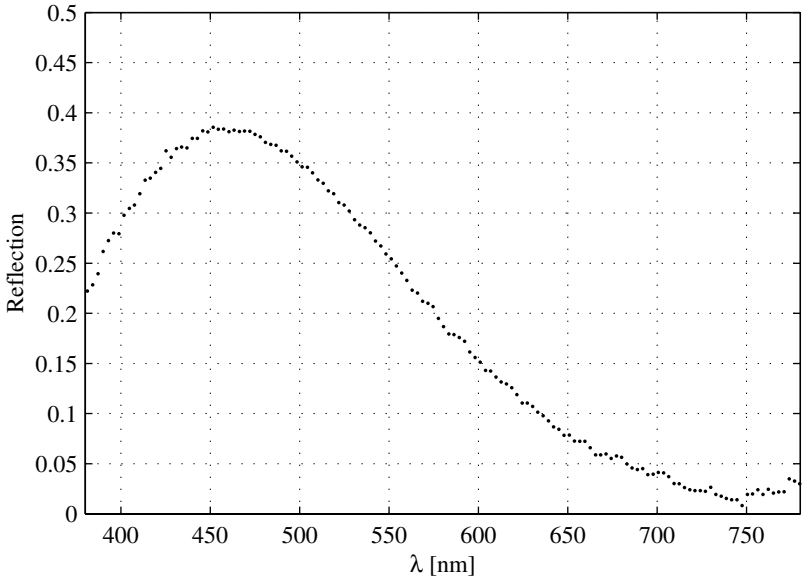


Figure 10.3 Reflection spectrum of the sensing element three-layer mirror.

imately $50 \times 90\,\mathrm{mm}^2$ of this is coated by a high-quality film. The substrate is then cut into small square pieces (around 20 mm square) and then ground with a diamond tool to a ø16 mm disc with a small wedge on the coated side (figure 10.5).

10.3 OPTICAL CONSTRUCTION

The optical construction of the instrument is shown in figure 10.6. A molded aspheric condenser lens with ø15 mm and $f = 12$ mm (Melles Griot catalogue number 01 LAG 001) is used as the focusing element. This lens has been chosen because of its good light collection properties.

The detector/emitter assembly support structure is milled from aluminium. There are two light-conducting rods in the system, one for emitted light and one for the detector. Figure 10.7 shows the assembly. The assembly (and the light conducting rods) has been coated black to reduce stray reflections. In order to reduce assembly stress the rods have been attached with silicone.

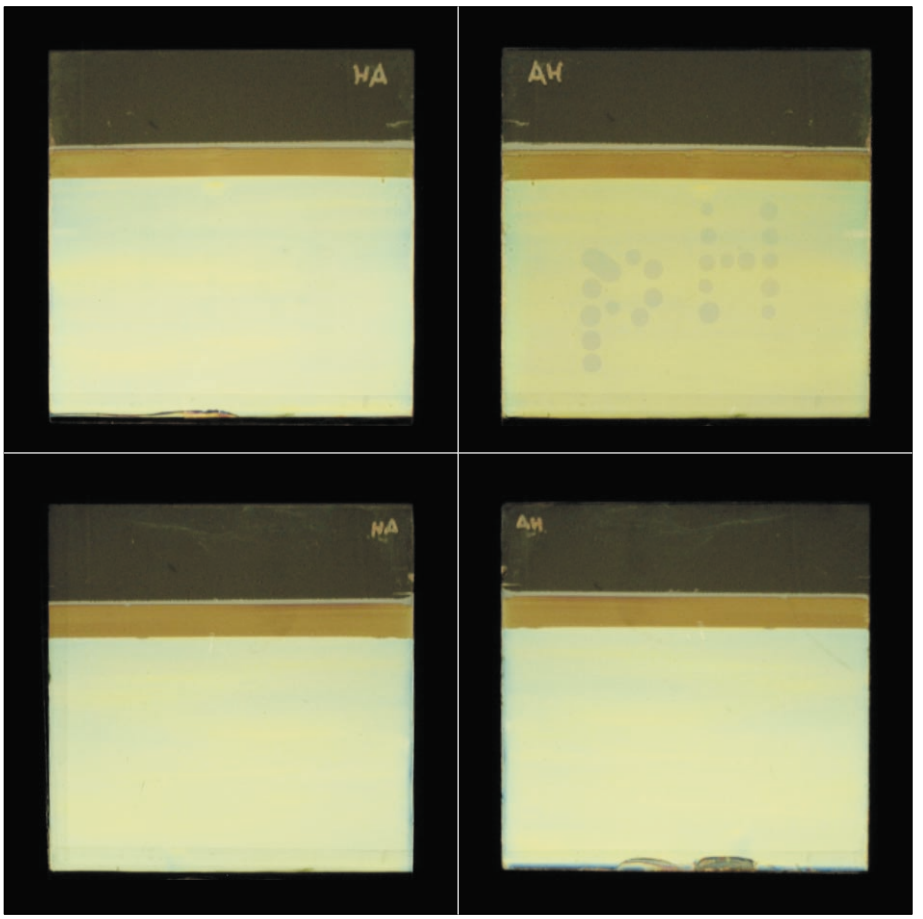


Figure 10.4 Substrates with similar mirrors photographed from the coating (left) and back side (right). Upper substrate has a mirror stack with pH-sensitive dye layer. The letters "pH" are written into the layer by drops of base.

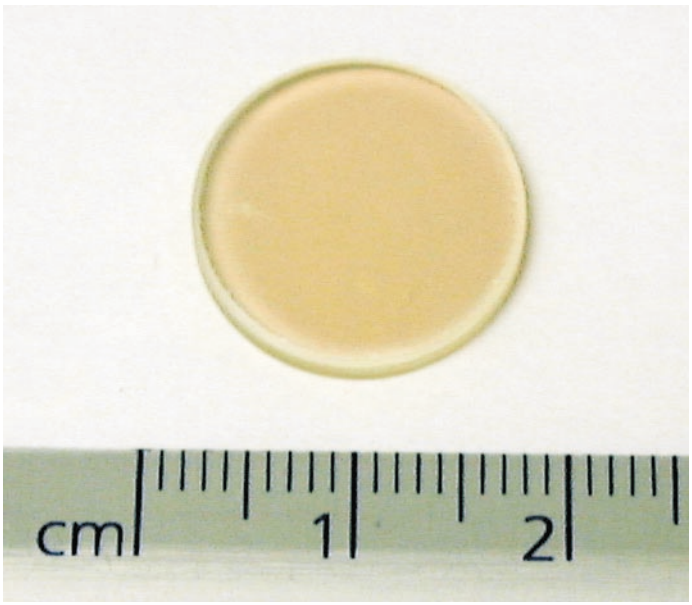


Figure 10.5 A sensor element in its final shape.

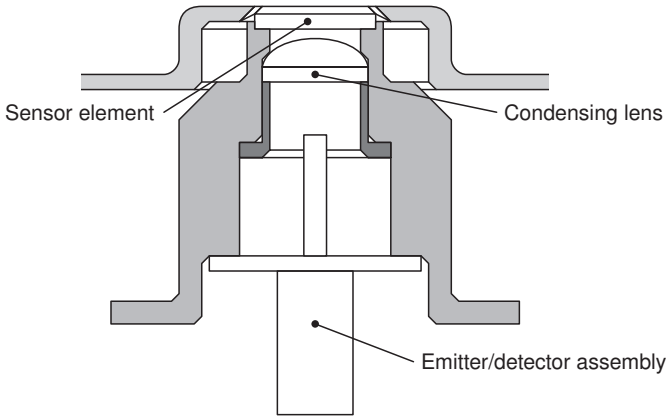


Figure 10.6 Optical construction of the instrument.

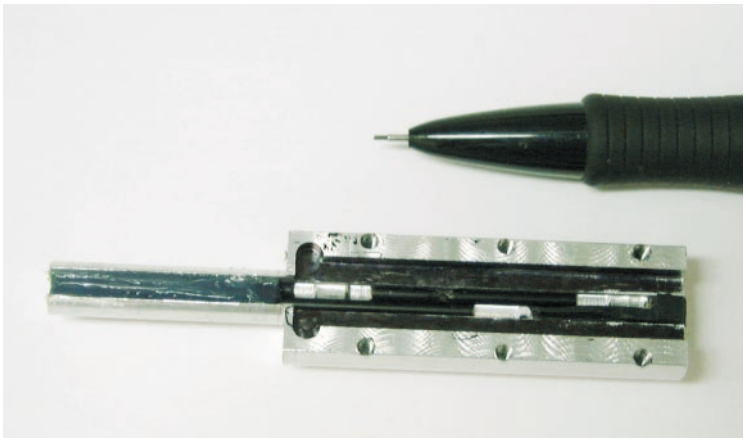


Figure 10.7 Emitter/detector assembly of the instrument. The pencil shown in the figure to give an indication of the size.

In this assembly the facet mirrors developed in section (9.3.3) are used in several places. The detector rod (figure 10.8) has a single mirror in its end to reflect the light to the detector which is mounted to the side of the rod. In addition to the reference signal mirror the emitter rod has a triple mirror in its end to reflect the light from three different LEDs to the rod. This way all emitters and detectors can be mounted on a single printed circuit board (PCB).

The emitter and detector light-conducting rods of the prototype were ground manually. All mirror and end surfaces were first ground with a customized thin-blade diamond saw near to their final shape. Scratches left by the diamond saw were then ground away manually with cerium oxide polish and bronze tools. Finally, the surfaces were carefully polished with fine cerium oxide paste and a puffing wheel. The black surfaces have been made with black paint.

The manual procedures in manufacturing the components do not yet produce optimal results. While the mirror surfaces do perform well enough, they still scatter some light due to suboptimal polishing results. In mass production the rods can be ground and polished with special tools to improve the result. It may also be possible to coat the mirror surfaces with metal to improve reflection. However, the small scattering does not affect the general results of the evaluation of system performance.

When the optical system is correctly aligned, the intensity distribution on the sen-

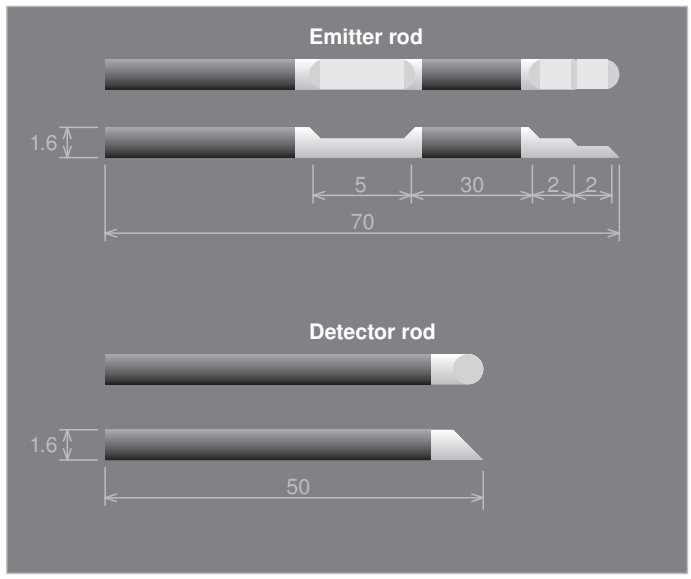


Figure 10.8 Dimensions of the emitter and detector rods.

sor element should be equivalent to the angular distribution of the rays emerging from the emitter rod end. To check this, a piece of "invisible" office tape was attached onto the sensor element. The tape has a diffuse surface which acts as a screen. A photograph thus obtained is shown in figure 10.9.

This intensity distribution is very close to the distributions predicted by simulations in section 9.3.2. Also, the center of the image is close the center of the sensor element which indicates there are no severe alignment errors.

There are no adjustable components in the system. The reflected light intensity changes only a few percents when the instrument is disassembled and reassembled, which would suggest assembly tolerances are not a problem in practice. Shaking the sensor does not change the measurement signals noticeably, i.e. the structure is rigid enough.

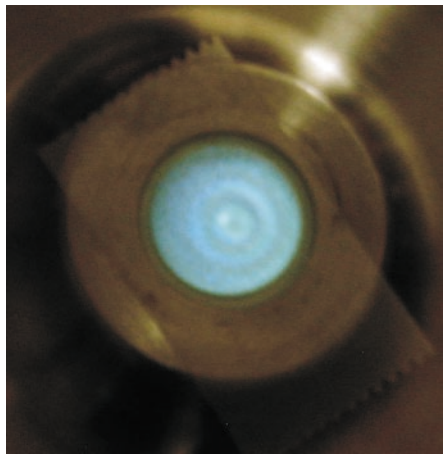


Figure 10.9 Intensity distribution on the sensor surface.

10.4 ELECTRONICS

As explained in section 8.4, only simple standard circuits are required in the measurement electronics. However, as the photodiode signals are high-impedance ones, some precautions are required to avoid the coupling of capacitive noise into the circuit.

The photodiode amplifiers are located very near to the photodiodes. Figure 10.10 shows the photodiode amplifier electronics attached onto the optical assembly. Figure 10.12 depicts the schematic diagram of the circuit. Also the LEDs are mounted on the same board so that they are on the same level with the photodiodes, as can be seen from figure 10.11.

Even the low-impedance voltage signals are sensitive to noise, because the voltage levels have to be measured accurately. The voltage signals are converted to digital signals on another circuit board located close to the photodiode amplifier board.

The AD board also carries the current sources for the LEDs. There are three independent current sources which are always supplying a constant current. This constant current can then be directed either to the LEDs or to load resistors.

It could also be possible to have only one current source which would be switched to one LED at a time. The scheme used here, however, enables powering several LEDs simultaneously, a property which is required in some detection schemes. Also, the use of individual current sources makes it possible to adjust the constant current of

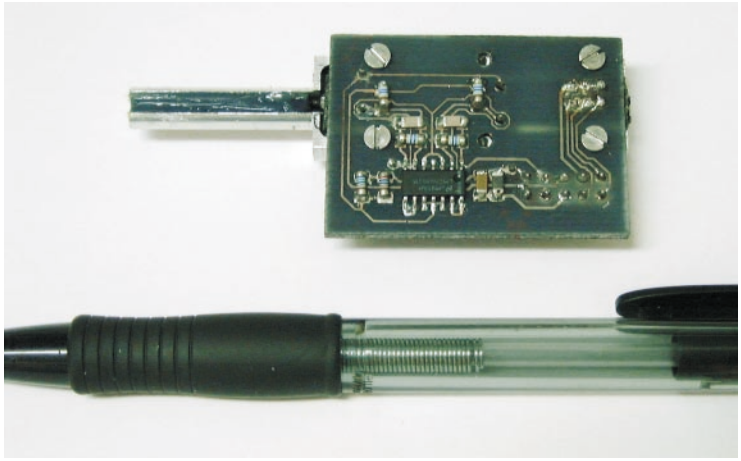


Figure 10.10 Photodiode amplifier electronics mounted on the optical assembly.

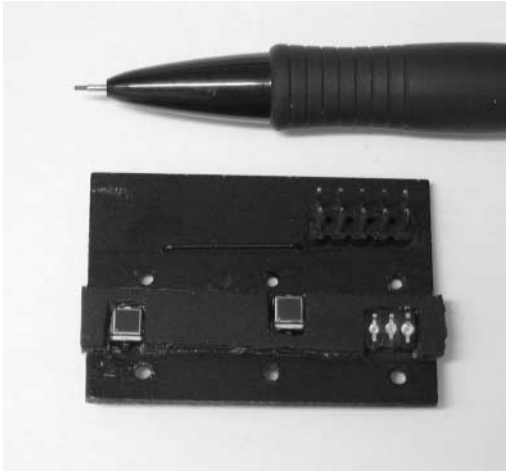


Figure 10.11 Photodiode side of the photodiode circuit board.

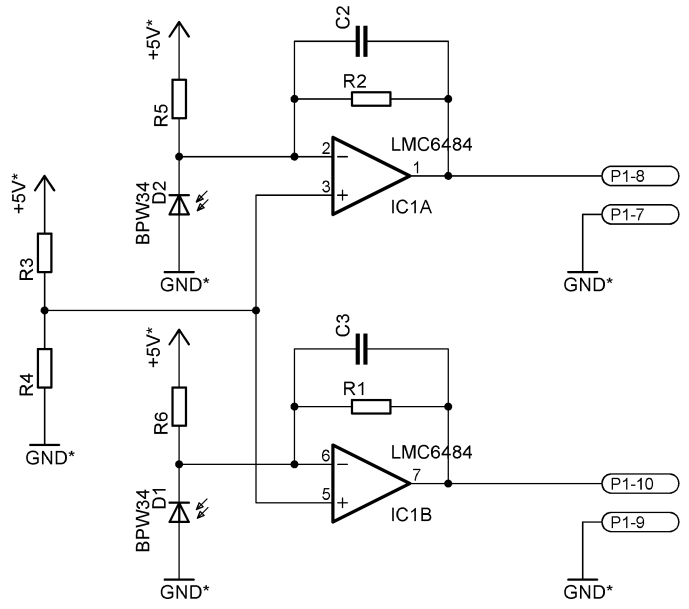


Figure 10.12 Circuit diagram of the photodiode amplifiers.

each LED so that the photodiode signals from different colors are on the same level.

In order to save some energy the current sources could be designed so that they are shut down when the corresponding LED is not lit. This, however, could produce spikes in the power source.

It should be noted that the circuit is entirely ratiometric. All voltage and current references are derived from the supply voltage. If the supply voltage drops, so does the LED current. The voltage signals from the photodetectors drop as the amount of light drops. This is compensated by the drop of reference voltage of the ratiometric AD converter.

The limits of this proportionality come from the non-linear current-to-intensity behavior of the LEDs. Fortunately, the supply voltage variations are in the order of a few thousandths, so this does not constitute a problem. Also, the reference measurement further cancels fluctuations in the intensity, i.e. there is another ratiometric loop built in the system.

The AD converters used in the circuit are Burr-Brown ADS1286 type successive approximation 12 bit converters. The operational amplifiers are National Semiconductor LMC6484 FET amplifiers which have very low leakage currents (typically less than a picoampere) and are able to perform rail-to-rail operation, i.e. the input common mode range extends over the full voltage range, and the output is able to sink or source current over the complete voltage range.

In the prototype the microcontroller is situated outside of the sensor housing. The microcontroller used in this application belongs to the AVR family of small 8 bit controllers. The measurement results are calculated by the controller and then sent to a computer for closer analysis through a serial interface.

10.4.1 Optoelectronic components

In order to have a good measurement signal the measurement wavelengths have to be chosen to fit the dye spectrum. The unavailability of certain wavelength LEDs limits the wavelength choice considerably, especially in the short wavelength end of the spectrum.

The wavelengths have been chosen to be approximately at the acid absorption (440 nm), isosbestic (495 nm), and base absorption (590 nm) wavelengths. These wavelengths are created by the LEDs: Agilent HLMP–DB25 ($\approx$ 430 nm peak wavelength), HLMP–CE23 (505 nm), and HLMP–EL24 (590 nm).

Traditionally, there have been very few blue-green LEDs on the market, and the existing devices have been very inefficient. The Agilent (former HP) cyan LED (505 nm) is one of the first cyan LEDs, and it has been aimed at traffic light use. Due to this, it was not available in a surface-mount package, and some customizations were required.

The three LEDs are all originally packaged in a 5 mm through-hole package. To fit them on the same board with the photodiodes, most of the epoxy of the package was ground off to give 5 mm $\times$ 5 mm $\times$ 2 mm package, as can be seen in figure 10.11. Also the leads were sawn off and the diodes were mounted as surface mount components.

Almost any visual range photodiodes would be adequate to be used as photodetectors in this application. Vishay BPW34 PIN detectors were chosen because of their suitable surface area, low capacitance, and good availability.

10.5 SIGNAL PROCESSING

One measurement cycle has four phases as shown in figure 10.13. Each LED is lit for 40 ms. The photodiode amplifier is let to settle for the first half of this period, the

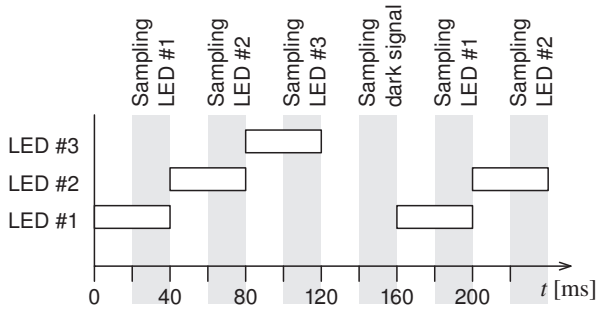


Figure 10.13 Led and data acquisition timing.

signals are sampled only during the latter half 20 ms.

Each signal is sampled 200 times with 100 µs intervals, and the samples are averaged. There are two reasons behind this averaging; reduction of random noise and reduction of the 50 Hz line frequency. There is a considerable amount of noise in the raw sampling results (in the order of several millivolts), and averaging seems to reduce that noise well in accordance with the square root law of white noise reduction.

As the photodiode measurement impedance is high (several megaohms), small stray capacitances inject millivolts of the 50 Hz line signal to the measurement. This can be practically eliminated by averaging over one full cycle of the line signal.[1]

Each measurement point consists of two readings, one from the detector photodiode (d_n) and one from the reference photodiode (r_n). The dark value of each detector (d_0, r_0) is subtracted from the measurement results, and the ratio of the two differences represents the ratio of emitted and detected light:

$$m_n = \frac{d_n - d_0}{r_n - r_0} \tag{10.1}$$

As the full measurement cycle takes 160 ms, it is fast compared to the changes in the dye film. In order to determine the noise of the instrument, 16 values m_n are collected from each detector. The average of these 16 values is sent for further processing along with the minimum and maximum values. Thus one sample of each color is sent to the computer every 2.56 s.

[1] Naturally, in the US and other 60 Hz regions the integration time has to be different. A well-known method of avoiding this problem is to integrate over 100 ms which is exactly 5 or 6 cycles depending on the line frequency. However, the prototype is intended only for test use, and the acquisition scheme has not been fully developed.

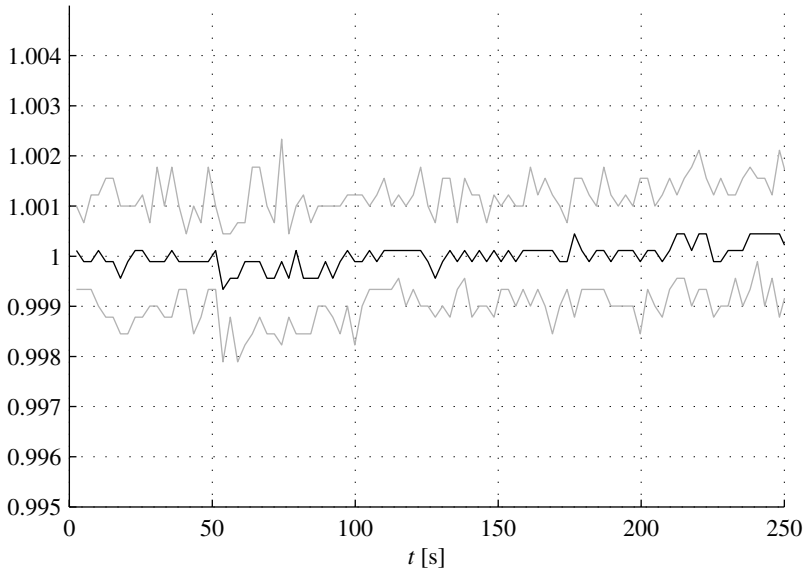


Figure 10.14 Reflectance values measured with the instrument. Average, minimum, and maximum of 16 consecutive measurements are plotted for each point.

The optoelectronic measurement has turned out to be drift-free and highly repeatable. Figure 10.14 shows some actual data recorded from the instrument on one wavelength. The y-axis is normalized so that the average value of the average values is unity.

It should be borne in mind that the 12 bit AD converters used in this system offer resolution down to 1/4096. While averaging does bring some more resolution, it does not compensate for possible integral non-linearities (INL) of the converters, and so any data beyond the INL specification of the converter (in this case one least significant bit) cannot be taken as generally valid. Compared to this specification the noise performance shown in figure 10.14 is good enough.

11 Measurements

The purpose of a pH measurement instrument is to indicate the number of active H_3O^+ ions per volume in a solution.

While this may sound trivial, there are several factors which may cause errors in the reading. Some of the general errors were already discussed in section 2.4. On the short list of these errors and deficiencies in any pH measurement instrument are:

- non-selective response to other ions,
- limited pH range,
- slow response,
- wear of sensor (erosion, leaching, chemical changes),
- errors induced by temperature,
- limited temperature range,
- limited pressure range,
- errors due to scaling (coating) of the sensor and
- drift.

In addition to these general pH measurement errors the novel measurement scheme introduced in this work may be vulnerable to errors caused by:

- changes in the external refractive index,
- color in the solution and
- error due to stray reflections from particles and bubbles.

The number of possible error sources is large. It becomes even larger when the combinatory nature of the errors is taken into account. For example, some wear phenomena may be especially pronounced at high temperature when there are certain ions present in the solution. Or, the sensor may be slow only at high pressures and in low-conductivity solutions.

It is evident that even decent practical knowledge of all parameters requires years of everyday use of the measurement method in real processes. While that kind of research belongs to the industrial research and development of a new sensor, it is beyond the scope of this work.

Emphasis has been set on studying the new properties and unique error sources of the thin film pH measurement. Naturally, the pH measurement accuracy is studied, as well as the dynamic response of the system. In addition to these, the errors caused by changing optical properties of the liquid have been under research.

Thermal properties of the measurement are not included in this research. All measurements have been carried out under room temperature conditions. Also, long term drift has not been measured; no leaching has been detected in the films during the measurements.

Pressure tests have also been excluded from this study. Theory suggests there should be very little change in the functioning as pressure rises. The Le Chatelier principle[1] states that higher pressures will favor reactions which decrease the overall volume of the reactants participating the reaction. The volume changes, however, are very small in these reactions, and as the pressure changes change liquid concentrations little, no significant changes are expected.

Chemical tests have been performed using well-known buffer solutions. These solutions are generally benign to pH measurement instruments. No cross-sensitivity tests are included in this work.

11.1 MEASUREMENT SETUP

All measurements (unless otherwise indicated) have been carried out with non-flowing liquids in a sample holder on top of the instrument. Figure 11.1 shows a diagram of this holder on top of the instrument.

The sample cavity is long (deep) and narrow. A long cavity is required because the distance from the sensing surface to the liquid surface has to be long enough to avoid the reflection from the liquid surface to cause errors to the measurement results.

[1] "A system at equilibrium, when subjected to a disturbance, responds in a way that tends to minimize the effect of the disturbance."

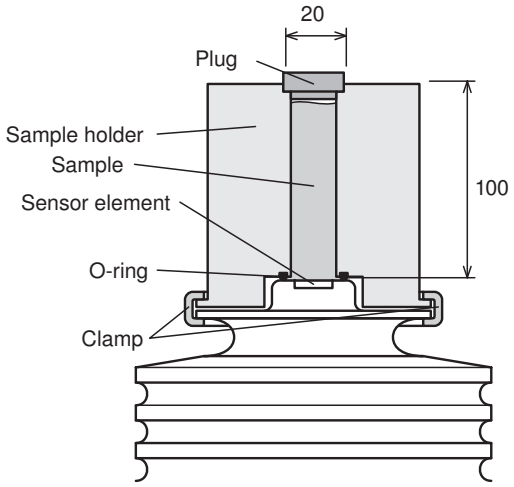


Figure 11.1 A diagram of the sample holder used in measurements. The general structure of the sensor head is visible in figure 10.1.

On the other hand, the smallish sample volume ($\approx$ 20 ml) is easy and quick to fill and empty.

A significant disadvantage associated with the narrow and long cavity is that there is very little liquid flow in the cavity. This may increase the response times and make the concentration distribution in the cavity uneven. Also, particle impurities in the sample liquid tend to deposit over the measurement surface and hence possibly cause measurement error.

The sample cavity is covered with a plug during the measurements. This is done to avoid external light errors in the measurement. Even though the measurement system is immune to non-altering external light, bright light may saturate the detector and thus cause significant errors. On the other hand, process pipes are dark inside, so the immunity against external light is not an important operational parameter.

After every sample the sample cavity has been sucked empty with a plastic pipette (glass pipettes might scratch the thin film surface) and rinsed with deionized water. While this procedure may not remove all traces of the previous solution, the procedure is sufficient when working with strongly buffered solutions.

All data from the instrument has been gathered to a computer through a serial interface, one measurement point every 2.56 s. This data has been manipulated with

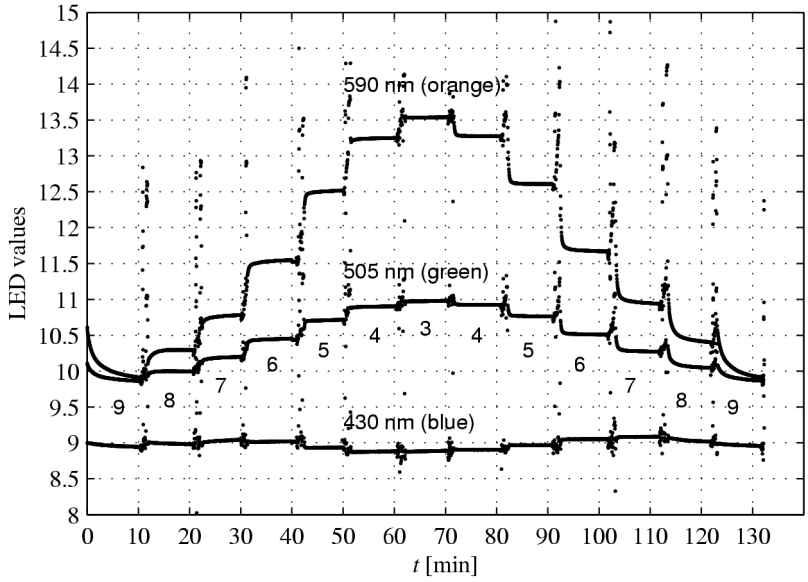


Figure 11.2 LED reflection values (arbitrary scale) obtained in a pH cycling test.

standard mathematical programs according to the formulae presented in section 8.1.

11.2 pH

The pH measurement accuracy was tested by cycling the pH from 9 to 3 and back with one pH unit steps so that each step takes 10 minutes. The LED values obtained in this test are shown in figure 11.2.

These measurement results are rather logical. However, comparing the desired LED positions to the BPB spectrum (11.3) the green LED value should remain stationary whereas the blue LED should react more to the pH changes. It seems that the dye spectrum has been shifted towards shorter wavelengths.

Fortunately, the measurement result interpretation scheme does not require any of the measurement points to be exactly at the isosbestic point. Nevertheless, the noise margin in the measurement is compromised with this mismatch between the LED wavelengths and the dye spectrum.

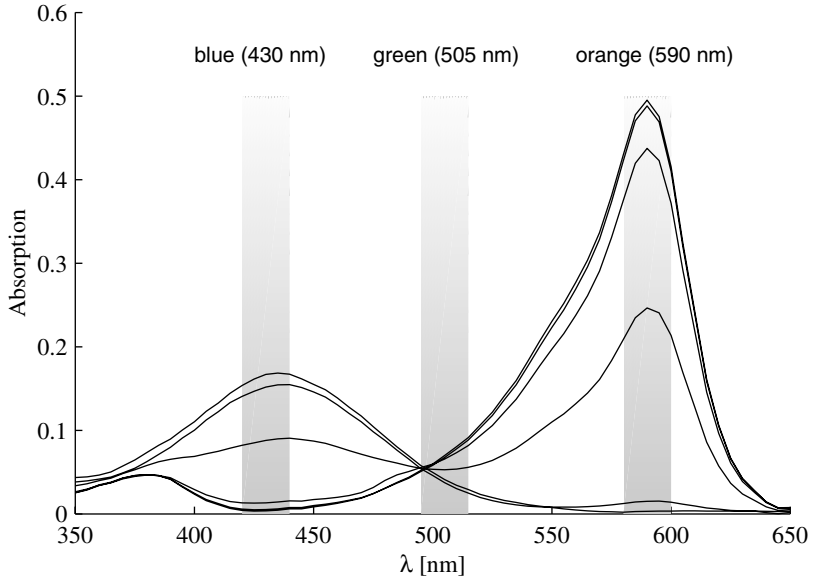


Figure 11.3 LED wavelengths and the dye (BPB) spectrum (aqueous solution).

The spurious points shown in the image result from the sample cavity emptying and cleaning, and they are not a result of any device malfunction.

There seems to be some drift in the measurement result. This is especially clearly visible in the blue reflection values when the measurement returns to the basic area at $t = 110$ min. This error seems to be similar at all wavelengths and may result from slight deposition on the sensor surface.

Whatever the reason for the drift it demonstrates that a single point measurement is not very useful in this context. The situation changes significantly when the points obtained by applying the equation (8.9) are plotted (figure 11.4):

$$f = \ln \frac{m_{430nm}}{m_{505nm}} + i \ln \frac{m_{590nm}}{m_{505nm}} \tag{11.1}$$

m_x refer to the raw measurement readings at different wavelengths.

According to the theory developed in section 8.1 these measurement points should lie a straight line if the points given by function f are drawn on the complex plane. The pattern thus obtained is shown in figure 11.5.

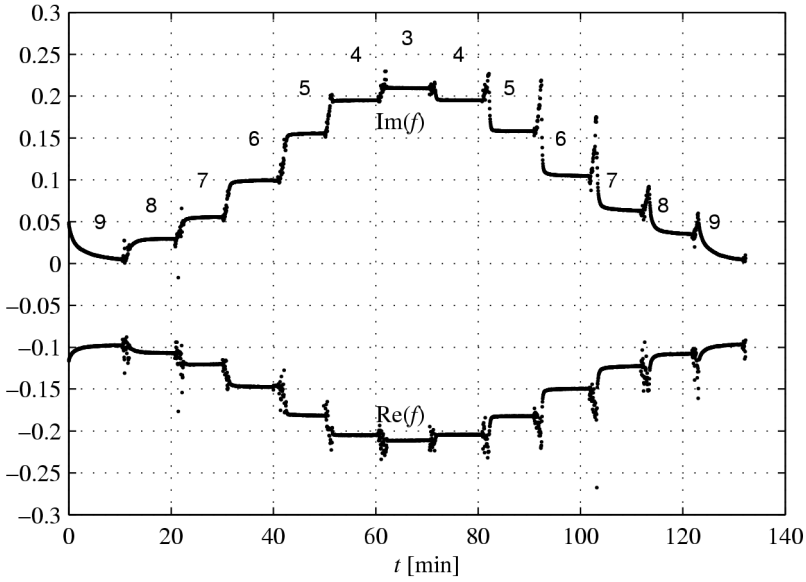


Figure 11.4 The values of Re(f) and Im(f) (equation (11.1)), calculated from the data depicted in figure 11.2.

The points follow a straight line quite well. The line seems to be slightly bent upwards in the basic end (right hand side in the figure). This bending is so small—and possibly just a coincidence—that the dye can be concluded to act according to the simple binary acid / base form model over its color changing pH range.

In order to evaluate the sensitivity of the thin film stack against dye concentration changes, the parameters k_1 and k_2 introduced in equation (8.9) have to be evaluated. The method outlined in section 8.1 may be utilized. However, the actual pH values at which the absorption on the different LED wavelengths are equal are not known as the exact spectral properties of the dye bound to the matrix are not known.

No leaching was observed during the experiments. Thus the sensitivity of the stack against dye concentration changes remains unclear and requires further experiments.

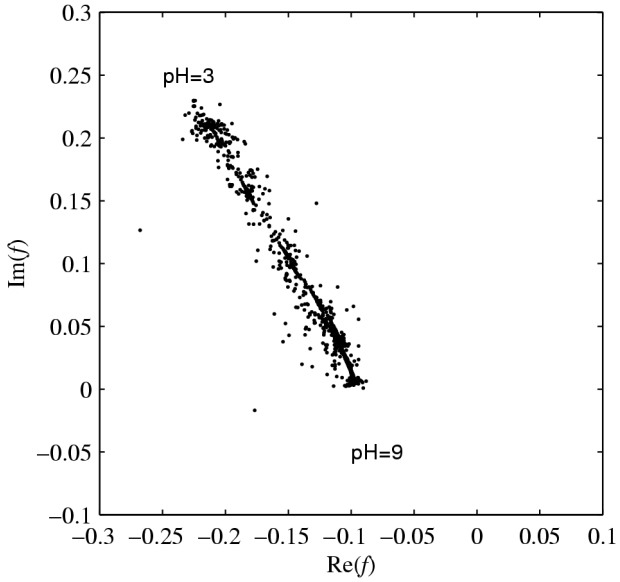


Figure 11.5 Locus of f during a pH cycle.

11.2.1 pH accuracy

If dye leaching is negligible, it is enough to look at either the real or the imaginary part of f. The following discussion will use the imaginary part of f.

Before the pH accuracy of the instrument can be determined, the measurement readings have to settle. To ensure this, the last readings before pH change are picked from the data depicted in figure 11.2, i.e. the instrument has had approximately 10 minutes time to settle before each reading.

The $\mathrm{Im}(f)$ calculated from these values is shown as a function of pH in figure 11.6.

There is still some hysteresis left in the curve. The hysteresis points seem to follow a clear pattern, and the amount of clearly visible randomness in the curve is very small. The s-shape of the curve is expected, as this is a typical saturation curve, acid and base forms dominate in the ends of the curve.

To obtain more quantitative estimates of the accuracy, a function c_{pH} which converts $\mathrm{Im}(f)$ to pH values has to be formed. As there is no knowledge of the intermediate points, and as there is no simple theory to explain the curve shape quantitatively,

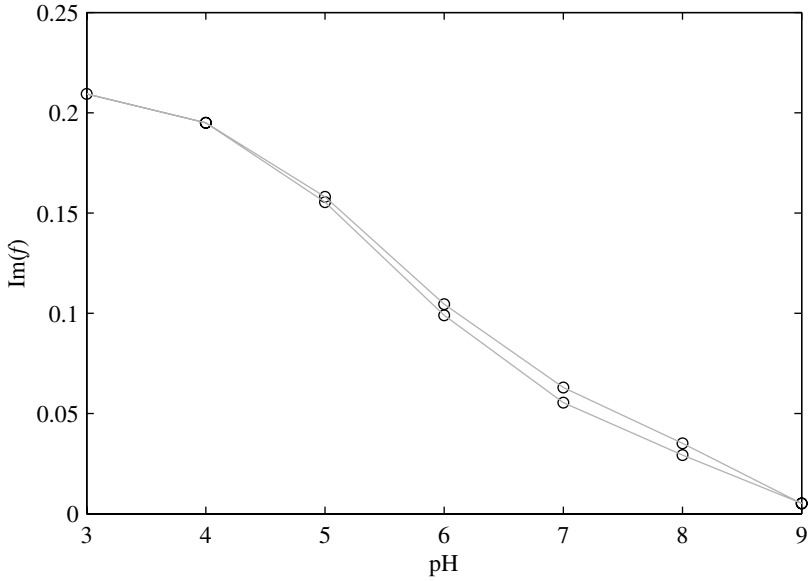


Figure 11.6 Im(f) plotted versus pH.

c_{pH} can be formed by using the data points. A linear interpolation would produce discontinuities to the differential of the function, so a cubic spline function is used. Almost any sufficiently smooth interpolation function is equally good in this context.

The pH values calculated by using the interpolation function c_{pH} are depicted in figure 11.7. The values are naturally very close to the real pH values as c_{pH} has been defined through these values. The value of this graph is, however, that it gives a realistic estimate of repeatability. It seems that after the settling time the hysteresis remains within ±0.1 pH for all measurement points.

An even more interesting feature to note is that even the values which exhibit considerable hysteresis (i.e. higher pH values) tend towards the expected value. This would propose that with a longer settling time the hysteresis would be smaller. The principle of operation of this sensor structure does not predict any hysteresis, and thus this observation is in accordance with the theory.

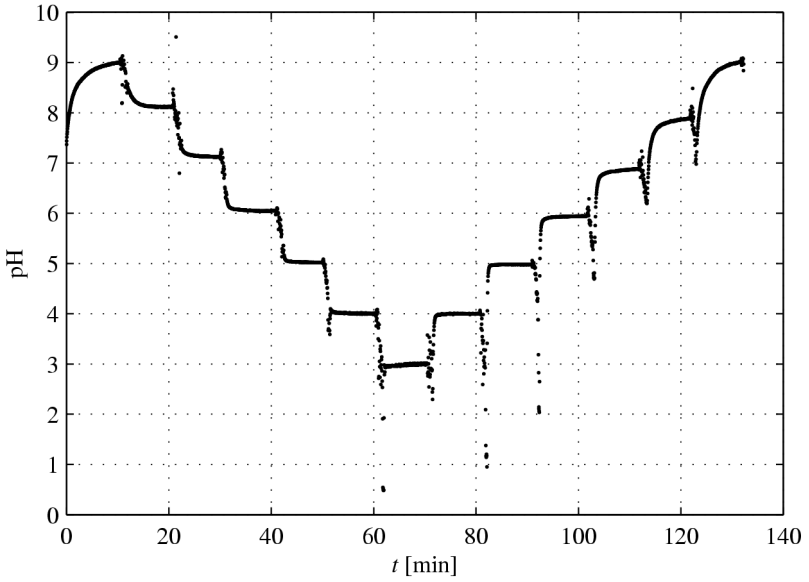


Figure 11.7 $c_{\mathrm{pH}}(\mathrm{Im}(f))$ during the pH cycle.

11.2.2 Dynamic response

There is an interesting dynamic feature visible in figure 11.7; the response time seems to be shorter at low pHs. At high pH values the response slows down very noticeably. While the time constant (to e^{-1} of the full step) of the step between pH = 3 and pH = 4 is approximately 10 s to 20 s, it is several minutes for the step between pH = 8 and pH = 9.

This change of time constant is even more clearly visible from a pH cycle which has been carried out with shorter settling times. The cycle shown in figure 11.8 has 5 minute settling times, of which approximately one minute is required to the sample cavity emptying, rinsing, and refilling. The function c_{pH} determined earlier has been used to convert the LED values to pH values.

From this graph it is evident that the 5 minute settling time is too short for values above pH = 6. This behavior is rather clear, but there is another feature which is more intriguing. The settling from below and from above is different. With increasing pH values the reading falls down during the sample cavity rinsing and comes up with

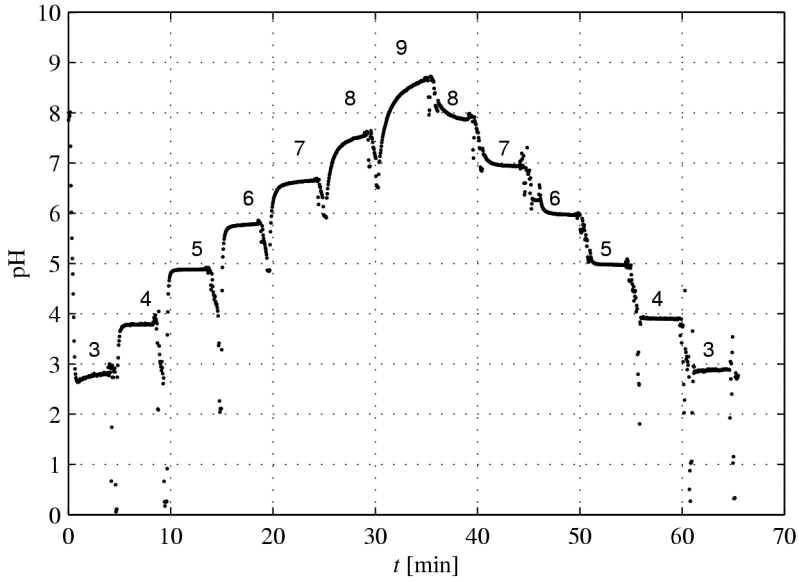


Figure 11.8 $c_{\mathrm{pH}}(\mathrm{Im}(f))$ during a faster pH cycle.

the slow settling delay. With decreasing pH values the readings decrease smoothly without noticeable jumps.

This asymmetric behavior cannot be explained by the rinsing procedure, as the rinsing procedure is performed similarly for each sample. While the time constant seems to be symmetric, the total settling time for upward steps is longer as the step starts from further down.

To study this phenomenon the sensor was tested with large step changes in pH. The pH was cycled repeatedly from 3 to 9 and back (figure 11.9).

The response time pattern follows those observed earlier; high pH values have slower response, and the five-minute settling time is too short. A new feature not visible in the previous pH cycles is the noticeable undershoot in high-to-low transitions. At pH = 9 the signal does not have enough time to settle completely and it falls short by approximately 0.5 pH but the low pH undergoes a considerable undershoot and then recovers slowly towards the correct value.

Neither of these asymmetric phenomena is important in a continuous process.

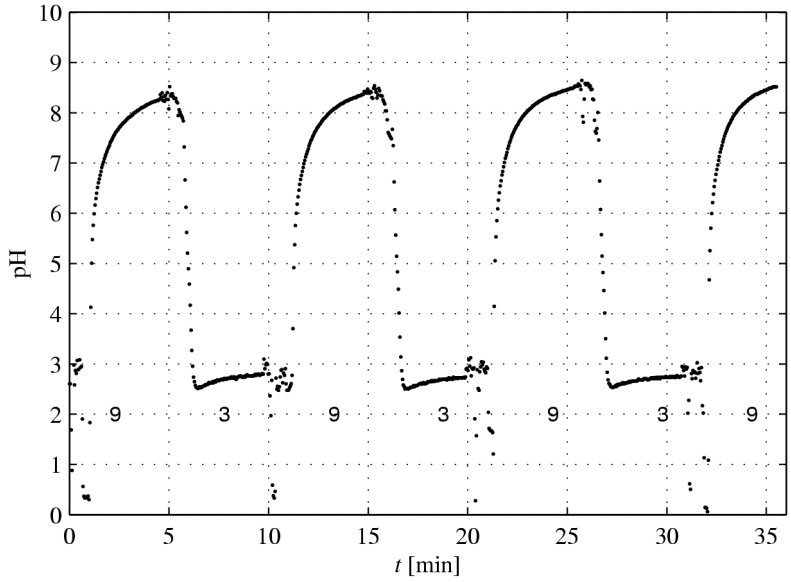


Figure 11.9 $c_{\mathrm{pH}}(\mathrm{Im}(f))$ during repeated large pH changes.

They are, however, an indication of some interesting dynamic phenomena which are not fully understood and require more research.

11.3 REFRACTIVE INDEX ERRORS

The optical measurement principle is sensitive to the optical properties of the liquid under measurement. Probably the most important source of these errors is the change in the process medium refractive index.

As discussed in sections 5.5 and 5.6, there are two different mechanisms for the external refractive index to change the measurement values. The change of external refractive index directly changes the reflection on the outer surface of the dielectric mirror, and the liquid filling the pores changes the film refractive indices.

In order to study these effects a sensor element without the indicator was utilized to separate pH and refractive index effects from each other. Figure 10.4 shows the two mirror stacks (one with and one without dye).

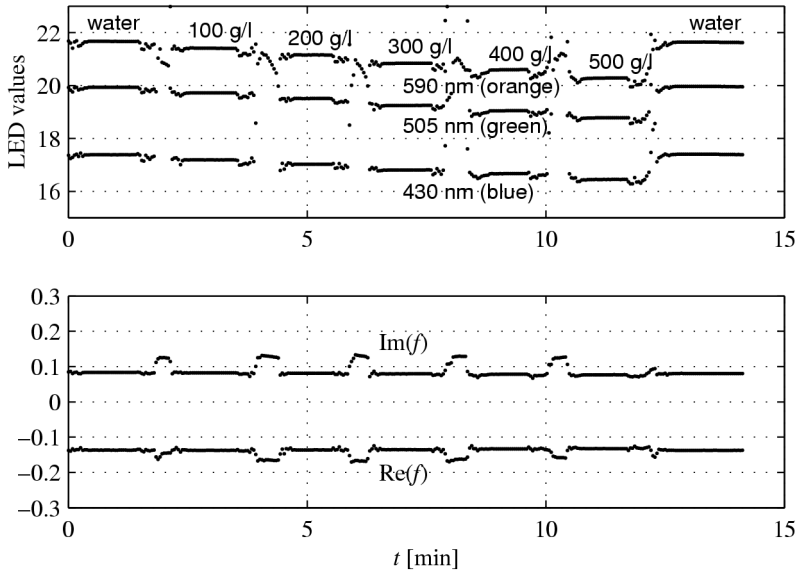


Figure 11.10 LED values and f with different sucrose concentrations.

To create known and realistic refractive index solutions, sucrose solutions with different concentrations (from 0 to 500 g/l) were prepared and measured. The LED values measured with these liquids are plotted in figure 11.10.

The refractive index range covered by the sucrose solutions is approximately 1.33 to 1.42. The overall change of LED values is very pronounced as a function of pH. However, when the function f is formed from these values, the points remain on a reasonably small area (figure 11.10).

Both the imaginary and the real part of f seem to change slightly as a function of external refractive index. By using the conversion function c_{pH} obtained in section 11.2.1 the changes in the indicated pH can be estimated to be in the order of 0.1 pH with the midrange pH values.

11.4 COLOR ERROR

The effects of external color were estimated by measuring two different concentrated color solutions (red and blue food dye) with the undyed mirror. The extinction coef-

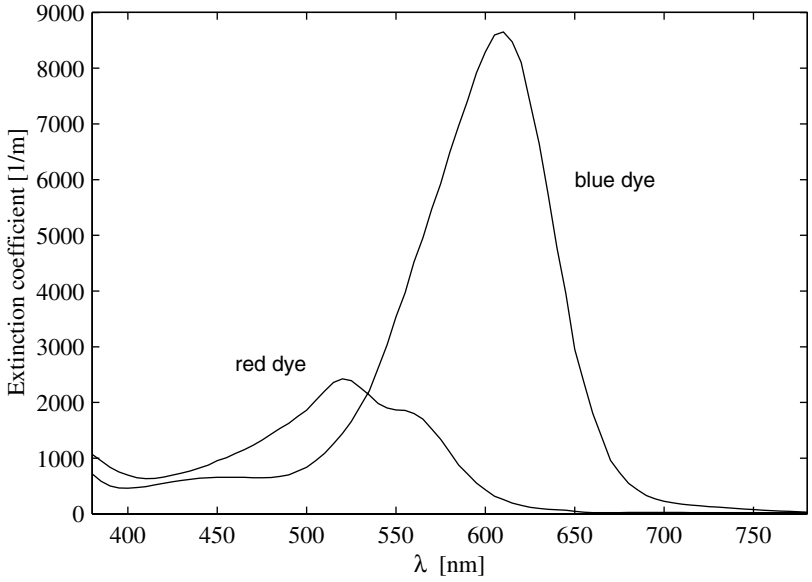


Figure 11.11 Extinction coefficients of the dyes used in color sensitivity measurements.

ficients of the two dyes are shown in figure 11.11. These extinction coefficients are measured with a spectrophotometer from 1:100 diluted solutions, as the concentrated solutions (figure 11.12) are too dark to measure even with a short path (l = 1 cm).

Figure 11.13 shows the LED values and f calculated from the values when the sample is repeatedly changed between water and dye. Similar results with the red dye is shown in figure 11.14.

A closer look at the graphs shows that there are some regular changes in f due to the dyes. These changes do not, however, correlate well with the dye spectrum, and they may be caused by, e.g. particles or small bubbles in the solutions. Also, these changes are small and partly time-dependent as shown in the magnification of Re(f) of the blue dye test (figure 11.15). Any pure optical extinction should be immediate, not delayed.

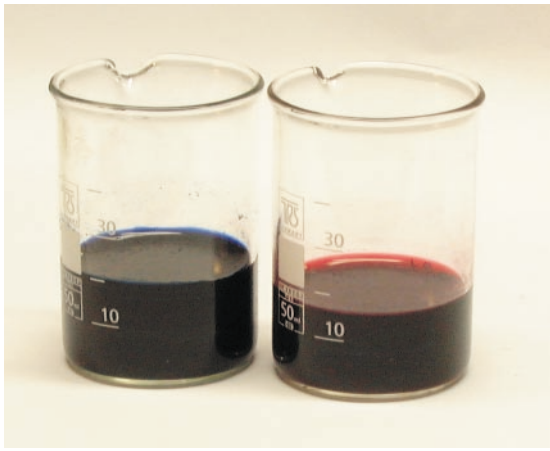


Figure 11.12 Concentrated dye solutions used in color sensitivity measurements.

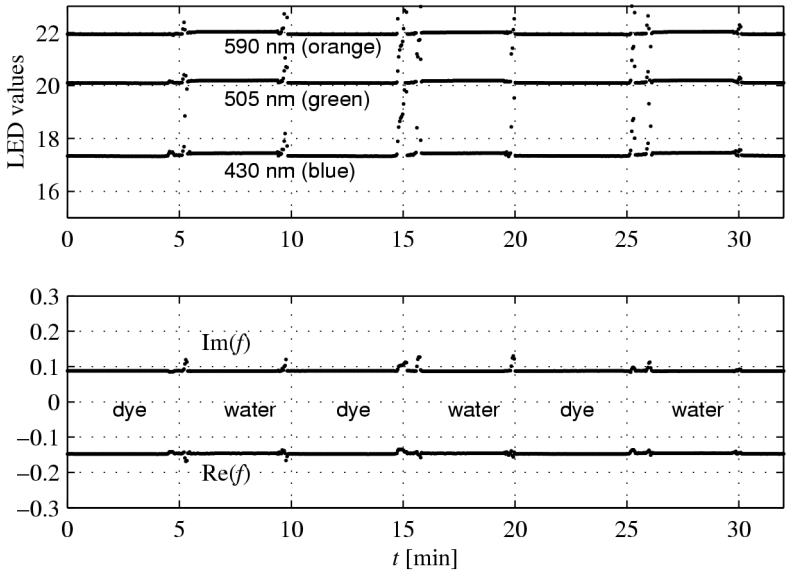


Figure 11.13 LED values and $\mathrm{Re}(f), \mathrm{Im}(f)$ when water and blue dye are cycled.

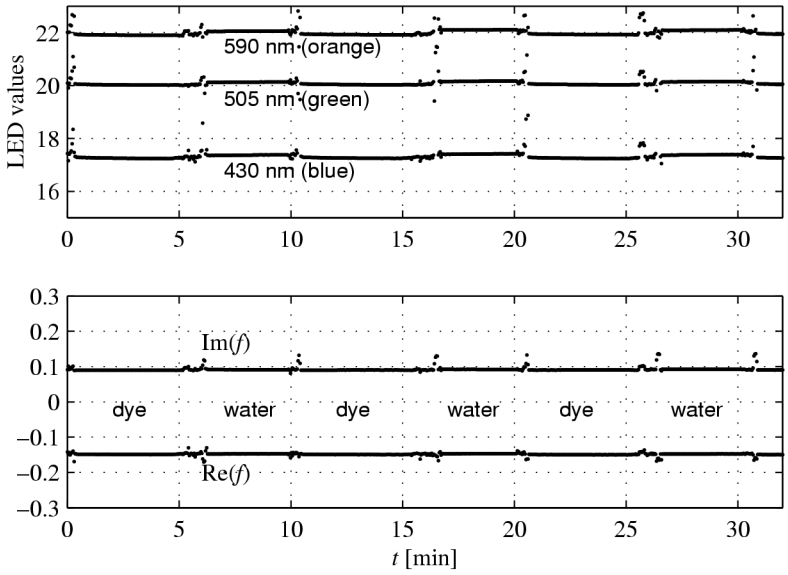


Figure 11.14 LED values and Re(f), Im(f) when water and red dye are cycled.

11.5 DISCUSSION

The measurement results are well in accordance with the theory developed earlier in this work. The sensor has been demonstrated to function as expected.

The pH range of the sensor extends at least from 3 to 9. Beyond this range there is still some response, but as the first differential of c_{pH} increases, measurement uncertainties grow quickly. The range is significantly wider than that of the aqueous indicator; the color change region of BPB is usually given to be from pH = 3.0 to pH = 4.6. [61]. This expansion of the range has been generally acknowledged to be the result of the dye being dissolved into the silica matrix.

The time constant of the sensor film is rather long at high pH values. One possible explanation to this is that the film is sensitive to H_3O^+ ions and there are few H_3O^+ ions available at high pH. This hypothesis does not, however, explain the undershoot noticed in some step pH changes.

While the sensor is slow at high pH values, its response and accuracy still compare favorably to many other optical sensor structures recently suggested in the literature

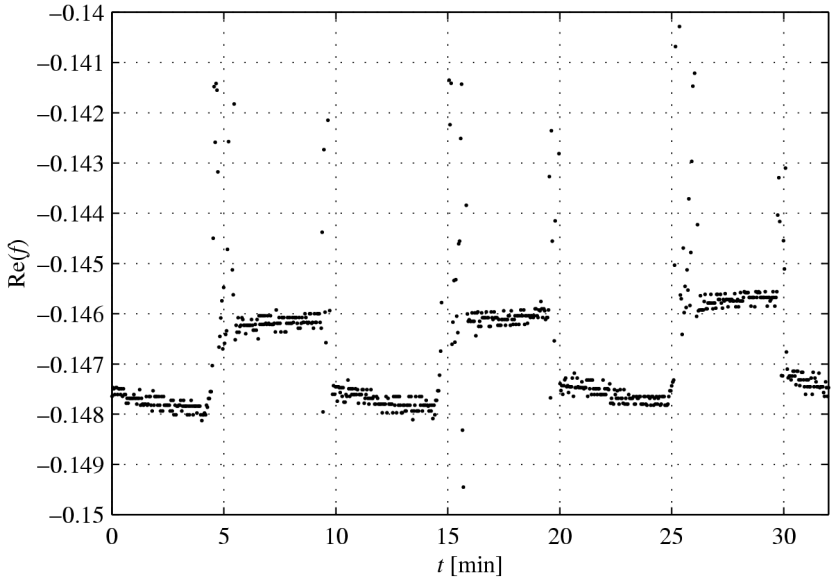


Figure 11.15 A magnification of Re(f) of the blue dye test (figure 11.13).

(e.g., [18]).

No leaching was noticed during the experiments, and the used films appear intact when visually compared to new films. However, it is still early to draw conclusions concerning the long-term durability of the measurement films in real process conditions with high temperatures and flow rates.

The measurement repeatability is better than ±0.1 pH after the sensor readings have settled. If this repeatability can be maintained in the long term, the instrument will surpass any existing pH measurement instruments in this context. While it is still too early to conclusively state the ultimate capability of this method in that respect, the simple working principle makes a good starting point for a reliable and accurate measurement method. Also, the instrument can be made so that replacing possibly aging damaged components is easy.

Changes in the optical properties of the solution under measurement do not seem to cause large errors to the measurement. Changes in the refractive index do cause a noticeable error if the refractive index values vary in a wide range. This error, however, is still small in the absolute scale. Changes in external color do not seem to produce

significant errors even when the liquid is heavily dyed.

The instrument itself performs as expected with the exception of the misplaced LED wavelengths. This error is probably due to dye spectrum shift when the dye is dissolved in the matrix filled with water. The absorbing film effects on reflection discussed in section 5.4 should shift the spectrum in the opposite direction, and thus do not seem to be applicable here.

There are several error sources discussed in the beginning of this chapter which have not been tested. Further research is still required on those points, but the measurements presented here provide a good starting point for the real application work.

12 Conclusions

The main goal of this thesis is to demonstrate the feasibility of making a pH measurement instrument which uses the new reflective measurement principle. This goal has been achieved, a prototype has been constructed and its measurement properties have been demonstrated.

The instrument outlined in this thesis has the potential to function as a reliable process instrument. The simplicity and small number of optical and electronics parts required in the instrument makes it simple and economical to manufacture. Also, the simple construction makes the instrument easier to make rugged enough to withstand the process conditions.

The new optical construction shown in this work uses some novel optical components which can be manufactured from a simple glass rod. These components have been demonstrated to function according to the theory and simulations developed in this thesis.

The measurements carried out with the prototype show that the new measurement method is potentially very accurate. If the instrument is able to keep its ± 0.1 pH repeatability in the long run without recalibration, it will compare favorably to any pre-existing pH sensor designs. However, more research is required in order to finally clarify the long term stability of the completed instrument.

The electronic structure of the new instrument is straightforward. It is possible to make the instrument so that it consumes only a few dozens of milliwatts. Consequently, the instrument is easy to design so that it can be used in explosion hazard environments, as well.

The new measurement method is sensitive to optical properties of the medium under measurement. This sensitivity can be eliminated almost completely by using the novel algorithms presented in this work, as demonstrated in the measurement part.

An essential part of the sensor element manufacturing process is the wet deposition of thin films by dipping. By incorporating a position measurement device and computer control to the dipping instrument (dipper), it is possible to manufacture a

large number of different thin film structures on a single substrate with a small number of process steps.

This thesis forms a starting point for industrial use of the new pH measurement method. Further research is under way to further clarify the behavior of the indicator structure. However, the results obtained this far are encouraging.

In the long run it is possible to make similar instruments which measure other chemical properties, such as other ions or redox potential. This can be done simply by changing the indicator molecule to an indicator which is sensitive to some other property; there is nothing pH specific in the measurement principle introduced in this thesis.

Bibliography

[1] S. Zumdahl, *Chemistry*, 3^{rd} ed., D.C. Heath and Co., 1993

[2] S.P.L. Sørenson, "Enzyme Studies II. The Measurement and Meaning of Hydrogen Concentration in Enzymatic Processes", *Biochemische Zeitschrift*, 21:131–200 (1909), English translation and excerpts in `http://dbhs.wvusd.k12.ca.us/Chem-History/Sorenson-article.html`

[3] P. Atkins, *Physical Chemistry*, 6^{th} ed., W.H. Freeman & Co., 1998

[4] J.G. Cummings and K. Torrence, "Chemical analysis — electrochemical techniques" in *Instrumentation Reference Book*, 2_{nd} ed., B.E. Noltingk (Ed.), Butterworth-Heinemann, 1995, pp. 105–133

[5] R. Keränen, "pH-mittaus — käytännön esimerkkejä ja kokemuksia", lecture notes, AEL/INSKO seminar on process measurements, 19 November 1998

[6] J.-P. Ylén, "pH-arvon mittaaminen", *Automaatioväylä*, 1/2001

[7] H. Jensen, L. Nielsen, "Uncertainty of pH measurements", *NT technical report 284*, Danish Institute of Fundamental Metrology, 1995

[8] *Determination of pH-value of water*, Finnish standard SFS3021, 1979

[9] K.S. Fletcher, "pH Analyzers", in *Analytical Instrumentation*, R.E. Sherman and L. Rhodes (Eds.). Instrumentation Society of America, 1996, pp. 431–456

[10] S.D. Moss, J. Janata, C.C. Johnson, "Potassium Ion-Sensitive Field Effect Transistor", *Anal. Chem.*, 47:2238–2243 (1975)

[11] A. Dybko, "Field effect transistors (FETs) as transducers in electrochemical sensors", *Chemia Analityczna*, 41:697 (1996), English translation in `http://www.ch.pw.edu.pl/~dybko/csrg/papers/index.htm`

[12] http://www.phmeters.com/Isfet_pH_Information.htm, IQ Scientific Instruments, Inc.

[13] G. Monkman, "München: pH and SAW", *Sensor Review*, Vol. 16 2:28-31 (1996)

[14] K. Nassau, *The physics and Chemistry of Color — The fifteen causes of color*, John Wiley & Sons, 1983

[15] K. Vuokila, *Sol-gel thin films and their applications*, Licentiate thesis, University of Joensuu, 1999

[16] M. Afromowitz et al., "Microfluidic Chemical Analytical Systems (μFCAS)", Paul Yager Research Group, University of Washington, http://faculty.washington.edu/yagerp/progressinmfcas.html

[17] O.S. Wolfbeis et. al., "Sol-Gels and Chemical Sensors", *Structure and Bonding 85: Optical and Electronic Phenomena in Sol-Gel Glasses and Modern Applications*, Springer-Verlag, 1996

[18] P.A. Wallace et al., "Development of a quasi-distributed optical fibre pH sensor using a covalently bound indicator", *Meas. Sci. Technol.* 12:882–886 (2001)

[19] B. Gupta and D. Sharma, "Evanescent wave absorption based fiber optic pH sensor prepared by dye doped sol-gel immobilization technique" *Optics Communications* 140:32–35 (1997)

[20] F. Baldini, "Critical review of pH sensing with optical fibers", *Chemical, Biochemical, and Environmental Fiber Sensors X*, SPIE proceedings no. 3540, 1999

[21] J. Peterson and S. Goldstein, *Fiber optic pH probe*, US Patent 4,200,110 (1980)

[22] L. Iyer and Y. Zhao, *Fiber-optic detectors with terpolymeric analyte-permeable matrix coating*, US Patent 5,640,470 (1997)

[23] http://www.oceanoptics.com/products/phsensor.asp, Ocean Optics, Inc.

[24] G. Wicks et al., *Tetraethyl orthosilicate-based glass composition and method*, US Patent 5,637,507 (1997)

[25] C.J. Brinker and G.W. Scherer, *Sol-gel science — the physics and chemistry of sol-gel processing*, Academic Press, 1990

[26] W. Geffcken and E. Berger, *Verfahren zur Änderung des Reflexionsvermögens optischer Gläser*, German Patent 736 411 (1943)

[27] H. Dislich and P. Hinz, "History and principles of the sol-gel process, and some new multicomponent oxide coatings", *J. Non-Cryst. Solids*, 48:11–16 (1982)

[28] B. Yoldas, "Preparation of glasses and ceramics from metal-organic compounds", *Journal of Material Science*, 12:1203–1208 (1977)

[29] J-L. Nogues, C. Balaban, and W. Moreshead, *Making sol-gel monoliths*, US Patent 5,076,980 (1991)

[30] A. Venkateswara Rao, G. Pajonk, and N. Parvathy, "Effect of solvents and catalysts on monolithicity and physical properties of silica aerogels", *Journal of Material Science*, 29:1807–1817 (1994)

[31] `http://www.nanopore.com/`, Nanopore, Inc.

[32] A. Hunt and M. Ayers, "A brief History of Silica Aerogels", Ernest Orlando Lawrence Berkeley National Laboratory, `http://eande.lbl.gov/ECS/aerogels/sahist.htm`

[33] B. Yoldas and D. Partlow, "Wide spectrum antireflective coating for fused silica and other glasses", *Applied Optics*, 23:1418–1424 (1984)

[34] B. Yoldas, "Investigations of porous oxides as an antireflective coating for glass surfaces", *Applied Optics*, 19:1425–1429 (1980)

[35] H. Pulker, *Thin Film Science and Technology 6: Coatings on Glass*, Elsevier Science Publishing, 1984, pp. 7–32

[36] H. Schmidt and M. Mennig, "Wet coating techniques for Glass", `http://www.solgel.com/articles/Nov00/mennig.htm`, Institut für Neue Material, 2000

[37] M.Langelet et al., "Glass and ceramic thin films deposited by an ultrasonically assisted sol-gel technique", *Thin Solid Films*, 221:44–54 (1992)

[38] D. Meyerhofer, "Characteristics of resist films produced by spinning", *Journal of Applied Physics*, 49:3993–3997 (1978)

[39] D. Birnie, "Common Defects Found When Spin Coating", `http://www.mse.arizona.edu/faculty/birnie/Coatings/Defects.htm`

[40] H. Dislich and E. Hussmann, “Amorphous and crystalline dip coatings obtained from organometallic solutions: procedures, chemical processes and products”, *Thin Solid Films*, 77:129–139 (1981)

[41] `http://www.geltech.com/`, Geltech, Inc.

[42] E. Nappi, “Aerogel and its application to RICH detectors”, *IFCA Instrumentation Bulletin*, Fall 1998

[43] `http://stardust.jpl.nasa.gov/spacecraft/`, NASA/Stardust mission, Jet Propulsion Lab

[44] `http://www.microvacuum.com/`, MicroVacuum Ltd.

[45] E. Przybylowicz and A. Millikan, *Integral analytical element*, US Patent 3,992,158 (1976)

[46] P. Hinz and H. Dislich, “Anti-reflecting light-scattering coatings via the sol-gel-procedure”, *Journal of Non-Crystalline Solids*, 82:411–416 (1982)

[47] P. Baumeister and G. Pincus, ”Optical interference coatings”, *Scientific American*, Dec. 1970, pp. 59–75

[48] D. Partlow and T. O’Keeffe “Thirty-seven layer optical filter from polymerized solgel solutions”, *Applied optics*, 29:1526–1529

[49] E. Hecht, *Optics*, 2^{nd} ed., Addison-Wesley, 1987

[50] M. Born and E. Wolf, *Principles of Optics*, 6^{th} ed., Pergamon Press, 1993

[51] `http://www.micro-e.com/`, MicroE Systems Inc.

[52] D. Mitchell and W. Thorburn, *Apparatus for detecting relative movement wherein a detecting means is positioned in the region of natural interference*, US Patent 5,486,923 (1996)

[53] D. Mitchell, *Apparatus for detecting relative movement*, US Patent 5,559,600 (1996)

[54] `http://www.microparts.de`, MicroParts GmbH

[55] *2000 Optics and optical instruments catalogue*, Edmund Industrial, 2000, p. 43

[56] *The Practical Application of Light*, Melles Griot Inc., 1999, p. 6.34

[57] *PIN Photodiode BPW34 Data Sheet*, Vishay-Telefunken GmbH

[58] S. Jennato and G. McKee, "What Color is My LED?", *Photonics Spectra*, May 2001

[59] *Cyan LED HLMP-CE23 Data Sheet*, Agilent Technologies

[60] V. Voipio, *Four-ray Process colorimeter*, Master's thesis, Helsinki University of Technology, 1997

[61] *CRC, Handbook of Chemistry and Physics*, 80^{th} ed., 1999–2000, pp. 8-16 – 8-18.

APPENDIX A Patent on reflective indicator measurement

The idea of using a dielectric mirror stack on top of an indicator dye layer is protected by a patent. This patent is shown on the following pages.

Along with the US patent shown here, there are parallel patents and applications worldwide: DE19927484 (Germany), JP2000046740 (Japan), and FI981424 (Finland).

US006208423B1

(12) **United States Patent**
Voipio et al.

(10) **Patent No.: US 6,208,423 B1**
(45) **Date of Patent: Mar. 27, 2001**

(54) **ARRANGEMENT AT MEASUREMENT OF PH OR ANOTHER CHEMICAL PROPERTY DETECTABLE BY DYE INDICATORS**

(75) Inventors: **Ville Voipio**; **Katri Vuokila**, both of Vantaa (FI)

(73) Assignee: **Janesko Oy**, Vantaa (FI)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/328,263**

(22) Filed: **Jun. 17, 1999**

(30) **Foreign Application Priority Data**

Jun. 18, 1998 (FI) 981424

(51) **Int. Cl.**[7] **G01N 21/47**
(52) **U.S. Cl.** **356/446**
(58) **Field of Search** 356/445, 446, 356/234, 432, 436, 409, 300, 326, 346, 345; 422/55, 57, 58, 82.05, 82.08, 82.09; 436/169

(56) **References Cited**

U.S. PATENT DOCUMENTS

3,992,158	*	11/1976	Przybylowicz et al.	23/253 TP
4,649,123		3/1987	Charlton et al.	436/79
5,039,491		8/1991	Saaski et al.	422/82.05
5,268,145		12/1993	Namba et al.	422/57
5,608,519	*	3/1997	Gourley et al.	356/318

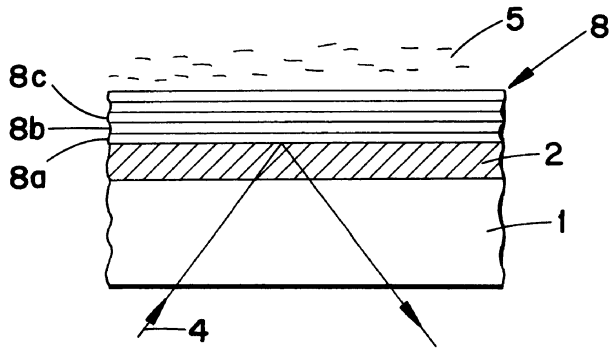
* cited by examiner

Primary Examiner—Frank G. Font
Assistant Examiner—Michael P. Stafira
(74) *Attorney, Agent, or Firm*—Burns, Doane, Swecker & Mathis, L.L.P.

(57) **ABSTRACT**

The invention relates to an arrangement at measurement of pH or another chemical property detectable by dye indicators. To provide a simple and safe solution, the measuring part is formed of a glass sheet or a similar substrate, the substrate being coated with a dye film, which is arranged to change its color in a manner known per se when a chemical property of the environment changes. The dye film is arranged to serve as a light reflective surface, whereby the chemical property of the solution to be measured can be measured as light reflection measurement.

7 Claims, 1 Drawing Sheet



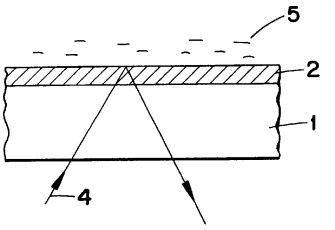


FIG. 1

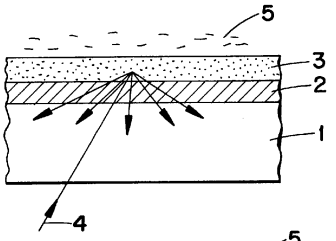


FIG. 2

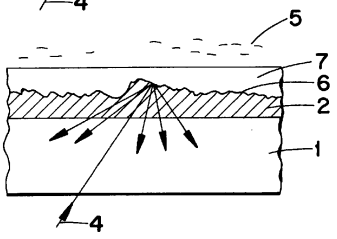


FIG. 3

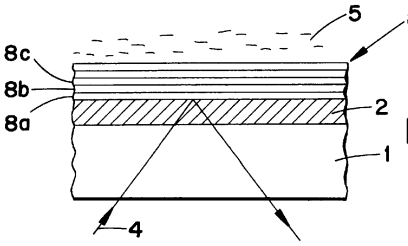


FIG. 4

ARRANGEMENT AT MEASUREMENT OF PH OR ANOTHER CHEMICAL PROPERTY DETECTABLE BY DYE INDICATORS

FIELD OF THE INVENTION

The invention relates to an arrangement at measurement of pH or another chemical property detectable by dye indicators, the arrangement comprising a glass sheet or a similar substrate to be immersed into a solution to be measured, the substrate being coated with a dye film arranged to change its colour when a chemical property of the environment changes, and means for leading light through the substrate to the dye film.

BACKGROUND OF THE INVENTION

At measurement of pH in solutions, strips made of litmus paper, for instance, are generally used, the strips being immersed into a solution to be measured, whereby the paper changes its colour in accordance with the environment, i.e. the pH value of the solution.

In principle, the above procedure functions at least in some situations, for instance in laboratory conditions. However, a problem is the difficulty and slowness of the measurement. In addition, the procedure is inconvenient for example in certain industrial conditions. Another drawback of the above technique is its one-time nature. At present, glass membrane sensors based on electrochemical phenomena are mainly used for pH measurement in the industry.

SUMMARY OF THE INVENTION

The object of the invention is to provide an arrangement, by means of which the drawbacks of the prior art technique can be eliminated. This has been achieved by means of the invention. The arrangement of the invention is characterized in that the dye film is covered with a structure, comprising at least one layer, allowing ions to pass through and reflecting light backwards, whereby a colour change in the dye film can be measured as light reflection measurement.

The main advantage of the invention is its simplicity, which makes the introduction and use of the invention advantageous. Another advantage of the invention is that the measurement can be automated in an advantageous manner and, additionally, the measurement can preferably be performed directly in a process pipe as well.

BRIEF DESCRIPTION OF THE DRAWING FIGURES

In the following, the invention will be described by means of preferred embodiments shown in the attached drawing, whereby

FIG. **1** shows the basic principle of a first embodiment of an arrangement of the invention,

FIG. **2** shows the basic principle of a second embodiment of the arrangement of the invention,

FIG. **3** shows the basic principle of a third embodiment of the arrangement of the invention, and

FIG. **4** shows the basic principle of a fourth embodiment of the arrangement of the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

As shown in the figures, the essential thing with the invention is that a glass sheet or a similar substrate **1** is coated with a dye film **2**, which changes its colour in a manner known per se, when a chemical property, e.g. the pH value, of the environment changes. The dye of the dye film **2** may have for instance two states, viz. acid state and alkali state, having different colours and strengths. However, it shall be noted that the pH measurement is not the only application field of the invention, but the invention may also be applied for example to the measurement of the concentration of different metal ions in a solution.

FIG. **1** shows the essential basic principle of the invention. Light **4** is led through the substrate **1** to the dye film **2**. The light can be a light beam or diffuse light. Reflection occurs from the interface between the dye film **2** and a solution **5** to be measured, such as process fluid. The reflection may be a partial or total reflection. Accordingly, the chemical property of the solution to be measured is measured by light reflection measurement, whereby only one optic window towards the process is needed. Consequently, the invention can be implemented in a rather simple manner.

The film mentioned above can preferably be fabricated of a solution synthesized by Sol-Gel method, described in more detail for instance in the book Sol-Gel Science, The Physics and Chemistry of Sol-Gel Processing, Academic Press, Inc. 1990. A Sol-Gel solution, i.e. a sol, is a solution which forms an inorganic polymer, glass, when drying on a glass surface. The glass sheet is coated for instance by immersing the sheet into the sol. The thickness and refractive index of the film are the most significant factors for the measuring optics. The thickness of the film is controlled by the viscosity of the sol and by immersion speed. The viscosity is changed by increasing or decreasing the amount of the solvent used and by changing the conditions of synthesis by means of a catalyst and the quantity of water, for example. Depending on the sol structure, the immersion speed has such an influence that, by slower immersion, a thinner film is produced, when the sol has a polymeric structure. If the sol has a particle structure, the film becomes thinner when the immersion speed increases. The refractive index of the film can be affected by the selection of sol precursor, i.e. the starting substance to be modified to metal oxide by other reagents, and by drying and filming conditions. The precursor forms an oxide, the refractive index of the oxide depending on the compound structure and metal atom. The refractive index of titanium oxide, for instance, is about 2, while the refractive index of silicon dioxide is about 1.5. Raising the drying temperature and making the film thinner raise the refractive index.

At Sol-Gel synthesis, as precursor serve in most cases metal alkoxides, which react easily with each other in the presence of a suitable catalyst or a gelatinating reagent and water. As catalyst can be used hydrochloric, nitric or sulphuric acid. Ammonia is also an often used catalyst. In turn, gelatinating reagents can be carboxylic acids, as an example of which acetic acid can be mentioned. The water can be added during the synthesis, it can be allowed to form through a reaction between a carboxylic acid and alcohol, or it can come from the humidity of air, the water reacting with the sol in the filming phase. By selection of catalyst, the polymerization and structure of the sol can be influenced. Alkali catalysts make a particulate sol, while acid catalysts make the sol polymeric. The quantity of water also affects the structure; if water is used in the molar ratio 2:1 with respect to an alkoxy mole and the catalyst is an acid, a polymeric sol is obtained.

Glass is the most generally used material in different optic components. Glass is often a stronger material than many plastics, which also are used relatively much in the optics at

present. In comparison to plastics, the thermal resistivity of glass is in a class of its own. In the examples of the figures, glass has been used, but it shall be noted that the invention is not restricted to the use of glass only, but suitable plastics can naturally be used in the same way as is generally done in optics nowadays.

As described above, reflection measurement is utilized in the invention. FIG. **1** shows the basic principle applicable in different ways. FIGS. **2** and **3** show a second and a third embodiment of the invention. In the examples of the FIGS. **2** and **3**, the essential thing is that the dye film **2** comprises a diffusely reflective surface disposed on it. In the case of FIG. **2**, the diffusely reflective surface is a coating **3** compounded with pigments, whereby light is reflected from pigment particals. In the case of FIG. **3**, the diffusely reflective surface is a rough surface **6**. The structure also comprises a cover film **7** or several films on the rough surface. The rough surface **6** can be the rough surface of the dye film, as shown in FIG. **3**, or the dye film surface can be smooth, but for instance the interface between the two following films rough, etc. Light **4** is reflected from the roughness of the interface diffusely. The layers as such are non-diffuse.

The pigment mentioned above can be either a substance added separately or pigment produced at the synthesis or in the manufacturing process. Instead of particles, small bubbles scattering light like particles are possible as well.

FIG. **4** shows an embodiment of the invention in which a dielectric mirror **8** is disposed on the dye film **2**, which mirror may comprise one or several layers **8***a*, **8***b*, **8***c*. . . . The dielectric mirror can be manufactured for instance of two materials having very differing refractive indexes. The film structure, i.e. pack film, can preferably be made to a multilayer structure, whereby the pack film should have at least three layers. It is especially preferable to form a pack in such a way that it alternately comprises a layer having a high refractive index and a layer having a low refractive index. A single layer of titanium oxide alone reflects about 20% of the incident light, but the reflection gets essentially better when layers are added. A five-layer pack provides a reflection of about 70% already.

The thickness of the above layers may have an influence on in which wavelength range the film is reflective. The pack may be designed for one or several wavelengths. If two light beams are used for the measurement such that there is a big difference between the wavelengths of the light beams to be used, i.e. measuring beams, it is possible to form even two pack films on each other, designed separately for each wavelength.

In the example of FIG. **4**, the dye film **2** is covered by a multilayer pack film, i.e. a dielectric mirror **8** composed of layers **8***a*, **8***b* and **8***c*. . . . The layers **8***a*, **8***b*, **8***c* . . . are arranged such that, next to the dye film **2**, there is a material **8***a* having a high refractive index, then a material **8***b* having a low refractive index and then a material having a high refractive index etc. Reflection naturally occurs also on the glass interfaces between air and glass sheet and between glass sheet and dye film in the travel direction of the beam. Interface reflection between the glass sheet **1** and the dye film **2** can be decreased by fabricating the dye film **2** in such a way that it has the same refractive index as the glass to be used. The interface reflection between air and glass sheet can be decreased by coating that side of the glass sheet **1** with a material having a lower refractive index than the glass.

The function of a reflective pack film is based on interference. Two electromagnetic waves can be subjected to a constructive or deconstructive interference, depending on the phase of the waves with respect to each other. At the

manufacture of a reflective film structure, the aim is to choose the thicknesses and refractive indexes in such a way that the interference will be constructive, whereby the reflected waves are in the same phase and the reflection is strong. If it is desirable to decrease the reflection, the thicknesses and refractive indexes are chosen such that the interference will be destructive, whereby the waves are in the opposite phases.

Layers disposed on the actual dye film can improve, besides reflective properties, also the mechanical resistance of the solution, by protecting the film against scratches and splits. In addition, they may lengthen the chemical duration of the film by decreasing the diffusion of dye molecules into the fluid to be measured. These layers can also be compounded with dye. As far as the porosity of the layers is concerned, the layers shall allow for instance the ions to be measured to pass through the film into contact with the dye of the dye film **2** and thus to react with it.

The above embodiments are by no means intended to restrict the invention, but the invention can be modified fully freely within the scope of the claims. It is thus clear that the arrangement of the invention or its details do not necessarily need to be just like shown in the figures, but solutions of another kind are also possible. Even if the examples of the figures present a glass sheet or the like, it is clear that this term shall be understood to cover, besides glass, also corresponding plastics and different parts; the glass sheet or the like can be a prism, for instance. The measuring beam or beams can be produced by means of any suitable light source. Any suitable indicator dyes can naturally be used in the dye film. Within the basic idea of the invention, layers on the dye film can be formed in any suitable manner, essential is only that the layers disposed on the dye film allow the OH and H ions of the solution to be measured to pass through and that the film simultaneously serves as a reflective surface, as described above. A diffusely reflective surface can be formed by means of different pigments, for instance. These materials can also be powdery additives, they can be produced from a reaction in accordance with the above, etc.

What is claimed is:

1. An arrangement at measurement of pH or another chemical property detectable by dye indicators, comprising:

a substrate to be immersed into a solution to be measured, the substrate being coated with a dye film arranged to change its color when a chemical property of the environment changes, wherein the dye film is covered by a dielectric mirror formed of several layers, allowing ions to pass through and reflecting light backwards, whereby a color change in the dye film can be measured as light reflection measurement.

2. The arrangement according to claim **1**, wherein the structure reflecting light backwards comprises a diffusely reflective surface.

3. The arrangement according to claim **2**, wherein the diffusely reflective surface is a coating compounded with pigments.

4. The arrangement according to claim **2**, wherein the diffusely reflective surface is a rough surface.

5. The arrangement according to claim **1**, wherein the layers are composed of two different materials having very differing refractive indexes.

6. The arrangement according to claim **5**, wherein the layers are arranged in such a way that, next to the dye film, there is a first material having a high refractive index, then a second material having a low refractive index.

7. The arrangement according to claim **1**, wherein the layers are arranged to reflect different wavelengths.

* * * * *

APPENDIX B Position sensor principle

This appendix shows the analysis of intensity signals associated with the position sensor used in chapter 6.

When a plane wave encounters a grating, it will be partially deflected. This deflection depends on the grating period and wavelength of the light. A simplified model of the grating consists of a large number of evenly spaced slits, where light can pass through the grating. These slits act as secondary light sources radiating light to all directions. In some directions the light will form plane waves in the far field as the light emitted from different slits interferes.

A grating has parallel slits with distance d. These slits are located so that the grating is in xz-plane and the slits are parallel to the z-axis. To simplify the calculations one of the slits goes through the origin. This assumption does not affect the generality of the calculations.

A plane wave E_1 comes to the interface and creates plane wave E_2 on the other side of the grating:

$$E_1 = e^{i(\omega t - \vec{k}_1 \cdot \vec{r})} \tag{B.1}$$

$$E_2 = e^{i(\omega t - \vec{k}_2 \cdot \vec{r})} \tag{B.2}$$

Only phases are taken into account in this calculation for simplicity.

If E_2 is created by E_1, the waves have to be in the same phase in each of the slits:

$$E_1(nd\vec{u}_x) = E_2(nd\vec{u}_x) \tag{B.3}$$

$$e^{ik_{1x}nd} = e^{ik_{2x}nd} \tag{B.4}$$

where n is an integer. This equation has to hold true for all n. k_{nx} denote the x components of the wave propagation vectors.

The phase shift between the two plane waves can then be either zero or an integral multiple of 2π:

$$k_{1x}nd = k_{2x}nd + mn2\pi \tag{B.5}$$

$$k_{1x} = k_{2x} + \frac{m2\pi}{d} \tag{B.6}$$

where m is an integer. The multiplier n has to be included in the phase shift term to make the equation valid for all n.

The absolute value of each wave number has to remain the same in all radiation, as there is no change in the medium or frequency. If the incident radiation E_1 is in angle α to the normal, and the transmitted wave in angle β, equation (B.1) can be written:

$$k\sin\alpha = k\sin\beta + \frac{m2\pi}{d} \tag{B.7}$$

$$\sin\alpha = \sin\beta + \frac{m2\pi}{kd} \tag{B.8}$$

$$\sin\alpha = \sin\beta + \frac{m\lambda}{d} \tag{B.9}$$

Naturally, the choice of m is limited so that the right hand side of the equation remains between minus and plus unity.

In the simplest case the incident angle is zero, and:

$$\sin\beta + \frac{m\lambda}{d} = 0 \tag{B.10}$$

$$\beta = -\sin^{-1}\frac{m\lambda}{d} \tag{B.11}$$

The position sensor is based on interference between the two first order ($m = \pm 1$) diffractions (see figure 6.3). When two waves interfere, the resulting field is a superposition of the two waves:

$$\vec{E} = \vec{E}_1 e^{i(\omega t - \vec{k}_1 \cdot \vec{r})} + \vec{E}_2 e^{i(\omega t - \vec{k}_2 \cdot \vec{r})} \tag{B.12}$$

The intensity of this field is given by:

$$I = \frac{1}{2}\vec{E}^* \cdot \vec{E} \tag{B.13}$$

where the asterisk is used to denote the complex conjugate. The impedance has been omitted for clarity (i.e., the unit of I is not physically correct in (B.13)).

For the first order diffraction the wave numbers k_{-1} and k_1 (representing $m = \pm 1$) are:

$$k_{-1} = \frac{2\pi}{\lambda}(-\sin\alpha\vec{u}_x + k_y\vec{u}_y + k_z\vec{u}_z) \tag{B.14}$$

$$k_{+1} = \frac{2\pi}{\lambda}(\sin\alpha\vec{u}_x + k_y\vec{u}_y + k_z\vec{u}_z) \tag{B.15}$$

where k_z and k_y are determined by the original direction of the incident radiation in the yz-plane. When (B.12), (B.14) and (B.13) are combined, the intensity is:

$$I = \frac{1}{2}\left[E_1^2 + E_2^2 + \vec{E}_1 \cdot \vec{E}_2\left(e^{-i(\vec{k}_1 \cdot \vec{r} - \vec{k}_2 \cdot \vec{r})} + e^{i(\vec{k}_1 \cdot \vec{r} - \vec{k}_2 \cdot \vec{r})}\right)\right] \tag{B.16}$$

$$= \frac{1}{2}\left[E_1^2 + E_2^2 + 2\vec{E}_1 \cdot \vec{E}_2 \cos\left((\vec{k}_1 - \vec{k}_2) \cdot \vec{r}\right)\right] \tag{B.17}$$

$$= \frac{1}{2}\left(E_1^2 + E_2^2 + 2\vec{E}_1 \cdot \vec{E}_2 \cos\frac{-4\pi x \sin\alpha}{\lambda}\right) \tag{B.18}$$

$$= \frac{1}{2}\left(E_1^2 + E_2^2 + 2\vec{E}_1 \cdot \vec{E}_2 \cos\frac{4\pi x}{d}\right) \tag{B.19}$$

The field is thus a constant intensity field with a sinusoidal component varying in the x direction. The period of one sinusoidal intensity variation is one half of the grating period.

The sensor is immune to misalignment in y and z directions. Rotational misalignment about the x-axis does not change the field pattern as the x component of the radiation is not changed. Rotational misalignment about the y-axis leaves the field intact but naturally rotates the pattern in respect to the sensor.

However, rotational error about z-axis does change the intensity pattern in the xy-plane. By looking at (B.9), it is clear that if the incident wave is non-normal ($\alpha \neq 0$), the deflected waves change their direction. If the sensor is rotated about z-axis, also the light source rotates, and hence the incident radiation is non-normal.

The angular change in the deflected beam is denoted by δ. Then (B.9) can be written:

$$\sin\alpha = \sin(\beta_0 + \delta) + \frac{m\lambda}{d} \tag{B.20}$$

$$\sin\alpha = \sin\beta_0\cos\delta + \cos\beta_0\sin\delta + \frac{m\lambda}{d} \tag{B.21}$$

where β_0 is the deflection angle at normal incidence ($\alpha = 0$).

In practical applications α is tried to be made as small as possible, so it can be assumed to be small enough for the approximation $\sin\alpha = \alpha$. Small changes in the

incoming radiation will produce small changes in the deflected waves, so δ can be assumed to be small, as well ($\sin\delta = \delta$, $\cos\delta = 1$):

$$\alpha = \sin\beta_0 + \delta\cos\beta_0 + \frac{m\lambda}{d} \tag{B.22}$$

$$\alpha = \delta\cos\beta_0 \tag{B.23}$$

$$\delta = \frac{\alpha}{\cos\beta_0} \tag{B.24}$$

Thus, if the angular error α is small, the deflected beams of the same order will tilt with equal amounts. This way the interference field tilts with similar angle. In practice, all angles associated in the process are quite small, as the grating period is long (20 µm) compared to the radiation wavelength (780 nm), which will make the diffractions angles small (first order diffraction is at $\approx 2.2°$).

Tilting of the field will make the sensor more sensitive to the y coordinate error. However, the sensitivity ratio between the x coordinate and the y coordinate is equal to the tangent of the field angle, so even with moderate errors in the z rotation the error produced by the travel in the y direction is small.

An important error arises from the zeroth order radiation, i.e. the plane wave which is not deflected by the grating. Even though the grating is designed so that it produces very little zeroth order, a small zeroth order wave is always present. So, the field becomes:

$$\vec{E} = \vec{E}_{-1}e^{i(\omega t - \vec{k}_{-1}\cdot\vec{r})} + \vec{E}_0 e^{i(\omega t - \vec{k}_0\cdot\vec{r})} + \vec{E}_{+1}e^{i(\omega t - \vec{k}_{+1}\cdot\vec{r})} \tag{B.25}$$

where $\vec{E}_0$ is the radiation advancing as the zeroth order.

The total intensity is then:

$$\begin{aligned} I = \frac{1}{2}\Big[& E_{-1}^2 + E_0^2 + E_{+1}^2 + \\ & 2\vec{E}_{-1}\cdot\vec{E}_0\cos(\vec{k}_{-1} - \vec{k}_0)\cdot\vec{r} + \\ & 2\vec{E}_{+1}\cdot\vec{E}_0\cos(\vec{k}_{+1} - \vec{k}_0)\cdot\vec{r} + \\ & 2\ \vec{E}_{+1}\cdot\vec{E}_{-1}\cos(\vec{k}_{+1} - \vec{k}_{-1})\cdot\vec{r}\Big] \end{aligned} \tag{B.26}$$

If the original wave is propagating at the normal of the grating, the wave vectors

are:

$$k_{-1} = \frac{2\pi}{\lambda}(-\sin\beta\vec{u}_x + \cos\beta\vec{u}_y) \tag{B.27}$$

$$k_0 = \frac{2\pi}{\lambda}\vec{u}_y \tag{B.28}$$

$$k_{+1} = \frac{2\pi}{\lambda}(\sin\beta\vec{u}_x + \cos\beta\vec{u}_y) \tag{B.29}$$

The intensity (B.26) can be written as:

$$\begin{aligned} I = &\frac{1}{2}\left\{E_{-1}^2 + E_0^2 + E_{+1}^2 + \right. \\ &2\vec{E}_{-1}\cdot\vec{E}_0\left[\cos(\vec{k}_{-1} - \vec{k}_0)\cdot\vec{r} + \cos(\vec{k}_{+1} - \vec{k}_0)\cdot\vec{r}\right] + \\ &2(\vec{E}_{-1}\cdot\vec{E}_0 - \vec{E}_{+1}\cdot\vec{E}_0)\cos(\vec{k}_{+1} - \vec{k}_0)\cdot\vec{r} + \\ &\left.2\vec{E}_{+1}\cdot\vec{E}_{-1}\cos(\vec{k}_{+1} - \vec{k}_{-1})\cdot\vec{r}\right\} \end{aligned} \tag{B.30}$$

In practice, the grating can be assumed to be symmetric. This implies:

$$\vec{E}_{-1}\cdot\vec{E}_0 = \vec{E}_{+1}\cdot\vec{E}_0 \tag{B.31}$$

Thus, the second last term in (B.30) is zero. The resulting intensity pattern resembles that of (B.19) but there is one extra term (the third last term):

$$2\vec{E}_{-1}\cdot\vec{E}_0\left[\cos(\vec{k}_{-1} - \vec{k}_0)\cdot\vec{r} + \cos(\vec{k}_{+1} - \vec{k}_0)\cdot\vec{r}\right] \tag{B.32}$$

The variable part of this term is:

$$\cos(\vec{k}_{-1} - \vec{k}_0)\cdot\vec{r} + \cos(\vec{k}_{+1} - \vec{k}_0)\cdot\vec{r} \tag{B.33}$$

$$= \cos\left(-x\frac{2\pi}{\lambda}\sin\beta + y\frac{2\pi}{\lambda}(\cos\beta - 1)\right)$$

$$+ \cos\left(x\frac{2\pi}{\lambda}\sin\beta + y\frac{2\pi}{\lambda}(\cos\beta - 1)\right) \tag{B.34}$$

$$= 2\cos\left(\frac{2\pi x}{\lambda}\sin\beta\right)\cos\left(\frac{2\pi y}{\lambda}(\cos\beta - 1)\right) \tag{B.35}$$

The angle β for the first order deflections is by (B.10):

$$\beta = \frac{\lambda}{d} \tag{B.36}$$

As β is small, $\cos\beta$ can be approximated by:

$$\cos\beta = \sqrt{1 - sin^2\beta} = \sqrt{1 - \left(\frac{\lambda}{d}\right)^2} \approx 1 - \frac{1}{2}\left(\frac{\lambda}{d}\right)^2 \tag{B.37}$$

and (B.35) can be written as:

$$2\cos\left(\frac{2\pi x}{d}\right)\cos\left[\frac{2\pi y}{\lambda}\left(-\frac{1}{2}\frac{\lambda^2}{d^2}\right)\right] \tag{B.38}$$

$$=2\cos\left(\frac{2\pi x}{d}\right)\cos\left(\frac{2\pi\lambda y}{d^2}\right) \tag{B.39}$$

This represents an intensity distribution which varies sinusoidally both in x and y directions. The variation in x direction has the same period as the grating (d), i.e. double that of the desired signal.

The intensity of this error signal depends on the y position, as well. This intensity variation along y-axis is not insignificant. With the real world values (λ = 780 nm, d = 20 μm) the period of the variation in y direction is approximately 500 μm, i.e. already a change of a few dozens of micrometers in the y position will change the amplitude of the zeroth order interference appreciably.

The sensor design partially compensates for this by using a periodic photosensitive element in detecting the intensity variations. The elements in the sensor are ordered with 2.5 μm pitch so that the adjacent elements will give $\pi/2$ phase difference. There are several elements, and the elements are connected so that every fourth element belong to the same group (figure B.1).

The configuration has several advantages. Larger number of elements will give larger measurement signal. As all four quadrature signals are available, sensor element biases can be compensated. Also, if the number of elements is an even multiple of four (i.e. a multiple of eight), the unwanted double-period signal is compensated. This is due to the fact that the signals given by every fourth element are in phase for the shorter period pattern but out of phase for the longer period pattern.

While this scheme compensates for most errors, it does not make the sensor immune to z rotation errors. In addition to the field tilting explained above the error interference from the zeroth order cannot be completely compensated if the sensor is not parallel to the grating.

Higher order diffractions will also produce interference with the first and zeroth order diffractions. However, if the distance between the grating and the sensor is long enough, this is not a problem due to geometrical reasons (figure B.2). Also, all higher order interference patterns are periodic at the grating period. So, even if there were

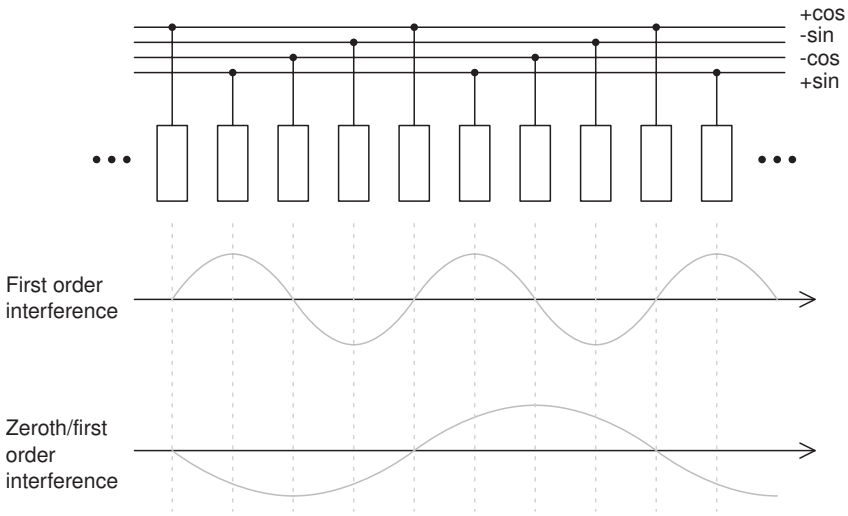


Figure B.1 Arrangement of sensor elements.

significant amounts of the higher orders in the signal, they can be compensated for with the method outlined in chapter 6.

The sensor used in this work is a reflective rather than a transmissive sensor. This does not change the principles described above but makes the sensor mounting easier. The grating is made of aluminized Zerodur glass. There are also transmissive sensors available with this measurement principle [51].

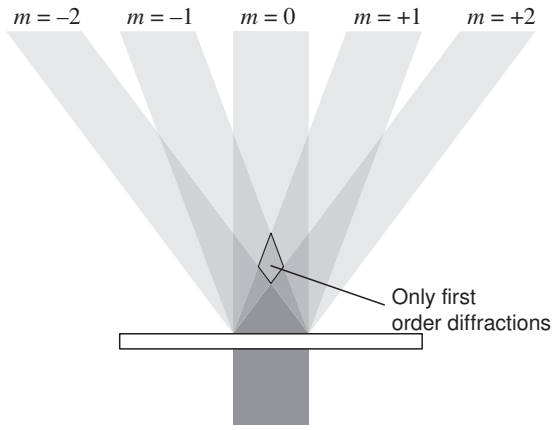


Figure B.2 Higher order diffractions are deflected out of the sensing area.