Applied Physics Thesis

Delft University of Technology

Breath analysis by quantum cascade laser spectroscopy

Author:
Olav Grouwstra

Direct supervisor:
A. Reyes Reyes
Supervisor:
N. Bhattacharya
Professor:
H.P. Urbach

January 30, 2015

Contents

1. Motivation for research

1.1 Gas analysis

1.1.1 Introduction: Why gas analysis?

Identification of constituent molecules of gasses is vital in practical applications such as medical diagnostics and environmental monitoring[7]. For medical diagnostics the general goal is to diagnose diseases with greater accuracy and ease. Many different ways of diagnosing the human body are available such as by blood sample or imaging techniques. An upcoming branch of diagnosis is that by analysis of human breath. Breath analysis has the benefit of being nonintrusive, thereby being much more easily implemented by medical professionals as nonintrusive medical instruments undergo less strict control by regulatory parties like the FDA and EMA.

1.1.2 How is gas analysis done?

Diagnosis by way of breath is performed by analysing metabolic molecules which diffuse from the bloodstream into the lungs and breath. Concentrations of these molecules result from certain chemical processes within the body, and as these chemical processes can be related back to diseases and general metabolism, the study of breath can be used as a diagnostic tool.

1.2 Quantum cascade laser spectroscopy

1.2.1 Why QCL spectroscopy?

One of the most common currently used techniques for gas analysis is gas chromatography – mass spectrometry (GC-MS). Although GC-MS achieves high accuracy, it has certain limitations. GC-MS is a specific test, meaning it requires knowledge of what to test for beforehand[8]. This means it is limited in use to only cases where symptoms are already showing, and thus requires pre-diagnosis. GC-MS also requires sample preparation, limiting its use to trained professionals. A gas analysis method using a quantum cascade laser for spectroscopy is treated in this paper. When properly developed this form of analysis should need little preparation and therefore be more easily accessible. A priori knowledge of what molecules to test for is also not necessary for the measurement itself,

which makes it a promising tool for laying statistical relations between concentrations of molecules and the health status of individuals.

1.2.2 Background to the research

A spectroscopy setup using a quantum cascade laser (QCL) is described and the subsequent data processing is extensively treated. The setup is made and extensively treated by A. Reyes Reyes as his PhD project funded by Fundamenteel Onderzoek Materie (FOM) foundation[11]. Z. Hou worked on the setup under supervision of A. Reyes Reyes and has also treated it extensively[8]. Measurements processed in this research are from breath samples of 35 healthy volunteers, and 35 asthmatic volunteers. Two different breath samples are used from each volunteer. These samples are from children from the Sophia Children's hospital in Rotterdam, all with their and/or their parents' informed consent. The goal of this work is twofold. Firstly to decrease the uncertainties with which the compounds and its concentrations are determined. Secondly the goal is to find a reliable way of distinguishing and identifying breath of healthy and asthmatic children by way of the compounds present and their concentrations, or otherwise.

2. Quantum Cascade Laser Spectroscopy Measurements

Measurements on the gas are performed with a spectroscopy setup using a quantum cascade laser. The various components and conditions of the measurement setup that influence the determination of the gas compounds and concentrations are explored. In order to understand how these components influence the eventual concentration, the path of light through the components is described starting from the quantum cascade laser (QCL) and ending at the CCD camera. For a more in depth analysis and reproducibility of the setup, refer to A. Reyes Reyes[11].

2.1 Obtaining concentrations from light

In order to obtain information on the gas through spectroscopy, the interaction between light and matter needs to be established. The relation of the concentrations of the compounds in the gas to the light passing through is given by the Beer-Lambert law:

$$I_o(\nu, C) = I_i(\nu) 10^{-A(\nu,C)} \tag{2.1}$$

with

$$A(\nu, C) = \sum_{\substack{c_{mol} \in C \\ \epsilon_{mol}(\nu) \in E(\nu)}} \epsilon_{mol}(\nu) c_{mol} l_{path} \tag{2.2}$$

where C and $E(\nu)$ denote the set of all concentrations and the set of all molar absorptivities respectively of the molecules in the gas. The Beer-Lambert law states a relation between light with intensity I_i entering a gas, its absorbance $A(\nu)$ by the gas, and its intensity I_o as it exits the gas. The absorbance is a function of the interaction length l_{path} of the light with the gas, and the concentration c_{mol} and the molar absorptivity $\epsilon(\nu)$ of the different molecules in the gas. The Beer-Lambert law shows that in the setup the interaction length and the intensity of the light going in and coming out of the gas must be measured.

2.2 Travel path of light

To find the concentration of compounds within a gas the spectroscopy setup as seen in Figure 2.1 is used.

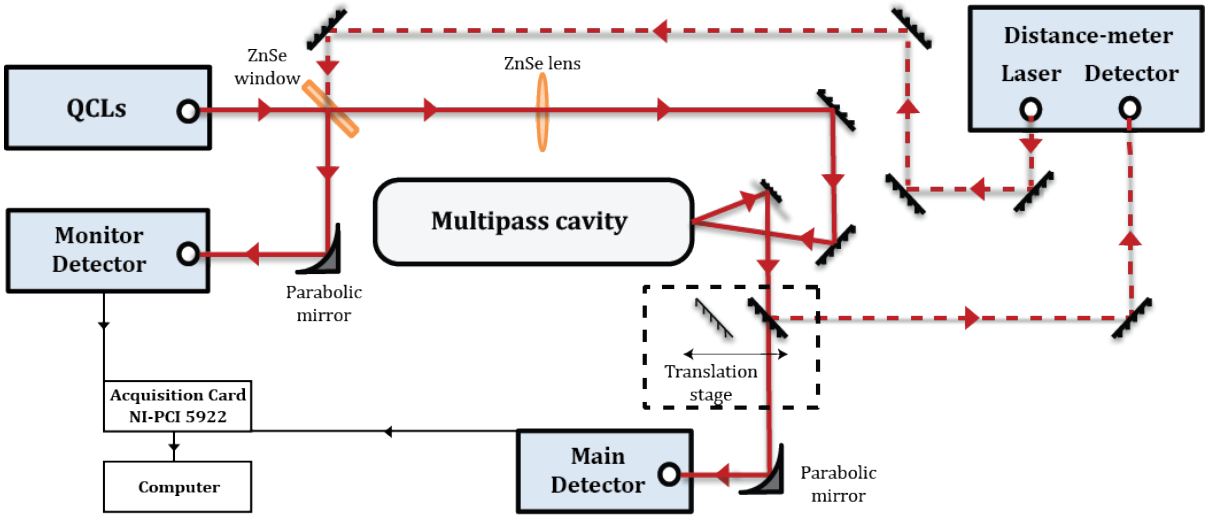


Figure 2.1: Diagram of the quantum cascade laser spectroscopy setup as used for measuring. Courtesy of Z.Hou[8]

2.2.1 Inside the QCL

By way of its design the QCL emits different wavenumbers of light simultaneously. The CCD cameras detect the intensity of the light without differentiating between the various wavenumbers. Since intensity per wavenumber is a necessity, a diffraction grating mounted on a piëzo is used to select a single wavenumber. The wavenumber allowed depends on the effective spacing of the grating, which is determined by the actual spacing and the incident angle of the light on the grating. The incident angle is controlled by applying a voltage over the piëzo. Two QCL's are arranged such that together they can scan over the wavenumber range of 832 cm^{-1} to 1263 cm^{-1}. In chapter 3 the measured signal and its noise is shown, most of which can be assigned to the hysteresis occurring in the piëzo element[11].

2.2.2 Outside the QCL

From the laser the light is branched using a beam splitter with one beam going to a detector, and the other send through the cavity in which the gas to examine sits. Part of the light is absorbed in the cavity as according to the absorbance profiles of the molecules constituting the gas. The remaining light exits the cavity and is measured at the next detector. These two detectors measure the light entering the gas I_i and the light exiting the gas I_o as necessary to determine the total absorbance in accordance with Equation 2.1. The length of interaction of the light with the gas l_{path} is measured using a laser distance meter[8].

3. Data Calibration

3.1 Wavenumber calibration

3.1.1 Four measurement signals

Multiple measurements are performed per sample using the two detectors as shown in the Figure 2.1. In order to get information on the gas, measurements are made using the main detector when a gas is inside the cavity. In order to reduce any effects due to the different paths to the detectors, measurements for the main detector are also done without gas in the cavity. This measurement without gas will henceforth be referred to as a reference measurement. Due to effects in the QCL, the signals also seem to vary between different measurements. Therefore both gas and reference measurements are also made with the monitor detector. The monitor signals are necessary to be able to reduce the effects of the QCL, as the QCL affects the signal of the monitor and the main detector similarly and can therefore be identified.

Table 3.1: Shows the different contributions to the measured signals. Wavenumber dependence is indicated by ν and ν‘, which show that different measurements can not be compared wavenumber-wise. This is a result from the uncertainty caused by effects in the QCL

	gas measurement	Reference
Main detector	$I_{QCL}(\nu`)f_{path1}f_{gas}(\nu`, C)$	$I_{QCL}(\nu)f_{path1}$
Monitor detector	$I_{QCL}(\nu`)f_{path2}$	$I_{QCL}(\nu)f_{path2}$

3.1.2 Pre-calibration

Ten measurements are made for both the monitor and main detector, for both gas and reference measurements. The measurements straight from the detector is very large resolution and too large to adequately save to disk. The processing is therefore largely done before the signal is saved, as described by Z. Hou[8]. The resulting measurements have a resolution of 0.05 cm^{-1}. A single of these processed signals looks as shown in Figure 3.1.

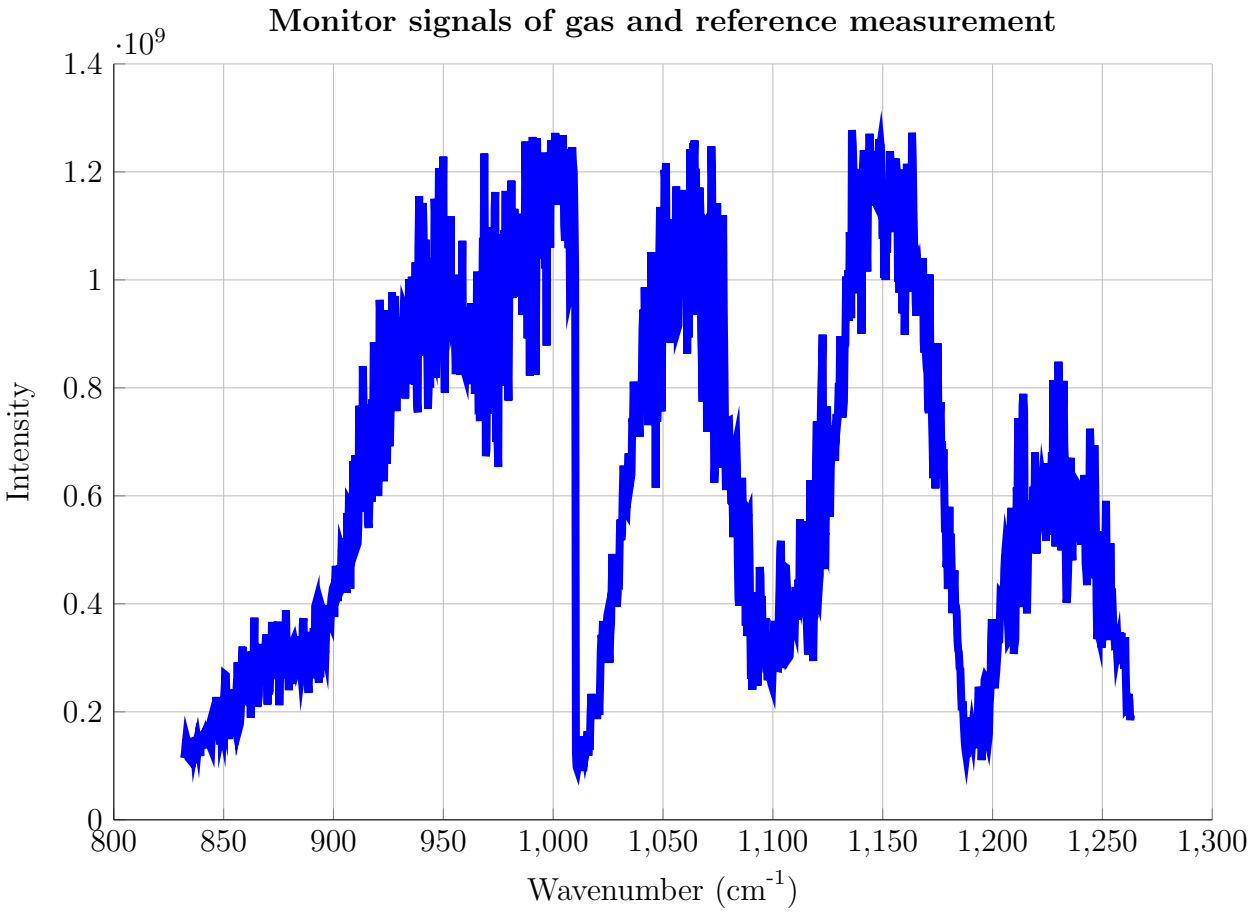


Figure 3.1: A signal as processed as described by Z. Hou[8]. The entire spectrum spans from 832 cm^{-1} to 1263 cm^{-1}. From Table 3.1 it can be seen that this monitor signal consists of the intensity from the laser times a factor caused by the path travelled. Since the influence from the path is quite small, the general shape of this signal portrays the intensity spectrum the laser outputs. The spectrum shown is down-sampled from 0.05 cm^{-1} resolution to about 0.35 cm^{-1} for amenable file size.

3.1.3 Need for additional calibration

Without any abnormalities the ten measurements of each signal would be averaged, and then the absorbance would be obtained by inverting Equation 2.1 such as:

$$A(\nu, C) = -\log_{10} \frac{I_o(\nu, C)}{I_i(\nu)}, \tag{3.1}$$

and then using the measured signals as in Table 3.1 for the following substitutions:

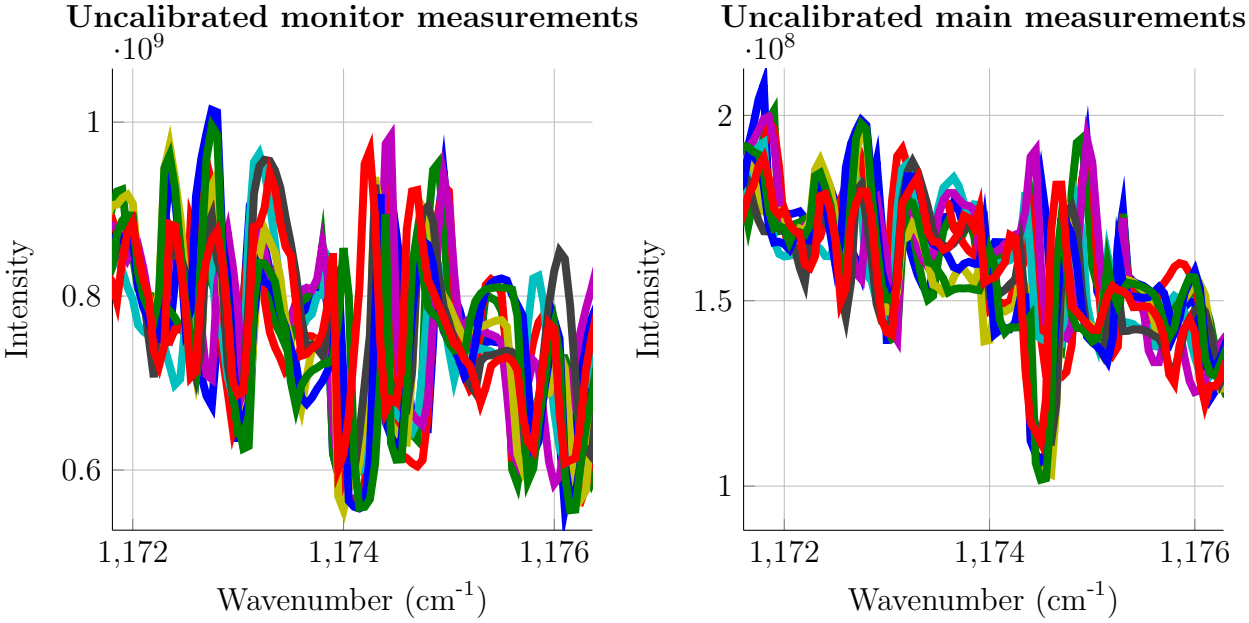
$$I_o(\nu, \nu', C) = \frac{I_{QCL}(\nu') f_{path1} f_{gas}(\nu', C)}{I_{QCL}(\nu) f_{path1}} \tag{3.2a}$$

$$I_i(\nu, \nu') = \frac{I_{QCL}(\nu') f_{path2}}{I_{QCL}(\nu) f_{path2}}, \tag{3.2b}$$

in order to get the absorbance $A(\nu, C)$ in terms of the gas contribution only:

$$A(\nu', C) = -\log_{10} f_{gas}(\nu', C). \tag{3.3}$$

This however requires the division of signals with an unknown relation between their wavenumbers ν and ν', and therefore doesn't work. When looking closely at Figure 3.2a and Figure 3.2b, this wavenumber discrepancy is seen as the measurements do not overlap.



(a) Uncalibrated measurements from the monitor detector.

(b) Uncalibrated measurements from the main detector.

Figure 3.2: Ten uncalibrated intensity measurements of a single sample. Five of these are from measurements with a gas in the cavity, and five are from a reference measurement.

The measurements exhibit an uncertainty in the wavenumber. This uncertainty is due to the hysteresis occurring in the piëzo on which the grating is mounted2.1. This local horizontal shift from one measurement to another results in an increase in standard deviation of the eventual absorbance and concentration. The absorbance of a single sample is calculated from multiple measurements with gas inside the cavity and the reference measurements without gas in the cavity. All these measurements need to be aligned to the same wavenumber values.

3.1.4 Wavenumber calibration

Every measurement of a sample consists of a main and a monitor part, of which the monitor is nearly identical for all measurements since it doesn't enter the cavity. This is where the primary use of the monitor signal comes in: it serves as a good choice for

a basis of calibration. Calibration is done towards the mean per wavenumber of all the monitor signal measurements: those with a filled cavity and with an empty cavity. The positions of the peaks and valleys of this mean are then determined.

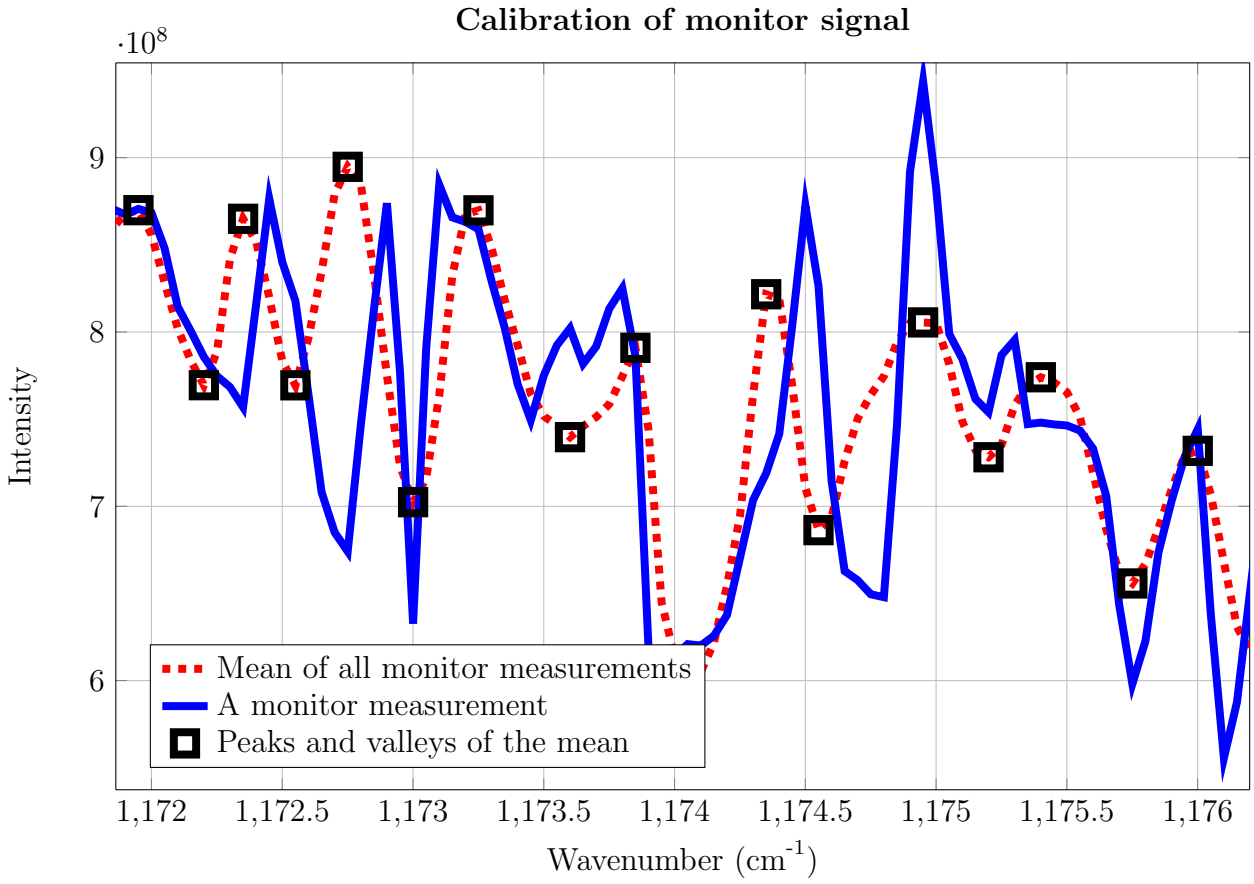


Figure 3.3: A single monitor signal is shown together with the mean of all monitor signals. Calibration is to be done towards the marked peaks and valleys of the mean.

For every monitor signal, the maximum value between two positions of valleys of the mean is taken. This maximum value is then taken as that signal's peak corresponding to the peak between the valleys of the mean mentioned. This peak of the monitor signal is then shifted towards the corresponding peak of the mean of all the monitors. This process is repeated for all valleys of the mean. The valleys of the monitor signal are determined in a similar fashion. Every signal is now locally shifted so its peaks and valleys match that of the mean signal. All data between peaks and valleys is stretched and compressed by means of resampling to fit the adjusted peaks and valleys. The result of this procedure performed on the monitor signal in Figure 3.3 can be seen in Figure 3.4.

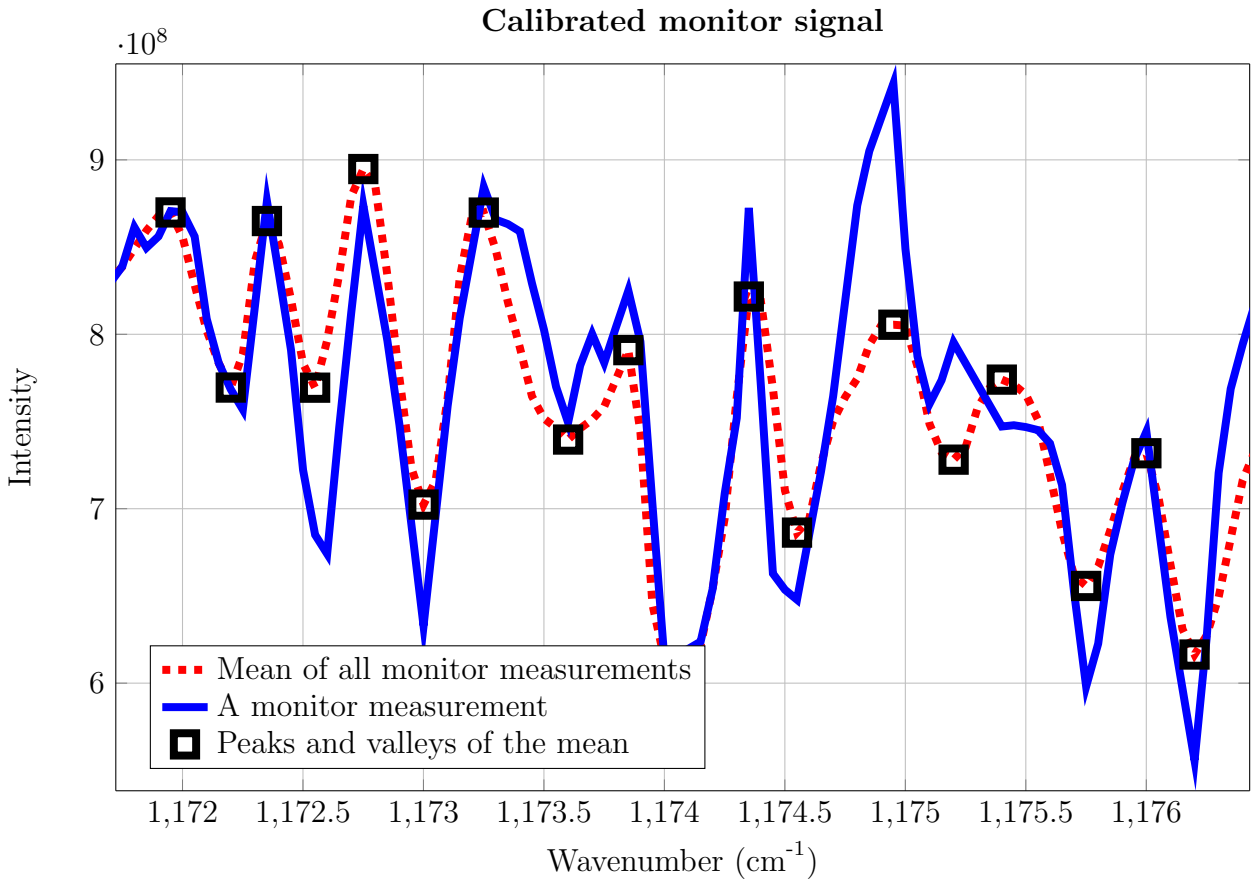
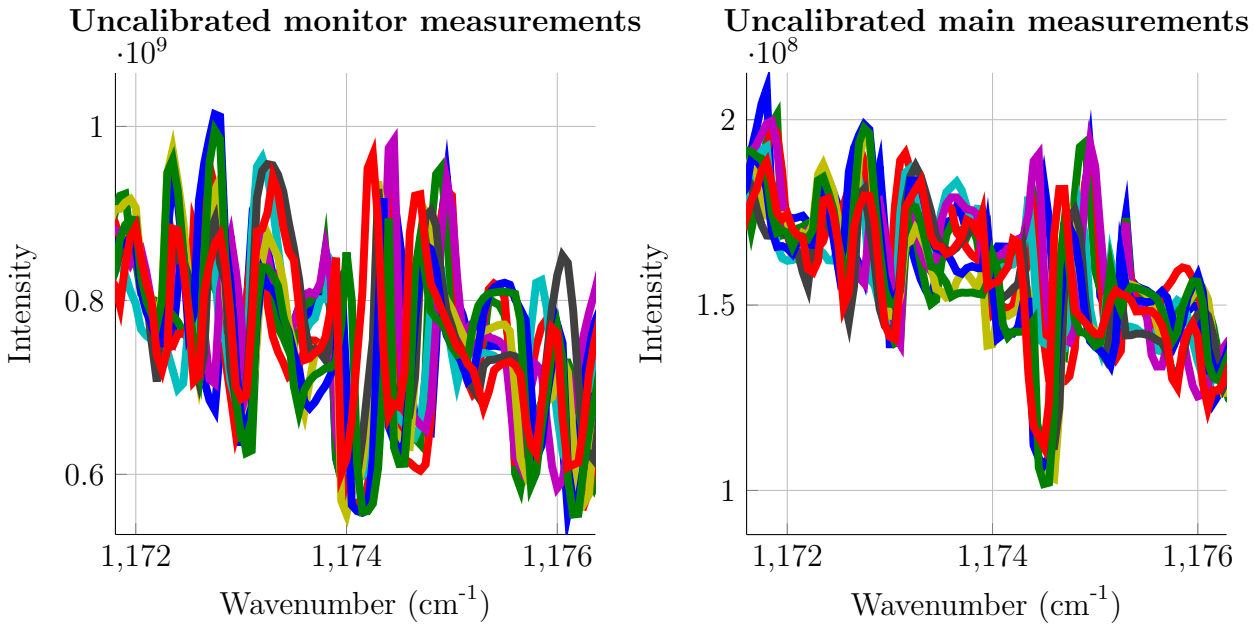


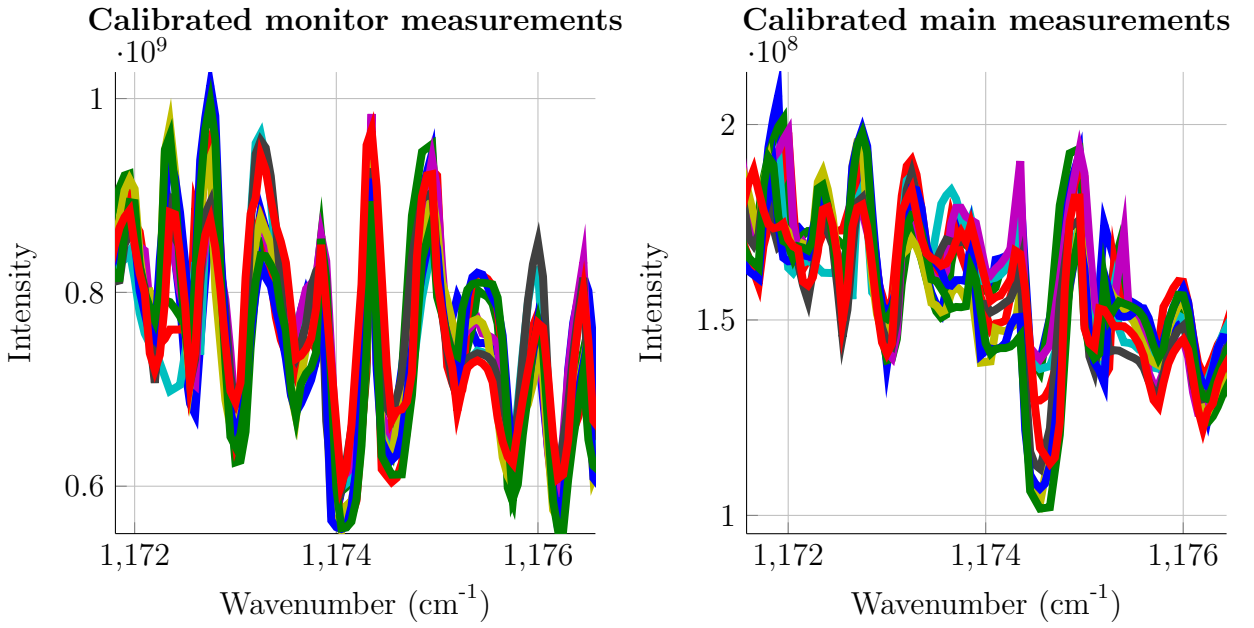
Figure 3.4: A single monitor signal is shown together with the mean of all monitor signals. Calibration is done towards the marked peaks and valleys of the mean.

The main signals of the reference and gas measurement are quite different, and hence calibration towards their mean cannot be applied to them. Since the monitor and main signal of a single measurement suffer from the same piëzo hysteresis effects, the transformation that calibrates the monitors also calibrates their respective main signals. Figure 3.5 demonstrates how this procedure calibrates all measurements of a sample towards each other.



(a) Uncalibrated measurements from the monitor detector.

(b) Uncalibrated measurements from the main detector.



(c) Calibrated measurements from the monitor detector.

(d) Calibrated measurements from the main detector.

Figure 3.5: Ten intensity measurements of a single sample. Five of these are from measurements with a gas in the cavity, and five are from a reference measurement. Figure 3.5c and Figure 3.5d are the wavenumber calibrated versions of Figure 3.5a and Figure 3.5b respectively.

Since all signals as given in Table 3.1 are now properly aligned, the absorbance can be calculated as according to Equation 3.1, Equation 3.2, and Equation 3.3. Absorbances are calculated this way for breath samples of 70 healthy and 70 asthmatic samples.

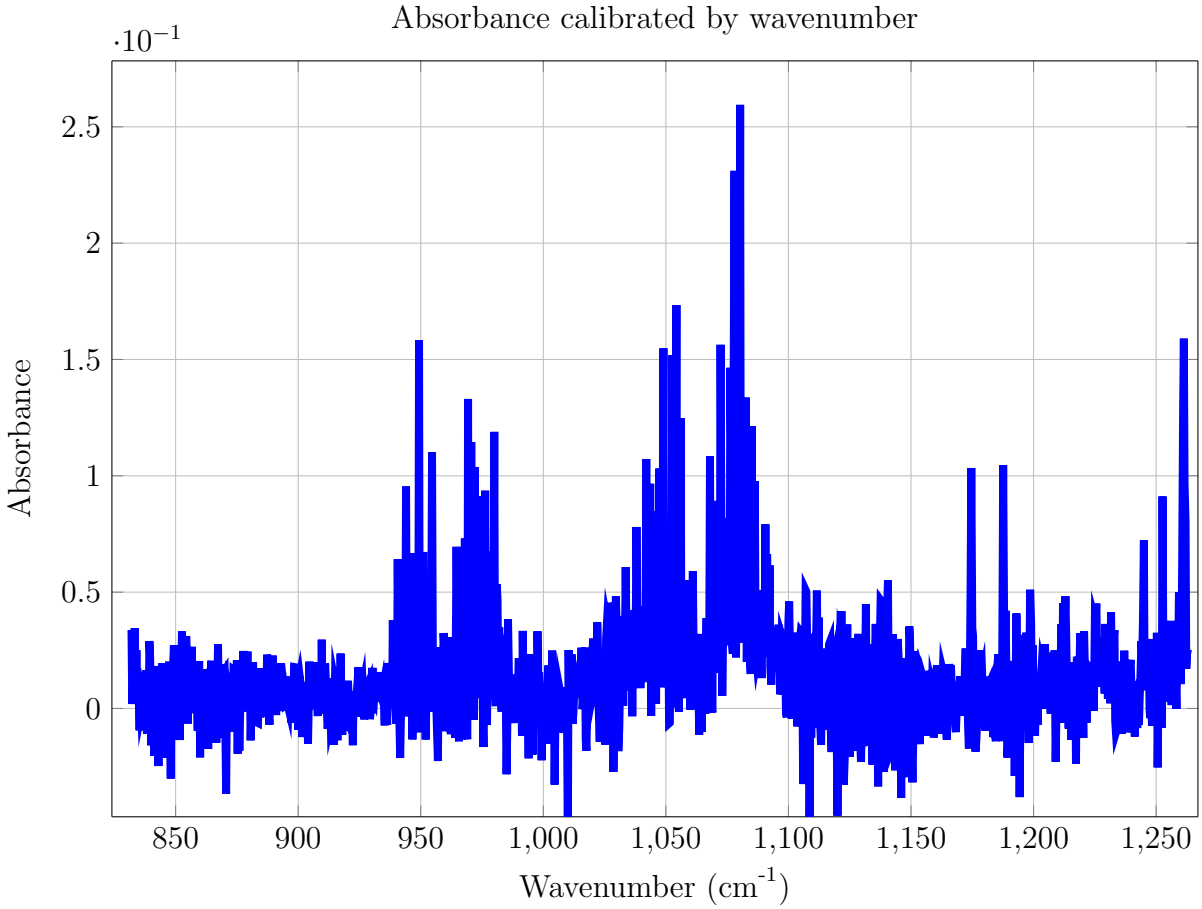


Figure 3.6: The wavenumber calibrated absorbance as calculated from the measurements of a single sample. The spectrum shown is down-sampled from 0.05 cm^{-1} resolution to about 0.3 cm^{-1} for amenable file size.

3.2 CO_2 and H_2O calibration

3.2.1 Motivation for CO_2 and H_2O calibration

When looking at the absorbance spectrum of Figure 3.6 there are a few regions that stand out by their relatively high absorbance. From Equation 3.1 it can be seen that high molar absorptivities $\epsilon(\nu)$ in a region and high concentrations of molecules with such molar absorptivities contribute to a high local absorbance. When checking for molecules with known high concentrations in breath such as CO_2, H_2O, and N_2, a cause of the high absorbance regions is easily identified.

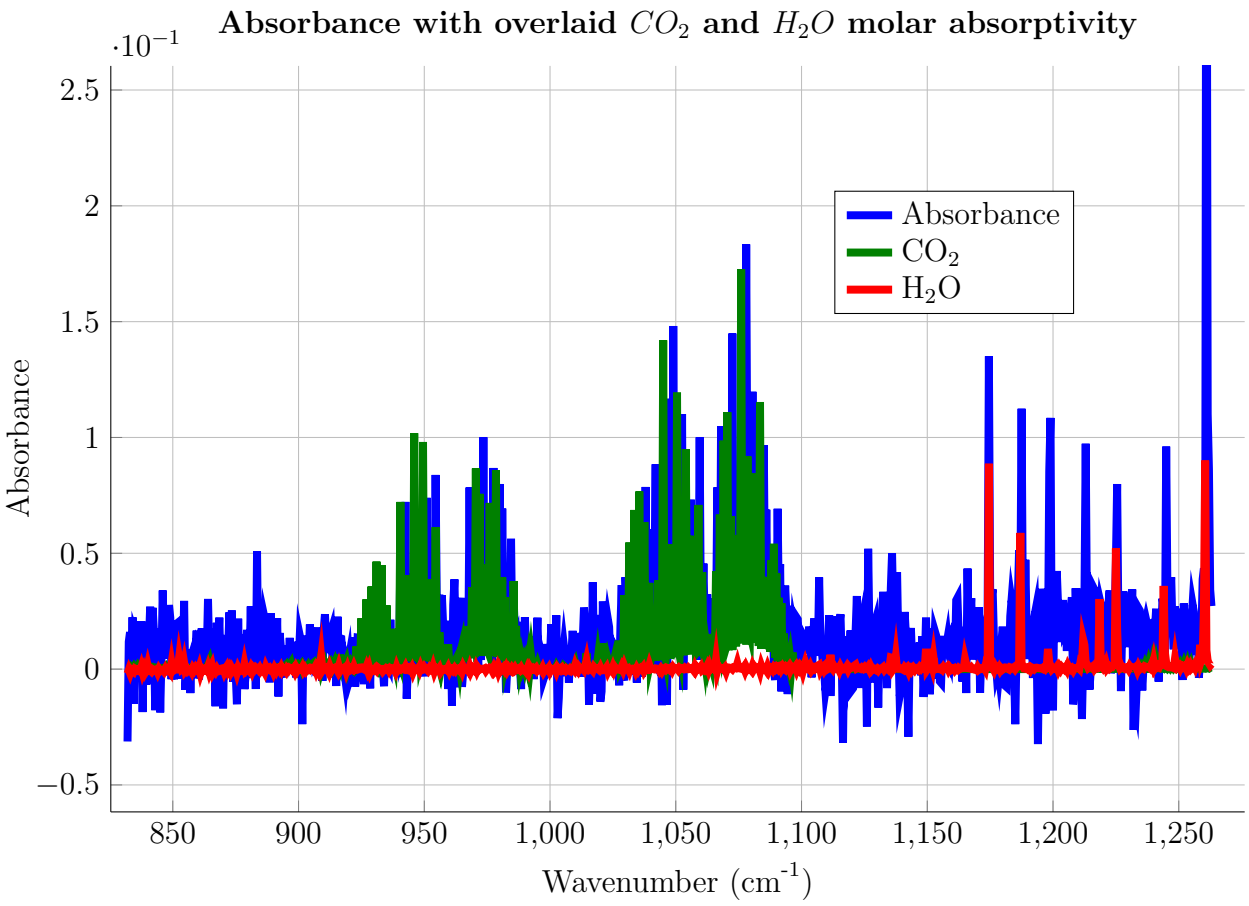


Figure 3.7: Absorbance from wavenumber calibrated data. The molar absorptivity overlays of CO_2 and H_2O are scaled with factors 60 and $5 \cdot 10^4$ respectively to clearly show the matching regions. CO_2 and H_2O molar absorptivities are taken from HITRAN database[2]. The three spectra shown are down-sampled to 0.5 cm^{-1} for more amenable file size.

When looking more closely at these matching regions between the absorbance spectrum and CO_2 and H_2O, it is seen in Figure 3.8 that even after the wavenumber calibration the absorbance spectrum doesn't match CO_2 and H_2O from literature.

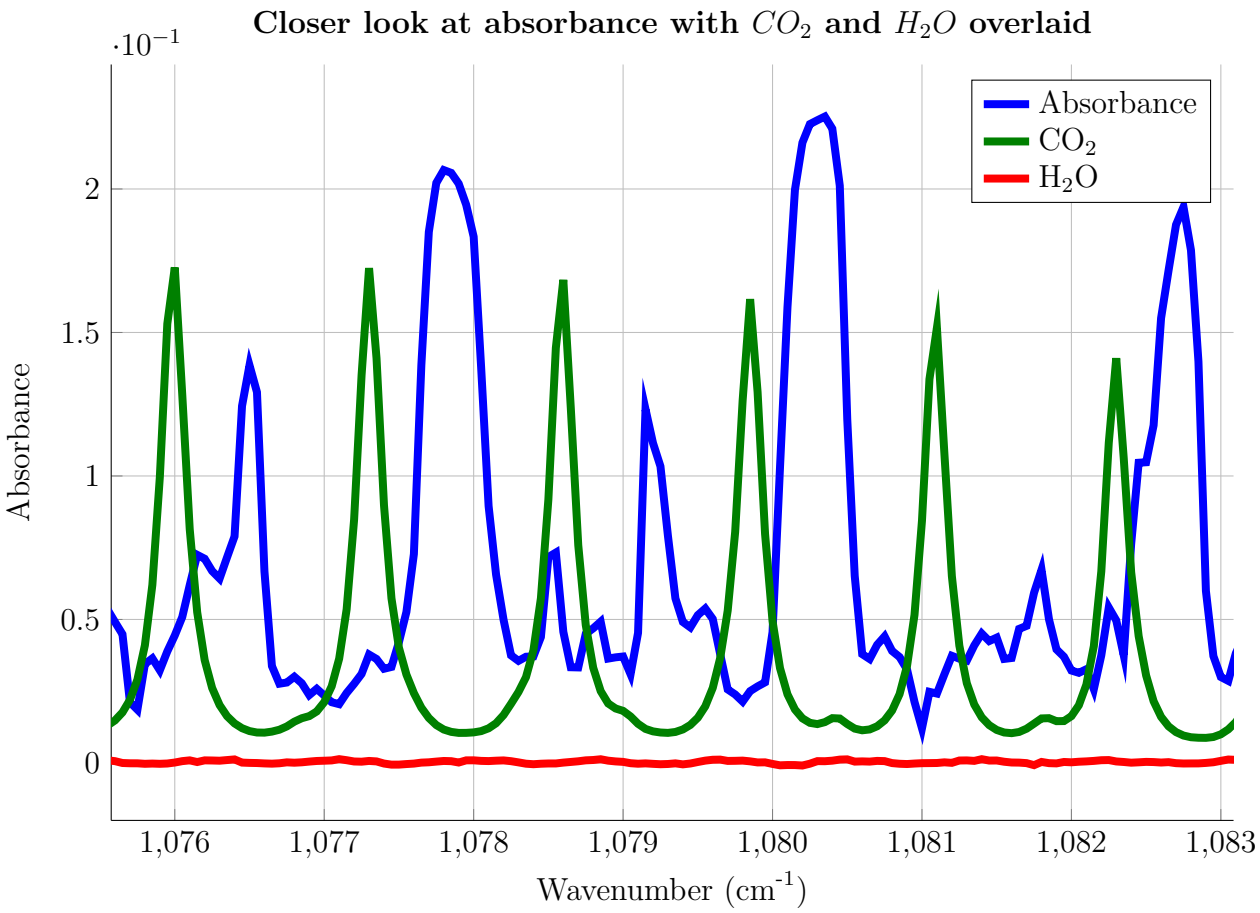


Figure 3.8: Absorbance from wavenumber calibrated data with molar absorptivity overlays of CO_2 and H_2O scaled by 60 and $5 \cdot 10^4$ respectively. H_2O is nearly absent since H_2O absorbs little in this region. CO_2 and H_2O molar absorptivities are taken from HITRAN database[2].

A method similar to the wavenumber calibration is used in order to calibrate the absorbance spectrum to CO_2 and H_2O.

3.2.2 Calibration to CO_2 and H_2O

Input and preprocessing

This method uses the absorbance spectrum as obtained from the wavenumber calibration, the accompanying wavenumber range, and molar absorptivity spectra from CO_2 and H_2O. Since the method used for the calibration to CO_2 and to H_2O are almost identical, only the CO_2 case is described here. The CO_2 spectrum is taken from HITRAN database[2], and interpolated to the range of 832 cm^{-1} to 1263 cm^{-1} and resolution of 0.05 cm^{-1}. Since HITRAN has its data ordered from high to low wavenumbers, the spectrum is checked and flipped if necessary to make sure they are in the same order as the measured data.

CO_2 peak and absorbance peak identification

To facilitate the calibration of the measured data to the CO_2 data, the peaks and valleys of the CO_2 spectrum serve as reference points. The peaks and valleys of the CO_2 spectrum are identified by using a peakfinder[3] algorithm. The peakfinder algorithm finds peaks above a certain adjustable threshold, based on its value compared to the surrounding data points. The peakfinder algorithm then returns the indices and intensities of all data points that it identifies. In case the first and last peak are before and respectively after the first and last identified valley, the first and last peak in the spectrum are removed. This way the range to adjusted always starts and ends with a valley of CO_2.

The first absorbance peak is identified by taking the maximum absorbance between the first two valleys of CO_2. The absorbance valley after this peak is identified by the minimum between the last identified absorbance peak and the next peak of CO2. The next valleys/peaks of the measurement are found by looking between the last peak/valley found and the next peak/valley from CO2, alternating between finding a peak and a valley. This process is repeated to the last identified valley of CO_2. This way all absorbance valleys are found between two peaks, and all absorbance peaks are found between two valleys. Thus about the same amount of absorbance peaks and valleys are identified as there are CO_2 peaks and valleys, making it easy for matching them[1].

Peak calibration

All the peaks and valleys of CO_2 are now put in a single list, and ordered by the index. The same is done for the peaks and valleys of the measurement. The peaks of the measurement list are now matched to coincide to the peaks of CO_2, and all the measurement data in between two peaks is interpolated to fit between the surrounding CO_2peaks. Since the algorithm messes up the region between the two major CO_2 region, this middle region is overwritten with the absorbance from that region before CO_2 calibration. This same procedure is applied to H_2O. The result of these procedures is seen in Figure 3.9.

[1]Since there is one more CO_2 valley than CO_2 peak, the amount of found absorbance peaks matches CO_2 peaks. However, two less absorbance valleys are found than there are CO_2 valleys. Hence the first and last CO_2 valley are removed.

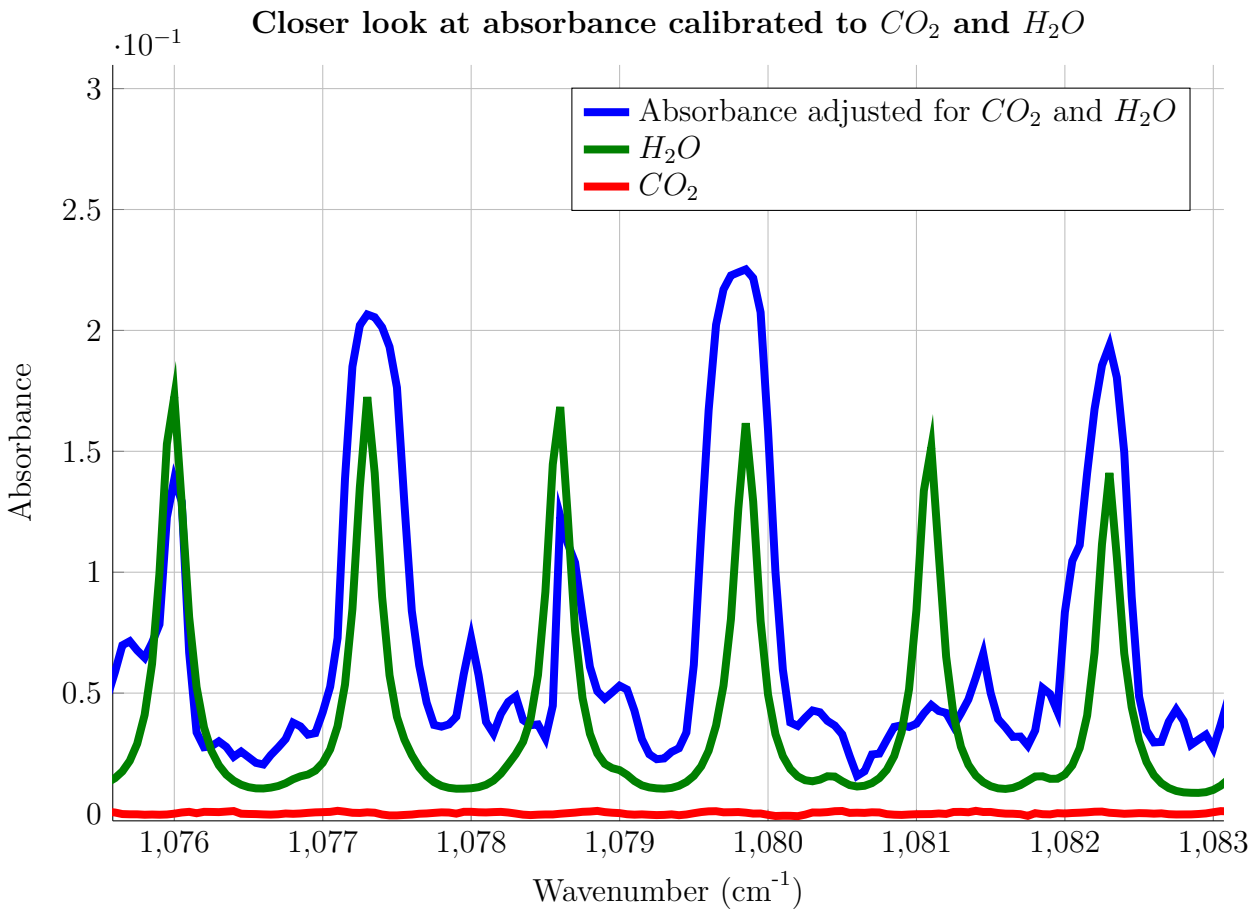


Figure 3.9: Absorbance calibrated to CO_2 and H_2O. H_2O is nearly absent since H_2O absorbs little in this region. Molar absorptivity overlays of CO_2 and H_2O are scaled by 60 and $5 \cdot 10^4$ respectively. CO_2 and H_2O molar absorptivities are taken from HITRAN database[2].

4. Molecules and Concentrations

4.1 Determination of molecules

4.1.1 Motivation for selecting molecules

In order to get the concentrations of molecules present in the gas, the measured absorbance of the sample is compared with absorbances of individual molecules. Absorbances of these individual molecules are calculated according to Equation 2.2 from the molar absorptivities. These molar absorptivities are taken from the Northwest Infrared database (NWIR)[5] and the high-resolution transmission molecular absorption database (HITRAN)[2], which are put together by Pacific Northwest National Laboratory (PNNL) and Harvard-Smithsonian Center for Astrophysics respectively. Currently these databases provide information on a total of 456 different molecules as listed in section C.1. Not all of these molecules are expected to be present in breath, since a lot of them are neither organic nor known to be in the atmosphere in an amount distinguishable by the current setup. Seeing as the absorbance still has uncertainty after the calibration, the concentrations to be determined cannot be without uncertainty. As a result the concentrations of non-present molecules will not necessarily be set to zero, and will therefore influence the concentrations determined for molecules that are present in measured gas. Thus before determining the concentrations it is vital to get a list of molecules to test against which excludes these excessive molecules. For the goal of distinguishing between the healthy and asthmatic groups, the list should also contain molecules likely to have large differences in concentration between the groups. This list is assembled in two steps: first by determining the molecules that are almost certainly present in breath, and second by selecting molecules likely to show different concentrations on the basis of analysis of p-values between the data sets. In the case when one is not interested in finding differences between data sets, and hence the following analysis based on comparing the data sets cannot be performed, the list of standard molecules should be extended in another way. This can be done by selecting molecules based on their absorptivity and a set minimal absorbance. Since the method based on comparing data sets filters on minimum absorbance and more specific wavenumber ranges, it is the preferred method when possible.

4.1.2 Standard molecules

This first group of molecules –the ones almost certainly present in breath– is a combination of atmospheric molecules[1] and molecules known to be in breath[6] with high absorbance in the relevant wavenumber region. Molecules can also be chosen by firstly obtaining the concentration as described in section 4.2 for a large list of molecules, and then picking out the most prominent ones. Molecules such as CO_2, H_2O, ethanol, and acetone are the major contributors to the absorption spectrum of the gas. These molecules must therefore be included in the list regardless of whether they are likely to have differences in concentration between the healthy and the asthmatic group. The list of molecules chosen this way can be found in section A.1.

4.1.3 Molecules selected to distinguish between data sets

To determine whether a gas was exhaled by someone with asthma or by someone without asthma, a recognizable difference in measurement results is desirable. P-values are used to quantify this difference between the measurements from healthy people and asthmatic people.

P-values

The p-value is defined as the probability of obtaining a certain result or more under the assumption of a certain hypothesis. In the context of these breath measurements the hypothesis is that the healthy and the asthmatic group have different concentrations of molecules, and thus most likely different absorption profiles. Thereby for each wavenumber the combined absorbances of multiple samples of healthy and asthmatic people are expected to be two groups of points spread around a different absorbance value. These two spreads can be modeled with different underlying probability distributions. The p-value is found by taking a distribution, such as the normal distribution, and fitting it to both the samples of the healthy group and the samples of the asthmatic group. Subtracting the overlap between these two distributions from one gives the p-value. The obtained p-value can be interpreted as the conditional probability of a randomly picked sample belonging distinguishably to one group or the other under the assumption that the samples in the groups have different and separate underlying probability distributions. The p-values between the 70 absorbances of healthy patients on the one hand and the 70 absorbances from asthmatic patients on the other hand are obtained this way for every wavenumber. This is done using an algorithm from E. Mastrigt[10] that uses the topTable function from the LIMMA package in *R*.

P-region analysis

In p-region analysis the compounds most likely to have different concentrations between the two data sets are sought. This is done by checking the p-values corresponding to the two data sets, and finding regions of consecutive low p-values. The reasoning behind this is as follows: organic molecules such as the ones expected in human breath are longer and more complicated than for example H_2O and CO_2, and thereby have a spectrum

consisting of many more peaks. These peaks can be so closely packed that they form a broad intensity band. If the concentration of such a molecule differs greatly from one data set to the other, this would be clear by low p-values in a broad band. Thus to find these molecules, the p-values corresponding to two data sets need to be below a certain threshold, for a certain region of consecutive wavenumber values. For these so-called p-regions, a list of compounds is checked for their absorbance. Compounds with a mean absorptivity above a certain threshold within any of these regions are likely candidates for the difference in the measured intensity between two sample groups, and are listed to be checked for their concentration in the data set. This list includes not only the molecules, but also the amount of regions in which they are present. Two p-regions are found using a minimal of 10 consecutive data points with p-values of 0.05 or higher. The complete set of compounds as given in section C.1 is checked for mean molar absorptivity above $10^{-4} ppm^{-1} m^{-1}$ in these regions. A table with the regions and a list of molecules found this way can be seen in section A.2.

4.1.4 Selection based on measurable concentration

The list of molecules can be shortened by pre-emptively checking their absorbances. Since molecules with a very low absorbance in the wavenumber region of interest will not be accurately estimated at the current noise level, these molecules can be eliminated. It is known from Z. Hou[8, p. 71] that the sensitivity of the system is in the order of 1 ppmv. This sensitivity is then used to establish a lower limit on the concentration for all molecules. In this case the chosen limit is even lower at 0.001 ppmv to make sure no molecules are eliminated unnecessarily. Absorbances are calculated for all molecules from this limit in concentration and their molar absorptivities. From an initial concentration estimation using only the standard molecules, a lower limit for the absorbance is set to $50 \cdot 10^{-7}$. For all molecules the absorbance calculated using the lower limit of concentration is then checked against this lower limit of absorbance. Molecules with absorbance that does not reach above the lower limit of absorbance at any point are dismissed, and the other molecules can be put on the list and further checked for their concentrations. A list of 199 molecules that should be present in breath is checked this way[6]. When the entire spectrum is checked this way, some 130 molecules pass. Adapting this method for the wavelength range of 1020 to 1100 cm^{-1}, where one of the major CO_2 regions is found, some 69 molecules pass as found in section B.1.

4.2 Concentration determination

4.2.1 method

Henceforth only the list of molecules to be checked for their concentration, as assembled according to methods mentioned in section 4.1 are used. At this stage the strictness of the list can be adjusted by only allowing molecules present in a certain amount or more of the previously determined regions of low p-values to be processed further. In order to process the molecules further and compare them to the measured data, the information

of the molecules are first transcribed to match the format of the measured data, i.e. in wavenumber range and resolution. This comes down to scrapping all the information outside of the relevant wavelength region, and matching the wavenumber spacing as used in the measured data through interpolation. Furthermore the measured data and molecules need to be in the same physical quantity, for which in this case absorbance is the most useful. The aggregated absorbance of the gas is determined by the terms of Equation 2.2 and for a single molecule by Equation 4.1

$$A(\nu, c_{mol}) = \epsilon_{mol}(\nu) c_{mol} l_{path}, \tag{4.1}$$

where the absorbance $A(\nu, c_{mol})$ is determined by the concentration c_{mol}, the optical path length of the laser in the gas l_{path}, and the molar absorptivity ϵ_{mol}. Since the path length is measured from the setup and the molar absorptivity of the molecules is obtained from the databases, the absorbance is left as a function of the concentration. All calculations are done over the set wavelength region of 832 cm^{-1} to 1263 cm^{-1} as determined by the range of QCL. The concentrations are determined through what is essentially a curve fitting problem as the absorbances of the molecules can be laid over the measured absorbance and then best fitted to match by tweaking the concentrations of the molecules. This curve-fitting problem can be expressed as a non-linear least squares problem as in Equation 4.2:

$$r(\nu, C) = A_{spectrum}(\nu) - \sum_{\substack{c_{mol} \in C \\ \epsilon_{mol}(\nu) \in E(\nu)}} \epsilon_{mol}(\nu) c_{mol} l_{path} \tag{4.2}$$

where C and $E(\nu)$ denote the set of all concentrations and the set of all molar absorptivities respectively of the molecules in the gas. Now take

$$S(C) = \sum_{\nu=832 cm^{-1}}^{1263 cm^{-1}} r(\nu, C)^2, \tag{4.3}$$

where r is a residue quantity, and S is the quantity to be minimized. This non-linear least squares problem is solved using the Levenberg–Marquardt algorithm which tweaks the concentrations from a given set of initial concentrations.

4.2.2 Input and results

The method as described in subsection 4.2.1 is applied on the molecules found by way of the p-region analysis as in section A.2, where all molecules with presence in one or more regions are checked for their concentration with the measured absorbance. The initial concentration as in section A.3 are used. The Levenberg-Marquardt algorithm is implemented through the *lsqnonlin* function in MATLAB, with its *option* argument as given in section A.3. The absorbances of six molecules with some of the highest concentrations are plotted over the measured absorbance in Figure 4.1:

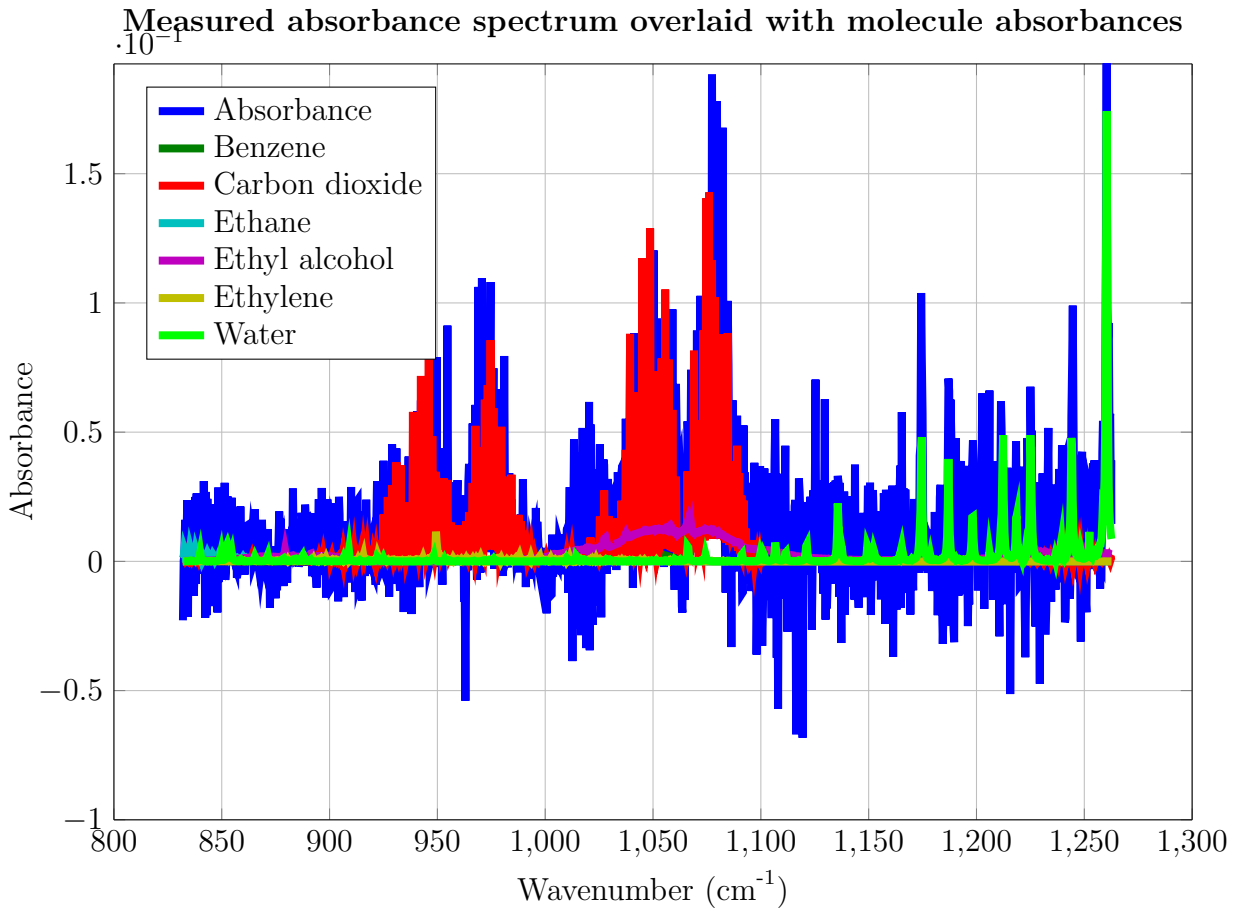


Figure 4.1: Measured absorbance of breath of a healthy patient. The measured absorbance is overlaid with absorbances of molecules with some of the highest concentrations. Concentrations found for all molecules for this sample are given at section A.3. The absorbances plotted are scaled down in resolution to 0.35 cm^{-1} for amenable file size.

4.3 Ethanol estimation by CO_2 removal

4.3.1 Motivation for CO_2 removal

When looking closer at the measured absorbance and the absorbance of CO_2 from the calculated concentrations as in Figure 4.2, it can be seen that they do not completely match.

The CO_2 peaks specifically are quite a bit off in some places, more so than just the noise level as judged from other regions. This also affects the estimation of concentration for the other major molecule present in this region, namely ethanol. Hence it could prove useful to remove CO_2 and its uncertainties in order to get a better estimate of the ethanol concentration, much in the same way as in [12].

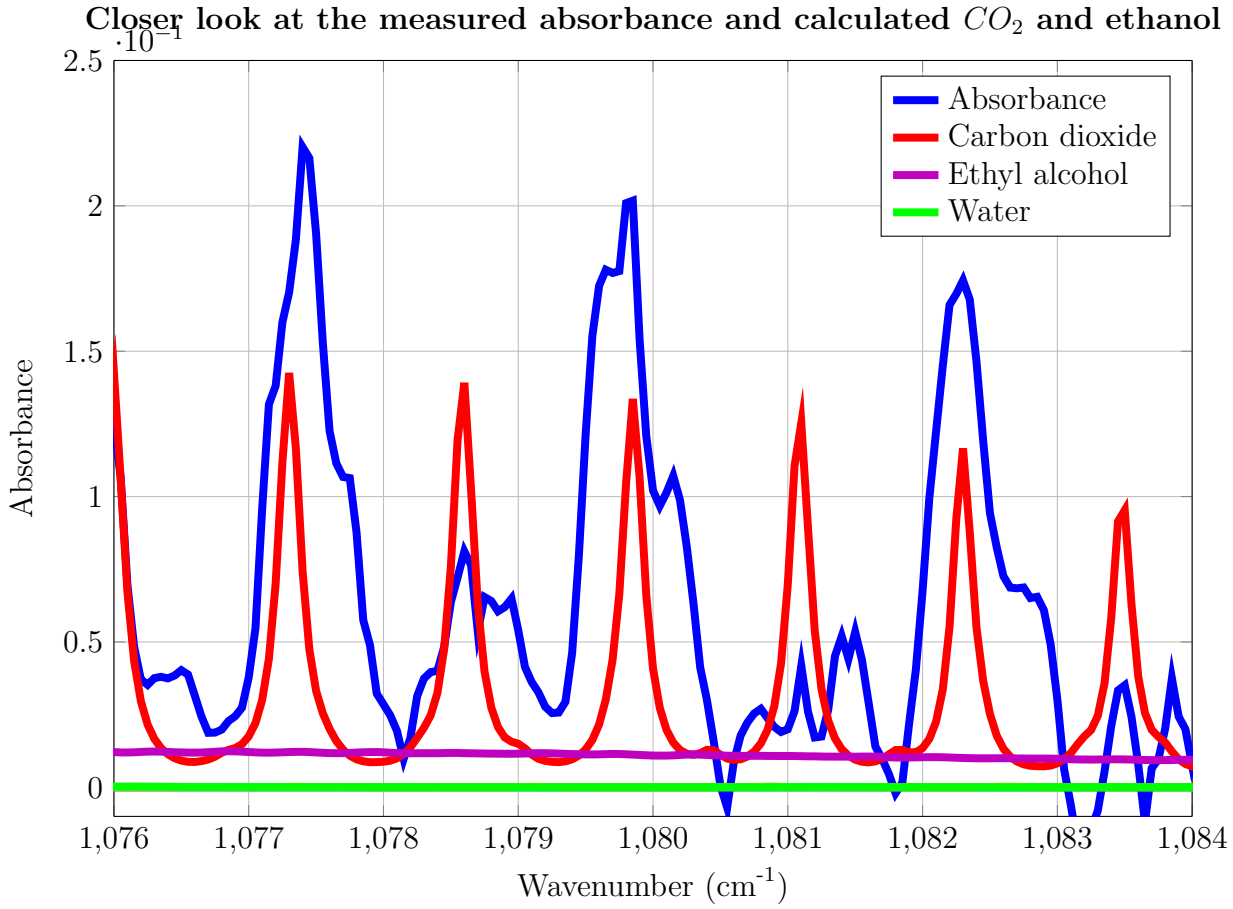


Figure 4.2: Measured absorbance of the breath of a healthy patient. The absorbance is overlaid with the spectra of water, carbon dioxide, and ethyl alcohol. This figure shows how the measured carbon dioxide peaks are far from the calculated ones.

4.3.2 Implementation of CO_2 removal for ethanol estimation

The concentrations as found using section 4.2 are used, and the molecules list as according to subsection 4.1.2 and subsection 4.1.4. The absorbance contributions of all molecules except for CO_2 are subtracted from the measured absorbance. All that should be left is noise around zero absorbance and the contribution of CO_2. The peaks from CO_2 are now identified using a peak detection algorithm, and then fitted by normal distributions. This way the actual measured CO_2 peaks are fitted including any deviations they have from the CO_2 spectrum according to the database. The fits are then subtracted from the signal, and all that should be left is the difference between the measurement and the absorbance contributions of all molecules as calculated using section 4.2. The contributions of all other molecules are now summed back on this in order to get the full measurement back minus any absorbance by CO_2. From this the concentrations are once again determined. The result is a drop of 2% of the calculated ethanol concentration.

5. Disease prediction by machine learning

5.1 Motivation for machine learning

For the breath samples measured, an analytical study based on p-values was performed to predict and to categorize the asthmatic and healthy samples[10]. This study did not get any definitive results for prediction. Multiple articles state their believe that methods more complicated than just using the p-values could give a better results[13][9]. Finding the molecules which differentiate between a healthy and non-healthy subject, and predicting diseases of a patient are both mentioned as possibilities. Since neither of these sources applies machine learning to actually predict disease, and [6] mentions that the application of machine learning performed on breath data is not widespread, an attempt to do just so is made. The analysis is performed on 70 absorbance spectra from healthy patients, and 70 absorbance spectra from asthmatic patients.

5.2 Machine learning for classification

Algorithms performing classification by machine learning do so by looking for pattern in the data of the different samples. Since for the current set of samples it is known whether they originate from healthy or asthmatic patients, supervised learning methods are applied. This means the machine learning algorithm is trained to recognize patterns of healthy and asthmatic samples by feeding it a part of all samples. In this case 98 of 140 samples are used to train the algorithms and the other 42 to test the accuracy of the algorithm.

5.3 Classification by trial-and-error

Since the scope of this project does not allow for learning the ins and outs of machine learning and the construction of an algorithm specifically made for these datasets, existing algorithms are used. Based on an extensive study of what kind of algorithms are most useful in the field of breath analysis, the focus initially lies on support vector machines (SVMs)[13]. A website specifically for uploading datasets, uploading machine learning algorithms, and running these algorithms on the datasets is used[4]. This website expresses the rate of success in a single number, the error, which is the amount of falsely

classified samples divided by the total amount of samples. Since most algorithm run in less than 5 minutes and they can run simultaneously, many different algorithms can be tested in a short amount of time. Aside from the complete set of absorbances of 140 samples over the wavelength range of 832 to 1263 cm^{-1}, various subsets of this are tested as well. Among these are the set of 36 wavenumbers for which the p-values are below 0.005, the set of concentrations of 167 compounds as derived from the 140 absorbances, and the set of 30 compounds with the highest concentrations as derived from the 140 absorbances.

5.4 Results

Results seem to generally be best for algorithms such as nearest neighbor, support vector machines, and rotation forests. A trend is also seen in that the larger data sets get lower misclassifications. The smallest errors are found for the unmodified data set of all absorbances over the entire wavelength range, as seen in Table 5.1.

Table 5.1: Best error rates for different algorithms run on the entire dataset of absorbances of 70 healthy and 70 asthmatic patients. Names of algorithms are as stated on the website[4].

Error rate	Name of algorithm	Description of algorithm
0.167	RBFNetwork_weka_nominal	Normalized Gaussian radial basis-basis function network
0.190	sgd-hinge-stepsize0.5-iter5 +Chi2	Stochastic gradient descent
0.190	svmlight_multiclass-linear	Linear support vector machine for multiclass classification
0.190	svmlight_multiclass	Support vector machine for multi-class classification
0.190	IBk_weka_nominal	K-nearest neighbours classifier
0.190	IB1_weka_nominal	Nearest neighbours classifier
0.214	RotationForest_weka_nominal +Chi2	Rotation forest
0.214	svmlight-linear	Support vector machine for binary classification

The similarity of the error rates from Table 5.1 can be explained by the small test set of 42 samples. The various nearest neighbour algorithms function quite similarly, and the same can be said for the support vector machines. As the results from Table 5.1 are the best from a total of 41 different algorithms, they corroborate the use of SVMs.

6. Discussion of results

6.1 Wavenumber calibration

Even after the wavenumber calibration the resulting absorbance spectrum doesn't match with CO_2 and H_2O peaks from literature. This is caused by the high variation of the twenty monitor measurements of which the mean is the basis of the wavenumber calibration. Regardless, wavenumber calibration is still useful since it is known that the different measurements do not match even though they should. The measurements should match since a cause of the variation between measurements is identified to be the piëzo element and its hysteresis, and this is expected to be the only major contributor.

6.2 Determination of molecules

A general disadvantage of choosing molecules based on their molar absorptivity is that some molecules might have a large impact on the eventual absorbance due to a high concentration, despite having a low absorptivity. They would therefore be falsely eliminated from the list of molecules to check for.

There is also the drawback of excluding molecules which might be an important part of the absorbance, but aren't selected since they do not have adequate molar absorptivity in the p-regions. This makes the subsequent determination of concentrations less accurate. It is possible to allow for more p-regions, either by decreasing the minimum amount of consecutive points needed or heightening the maximum p-values of which these regions consist. This way more molecules will have some absorptivity in one of the regions and therefore won't be scrapped.

Both problems to do with selecting molecules based on intensity can be resolved in advance by putting molecules that play an important role in the eventual absorbance in the standard list of molecules. This requires a priori knowledge of the important molecules which is obtained from literature[6] and simply by performing the least squares method to get the concentration on a large list of molecules. The concentrations won't be as accurate for a large list, but all molecules that contribute a lot to the spectrum can be picked out for later runs with strict molecule lists.

6.3 Determination of concentrations

From the shapes of CO_2 and H_2O in Figure 4.1 it can be seen that the concentrations calculated using the *lsqnonlin* function are definitely in the right order of magnitude. The signal-to-noise for the rest of the measured absorbance is however still quite large, and the calculations for the molecules with lower concentrations are therefore not very accurate.

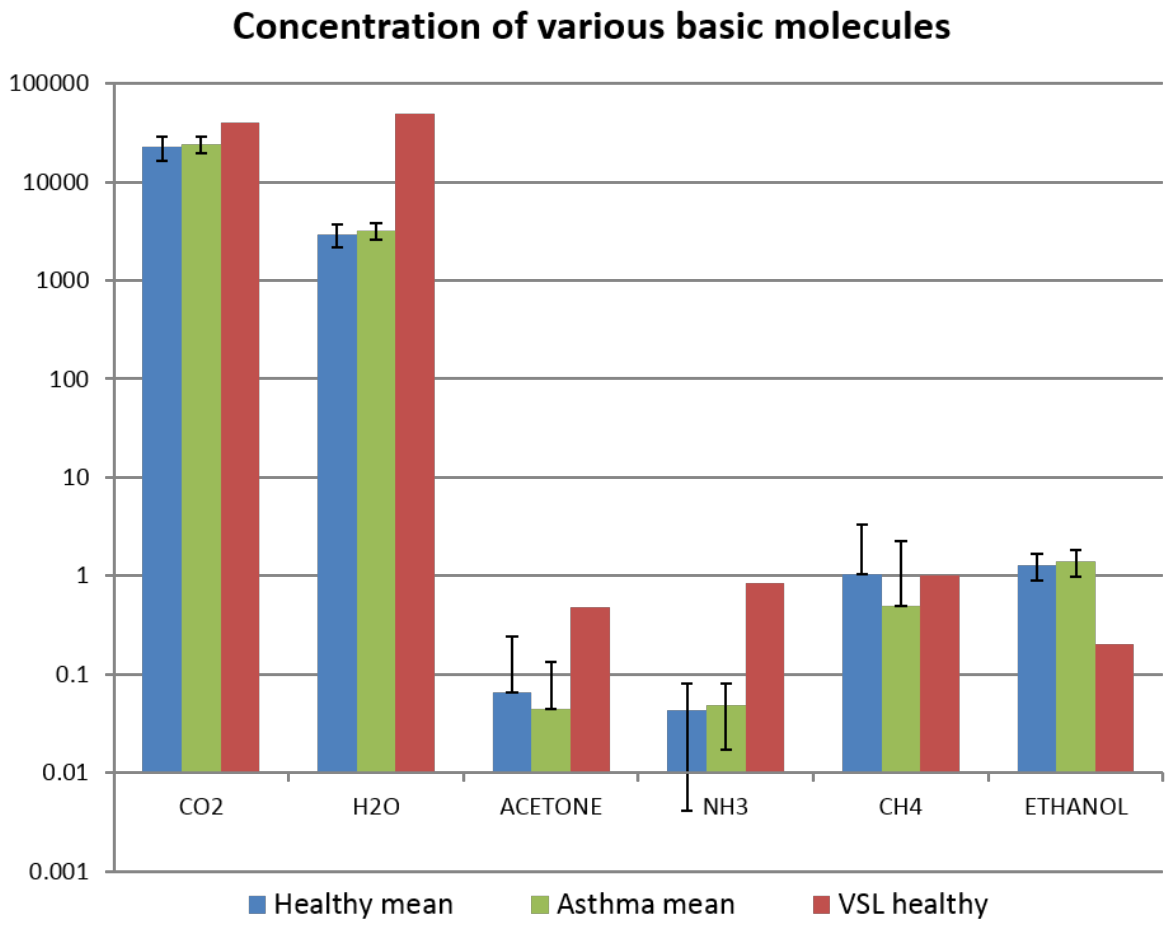


Figure 6.1: Average concentrations of molecules for healthy and asthmatic children compared to VSL data of healthy adults

Currently the concentration determination is not yet accurate enough to use for distinguishing between healthy and asthmatic children, as can be seen in Figure 6.1. The uncertainty of the measurements themselves can be improved by finding a moving element of which the displacement over voltage can be better characterized than the current piëzo. Alternatively an attempt can be made to calibrate the signals as they come straight from the measurement apparatus, before even the pre-calibration.

6.4 Categorizing health status by classification

The accuracy of the classification can possibly be improved by looking more closely into the used algorithms, and improving them using a-priori knowledge of the samples such as regions of relatively high signal-to-noise ratio. Wavenumber regions can be given weights based on the signal-to-noise ratios so the algorithm depends more heavily on these accurate regions. The error rate could also tell a lot more if the amount of false positives and false negatives is given. This would give separate error rates based on whether the algorithm judges the sample as healthy or asthmatic. It is possible that one of these error rates is much lower than the other, thereby increasing the certainty of the result above the total error rate.

Bibliography

[1] Atmosphere of earth - http://en.wikipedia.org/wiki/atmosphere_of_earth.

[2] Hitran database http://www.cfa.harvard.edu/hitran/.

[3] Matlab peakfinder code http://www.mathworks.com/matlabcentral/fileexchange/25500-peakfinder.

[4] Mlcomp - http://mlcomp.org/.

[5] Northwest infrared https://secure2.pnl.gov/nsd/nsd.nsf/welcome.

[6] B de Lacy Costello, A Amann, H Al-Kateb, C Flynn, W Filipiak, T Khalid, D Osborne, and N M Ratcliffe. A review of the volatiles from the healthy human body. *Journal of Breath Research*, 8(1):014001, 2014.

[7] I. Horvath and J.C. de Jongste. *Exhaled Biomarkers*. European Respiratory Society, 2010.

[8] Zhe Hou. Breath analysis using infrared spectroscopy. 2013.

[9] J. D. Malley, A. Dasgupta, and J. H. Moore. The limits of p-values for biological data mining. *BioData Min*, 6(1):10, 2013.

[10] Esther Van Mastrigt, Adonis Reyes-Reyes, Karl Brand, Nandini Bhattacharya, H.P. Urbach, Andrew Stubbs, Johan C. De Jongste, and Marielle W. Pijnenburg. *Exhaled Breath Profiling In Children With Asthma And Cystic Fibrosis: A Laser Spectroscopy-Based Method Focused On Hydrocarbons*, pages A5639–A5639–. American Thoracic Society, May 2014.

[11] A Reyes-Reyes, Z Hou, E Van Mastrigt, R C Horsten, J C De Jongste, M W Pijnenburg, H P Urbach, and N Bhattacharya. Multicomponent gas analysis using broadband quantum cascade laser spectroscopy. *Optics express*, 2014.

[12] Adonis Reyes-Reyes, Roland C. Horsten, H. Paul Urbach, and Nandini Bhattacharya. Study of the exhaled acetone in type 1 diabetes using quantum cascade laser spectroscopy. *Analytical Chemistry*, 87(1):507–512, 2015. PMID: 25506743.

[13] A Smolinska, A-Ch Hauschild, R R R Fijten, J W Dallinga, J Baumbach, and F J van Schooten. Current breathomics—a review on data pre-processing techniques

and machine learning in metabolomics breath analysis. *Journal of Breath Research*, April 2014.

Appendices

A. Results for a sample of a healthy patient

A.1 Standard molecules

Table A.1: List of standard molecules used.

Molecule
Acetic_acid
Acetone
Ammonia_anhydrous
Benzene
Carbon_dioxide
Carbon_monoxide
Ethane
Ethyl_alcohol
Ethylene
Formaldehyde__monomer
Methane
Pentane
Propane
Water

A.2 Molecules found by p-region analysis

Table A.2: Table of molecules that are present in either or both of the p-regions. The p-regions were found using a maximum p-value of 05 for at least 10 consecutive data points or 0.5 cm^{-1}. These molecules have a mean absorptivity in this region of at least 10^{-4}.

	1181.8-1182.55	**1261.4-1262.05**
1-Chloro-1_1_2_2-tetrafluoroethane_(R124A)	1	1
1-Hexanoic_acid	1	0
1_1_1-Chlorodifluoroethane_(Freon-142B)	1	0

Table A.2: Table of molecules that are present in either or both of the p-regions. The p-regions were found using a maximum p-value of 05 for at least 10 consecutive data points or 0.5 cm^{-1}. These molecules have a mean absorptivity in this region of at least 10^{-4}.

	1181.8-1182.55	**1261.4-1262.05**
1_1_1-Trifluoroacetone	1	0
1_1_1-Trifluoroethane_(R143A)	1	1
1_1_1_2_3_3_3-Heptafluoropropane_(HFC227EA)	1	1
1_2-Dibomotetrafluoroethane_(Freon-114B2)	1	0
1_2-Dibromoethane_(EDB)	1	0
1_2-Dichloro-1-fluoroethane_(R141)	1	0
1_2-Dichloro-1_1_2-trifluoroethane_(F132A)	1	0
1_2-Dimethoxyethane	1	0
1_4-Dioxane	0	1
2-Chloro-1_1_1-trifluoroethane_(R133A)	1	1
2-Chloropropane	0	1
2-Chlroro-1_1_1_2-tetrafluoroethane_(HCFC-124)	1	0
2-Ethoxyethyl_acetate	1	1
2-Methoxyethanol	1	0
2H__3H-Perfluoropentane	1	1
2_2-Dichloro-1_1_1-trifluoroethane_(Freon-123)	1	1
2_2-Difluorotetrachloroethane_(F112A)	1	0
2_2_2-Trifluoroethanol	1	1
2_2_4-Trimethyl-1_3-pentanediol_isobutyrate_(Texanol)	1	1
3-Methyl-2-pentanone	1	0
Acetic_acid	1	1
Acetic_anhydride	1	0
Acetid_acid_dimer	0	1
Acetone_cyanohydrin	1	0
Aniline	0	1
Benzyl_chloride	0	1
Bromochlorodifluoromethane_(Freon-12B1)	1	0
Butyl_acetate	0	1
Butyric_acid	1	0
Carbonyl_fluoride	0	1
Chloromethyl_ethyl_ether	1	1
Chloromethyl_methyl_ether	1	0
Chloropentafluoroethane_(R115)	1	1
Chlorosulfonyl_isocyanate_(CSI)	1	0
Diborane	1	0
Dichloromethane	0	1
Diethyl_ether	1	0

Table A.2: Table of molecules that are present in either or both of the p-regions. The p-regions were found using a maximum p-value of 05 for at least 10 consecutive data points or 0.5 cm^{-1}. These molecules have a mean absorptivity in this region of at least 10^{-4}.

	1181.8-1182.55	**1261.4-1262.05**
Diethyl_sulfide	0	1
Diisopropyl_ether	1	0
Diisopropylamine	1	0
Diketene	0	1
Dimethoxymethane	1	0
Dimethyl_carbonate	0	1
Dimethyl_ether	1	0
Dimethylcarbamoyl_chloride	0	1
Dipropylene_glycol_methyl_ether	1	1
Ethyl_acetate	0	1
Ethyl_acrylate	1	1
Ethyl_bromide_(Halon-2001)	0	1
Ethyl_butyrate	1	1
Ethyl_chloroformate	1	0
Ethyl_formate	1	0
Ethyl_methyl_ether	1	0
Ethyl_tert-butyl_ether	0	1
Ethyl_trifluoroacetate	1	1
Ethylvinyl_ether_(EVE)	1	0
Formic_acid__dimer	1	0
Freon-113_(1_1_2-Trichlorortrifluoroethane)	1	0
Freon-114_(1_2-dichlorortetrafluoroethane)	1	1
Freon-134a_(1_1_1_2-tetrafluoroethane)	1	0
Freon-218_(octafluoropropane)	0	1
Freon-C318_(octafluorocyclobutane)	0	1
Furan	1	0
Glycoaldehyde	0	1
Guaiacol	1	1
Hexachloro-1_3-butadiene	1	0
Hexafluoroacetone	1	1
Hexafluoroethane_(Freon-116)	0	1
Hexafluoroisobutylene	1	1
Hexafluoropropene	1	0
Hydrogen_peroxide	0	1
Isobutyl_acetate	0	1
Isopentyl_acetate	0	1
Isopropyl_acetate	0	1

Table A.2: Table of molecules that are present in either or both of the p-regions. The p-regions were found using a maximum p-value of 05 for at least 10 consecutive data points or 0.5 cm^{-1}. These molecules have a mean absorptivity in this region of at least 10^{-4}.

	1181.8-1182.55	**1261.4-1262.05**
Methacryloyl_chloride	0	1
Methanesulfonyl_chloride	1	0
Methyl_acetate	0	1
Methyl_acrylate	1	1
Methyl_benzoate	1	1
Methyl_formate	1	0
Methyl_iodide	0	1
Methyl_isoamyl_ketone	1	0
Methyl_isobutyl_ketone_(MIBK)	1	0
Methyl_isobutyrate	1	1
Methyl_methacrylate	1	0
Methyl_pivalate	1	0
Methyl_propionate	1	1
Methyl_propyl_ketone	1	0
Methyl_salicylate	1	1
Methylchloroformate_(MCF)	1	0
Methyldichlorodisilanes_(mixed_isomers)	0	1
Methylethyl_ketone	1	0
Methyltrichlorosilane	0	1
Methylvinyl_ketone	1	1
N_N'-Dimethyl_formamide_(DMF)	0	1
N_N-Diethylaniline	0	1
N_N-diethyl_formamide	0	1
Nitrogen_dioxide_and_dinitrogen_tetroxide	0	1
Octanoic_acid	1	0
Paraldehyde	1	0
Pentafluoroethane_(f125)	1	0
Perfluorobutane	1	1
Perfluoroisobutylene_(PFIB)	1	0
Phenol	1	1
Propargyl_chloride	0	1
Propionic_acid_(and_some_dimer)	1	0
Propyl_acetate	0	1
Propylene_glycol	0	1
Sulfuryl_chloride	1	0
Sulfuryl_fluoride	0	1
Tetrafluoromethane	0	1

Table A.2: Table of molecules that are present in either or both of the p-regions. The p-regions were found using a maximum p-value of 05 for at least 10 consecutive data points or 0.5 cm^{-1}. These molecules have a mean absorptivity in this region of at least 10^{-4}.

	1181.8-1182.55	**1261.4-1262.05**
Trichlorofluroethylene	1	0
Trifluoroacetic_acid	1	1
Trifluoroacetic_anhydride	1	1
Trifluoroacetyl_chloride	1	1
Trifluoromethane_(Freon-23)	1	0
Trifluoromethylsulfur_pentafluoride	0	1
Trifluoronitrosomethane	1	1
Trimethylamine	1	0
Vinayl_acetate	0	1
bis(2-Chloroethyl)_ether	0	1
cis-1_3-Dichloropropene	0	1
m-Cresol	1	0
n-Amyl_acetate	0	1
o-Toluidine	0	1
tert-Amyl_methyl_ether	1	0

A.3 Concentrations found for a single healthy patient

In order to find the concentrations, the following *options* argument is used for the *lsqnonlin* function in MATLAB:
options = optimset('Display','off','TolFun',1e-15)

Table A.3: Table of initial concentrations used to calculate the concentrations as given in section A.3.

Molecule	**Concentration (ppmv)**
Carbon dioxide	1000
Water	1000
Standard molecules	1
Other molecules	0.01

Table A.4: Concentrations found using the method described in section 4.2. These are of a single sample from a healthy patient. The standard list of molecules as given in section A.1 and the list of molecules as given by section A.2 are used.

Molecule	**concentration (ppmv)**
Acetic_acid	2.22045e-14
Acetone	2.22045e-14
Ammonia_anhydrous	2.22045e-14
Benzene	1.79671
Carbon_dioxide	0.912382
Carbon_monoxide	1538.28
Ethane	2.66622
Ethyl_alcohol	0.946815
Ethylene	0.109119
Formaldehyde__monomer	2.22045e-14
Methane	2.22045e-14
Pentane	2.22045e-14
Propane	2.22045e-14
Water	0.0988024
1-Chloro-1_1_2_2-tetrafluoroethane_(R124A)	2.22045e-14
1-Hexanoic_acid	2.22045e-14
1_1_1-Chlorodifluoroethane_(Freon-142B)	0.00478773
1_1_1-Trifluoroacetone	2.22045e-14
1_1_1-Trifluoroethane_(R143A)	2.22045e-14
1_1_1_2_3_3_3-Heptafluoropropane_(HFC227EA)	2.22045e-14
1_2-Dibomotetrafluoroethane_(Freon-114B2)	2.22045e-14
1_2-Dibromoethane_(EDB)	2.22045e-14
1_2-Dichloro-1-fluoroethane_(R141)	2.22045e-14
1_2-Dichloro-1_1_2-trifluoroethane_(F132A)	2.22045e-14
1_2-Dimethoxyethane	2.22045e-14
1_4-Dioxane	2.22045e-14
2-Chloro-1_1_1-trifluoroethane_(R133A)	2.22045e-14
2-Chloropropane	2.22045e-14
2-Chlroro-1_1_1_2-tetrafluoroethane_(HCFC-124)	2.22045e-14
2-Ethoxyethyl_acetate	2.22045e-14
2-Methoxyethanol	2.30625e-14
2H__3H-Perfluoropentane	2.22045e-14
2_2-Dichloro-1_1_1-trifluoroethane_(Freon-123)	2.22045e-14
2_2-Difluorotetrachloroethane_(F112A)	0.0104979
2_2_2-Trifluoroethanol	2.22045e-14
2_2_4-Trimethyl-1_3-pentanediol_isobutyrate_(Texanol)	2.22045e-14
3-Methyl-2-pentanone	2.22045e-14
Acetic_acid	2.22045e-14

Table A.4: Concentrations found using the method described in section 4.2. These are of a single sample from a healthy patient. The standard list of molecules as given in section A.1 and the list of molecules as given by section A.2 are used.

Molecule	**concentration (ppmv)**
Acetic_anhydride	2.22045e-14
Acetid_acid_dimer	2.22045e-14
Acetone_cyanohydrin	2.22045e-14
Aniline	0.692194
Benzyl_chloride	2.22045e-14
Bromochlorodifluoromethane_(Freon-12B1)	2.22045e-14
Butyl_acetate	2.22045e-14
Butyric_acid	2.22045e-14
Carbonyl_fluoride	2.22045e-14
Chloromethyl_ethyl_ether	2.22045e-14
Chloromethyl_methyl_ether	2.22045e-14
Chloropentafluoroethane_(R115)	0.0115797
Chlorosulfonyl_isocyanate_(CSI)	2.22045e-14
Diborane	0.114448
Dichloromethane	2.22045e-14
Diethyl_ether	0.0152475
Diethyl_sulfide	2.22045e-14
Diisopropyl_ether	2.22045e-14
Diisopropylamine	2.22045e-14
Diketene	2.22045e-14
Dimethoxymethane	2.22045e-14
Dimethyl_carbonate	2.22045e-14
Dimethyl_ether	2.22045e-14
Dimethylcarbamoyl_chloride	2.22045e-14
Dipropylene_glycol_methyl_ether	2.22045e-14
Ethyl_acetate	2.22045e-14
Ethyl_acrylate	0.039846
Ethyl_bromide_(Halon-2001)	2.22046e-14
Ethyl_butyrate	2.22045e-14
Ethyl_chloroformate	2.22045e-14
Ethyl_formate	2.22045e-14
Ethyl_methyl_ether	2.22045e-14
Ethyl_tert-butyl_ether	0.195303
Ethyl_trifluoroacetate	2.22045e-14
Ethylvinyl_ether_(EVE)	2.22045e-14
Formic_acid__dimer	2.22045e-14
Freon-113_(1_1_2-Trichlorortrifluoroethane)	2.22045e-14
Freon-114_(1_2-dichlorortetrafluoroethane)	2.22045e-14

Table A.4: Concentrations found using the method described in section 4.2. These are of a single sample from a healthy patient. The standard list of molecules as given in section A.1 and the list of molecules as given by section A.2 are used.

Molecule	**concentration (ppmv)**
Freon-134a_(1_1_1_2-tetrafluoroethane)	2.22045e-14
Freon-218_(octafluoropropane)	2.22045e-14
Freon-C318_(octafluorocyclobutane)	2.22045e-14
Furan	2.22045e-14
Glycoaldehyde	2.22045e-14
Guaiacol	2.22045e-14
Hexachloro-1_3-butadiene	2.22045e-14
Hexafluoroacetone	2.22045e-14
Hexafluoroethane_(Freon-116)	2.22045e-14
Hexafluoroisobutylene	2.22045e-14
Hexafluoropropene	2.22045e-14
Hydrogen_peroxide	2.22045e-14
Isobutyl_acetate	2.22045e-14
Isopentyl_acetate	2.22045e-14
Isopropyl_acetate	2.22045e-14
Methacryloyl_chloride	0.101246
Methanesulfonyl_chloride	0.0190093
Methyl_acetate	2.22045e-14
Methyl_acrylate	2.22045e-14
Methyl_benzoate	2.22045e-14
Methyl_formate	2.22045e-14
Methyl_iodide	2.22045e-14
Methyl_isoamyl_ketone	2.22045e-14
Methyl_isobutyl_ketone_(MIBK)	2.22045e-14
Methyl_isobutyrate	2.22045e-14
Methyl_methacrylate	2.22045e-14
Methyl_pivalate	0.0277697
Methyl_propionate	2.22045e-14
Methyl_propyl_ketone	2.22045e-14
Methyl_salicylate	2.22045e-14
Methylchloroformate_(MCF)	2.22046e-14
Methyldichlorodisilanes_(mixed_isomers)	2.22045e-14
Methylethyl_ketone	2.22045e-14
Methyltrichlorosilane	2.22046e-14
Methylvinyl_ketone	2.22045e-14
N_N'-Dimethyl_formamide_(DMF)	2.22045e-14
N_N-Diethylaniline	2.22045e-14
N_N-diethyl_formamide	2.22045e-14

Table A.4: Concentrations found using the method described in section 4.2. These are of a single sample from a healthy patient. The standard list of molecules as given in section A.1 and the list of molecules as given by section A.2 are used.

Molecule	**concentration (ppmv)**
Nitrogen_dioxide_and_dinitrogen_tetroxide	2.22045e-14
Octanoic_acid	2.22045e-14
Paraldehyde	2.22045e-14
Pentafluoroethane_(f125)	2.22045e-14
Perfluorobutane	2.22045e-14
Perfluoroisobutylene_(PFIB)	0.000842043
Phenol	2.22045e-14
Propargyl_chloride	0.250046
Propionic_acid_(and_some_dimer)	2.22045e-14
Propyl_acetate	2.22045e-14
Propylene_glycol	2.22045e-14
Sulfuryl_chloride	2.22045e-14
Sulfuryl_fluoride	2.22046e-14
Tetrafluoromethane	2.22045e-14
Trichlorofluroethylene	2.22045e-14
Trifluoroacetic_acid	2.22045e-14
Trifluoroacetic_anhydride	0.00520583
Trifluoroacetyl_chloride	2.22045e-14
Trifluoromethane_(Freon-23)	2.22045e-14
Trifluoromethylsulfur_pentafluoride	2.22045e-14
Trifluoronitrosomethane	2.22045e-14
Trimethylamine	2.22045e-14
Vinayl_acetate	2.22045e-14
bis(2-Chloroethyl)_ether	2.22045e-14
cis-1_3-Dichloropropene	2.22045e-14
m-Cresol	2.22045e-14
n-Amyl_acetate	2.22045e-14
o-Toluidine	2.22045e-14
tert-Amyl_methyl_ether	2.22045e-14

B. List of molecules

B.1 Breath molecules with absorbance in the 1020 to 1100 cm^{-1} region.

Table B.1: Molecules as filtered using subsection 4.1.4 in the 1020 to 1100 cm^{-1} region on 199 molecules that should be present in breath[6] .

Molecules
1-Hexanoic_acid
1-Pentanol_(n-amyl_alcohol)
1-Propanol
1_2-Diclorobenzene
1_3-Dichlorobenzene
1_4-Dichlorobenzene
1_4-Dioxane
2-Butoxyethanol
2-Ethoxyethyl_acetate
2-Ethyl-1-hexanol
2-Methoxyethanol
2-methyl-1-propanol_(IBA__isobutanol)
2-Methyl-2-propanal_(Isobutenal)
2-Methylfuran
2_3-Butanedione
3-methylfuran
Acetic_acid
Acetic_anhydride
Acetol
Allene_(1_2-propadiene)
Allyl_alcohol
Ammonia_anhydrous
Benzene
Benzyl_alcohol
Butyl_acetate
Butyric_acid

Table B.1: Molecules as filtered using subsection 4.1.4 in the 1020 to 1100 cm^{-1} region on 199 molecules that should be present in breath[6] .

Molecules
Cadaverine
Chlorobenzene
Cyclopentene
Cyclopropane
Diethyl_ether
Dimethoxymethane
Dimethyl_ether
Dimethyl_sulfoxide
Dipropylene_glycol_methyl_ether
Ethyl_acetate
Ethyl_acrylate
Ethyl_alcohol
Ethyl_butyrate
Ethyl_tert-butyl_ether
Ethylene
Ethylvinyl_ether_(EVE)
Furan
Furfural
Furfuryl_alcohol
Glycoaldehyde
Guaiacol
Isobutyl_acetate
Isopentyl_acetate
Isopropyl_acetate
Isopropyl_alcohol
Methyl_acetate
Methyl_alcohol
Methyl_methacrylate
Methyl_propionate
Methylamine
n-Butyl_alcohol
N_N'-Dimethyl_formamide_(DMF)
N_N-diethyl_formamide
Octanoic_acid
Propionic_acid_(and_some_dimer)
Propylene_glycol
Propylene_sulfide
Pyridine
sec-Butyl_alcohol

Table B.1: Molecules as filtered using subsection 4.1.4 in the 1020 to 1100 cm^{-1} region on 199 molecules that should be present in breath[6] .

Molecules
tert-Butyl_methyl_ether
Thiophene
trans-Crotonaldehyde
Trimethylamine

C. List of molecules

C.1 All molecules in HITRAN and PNNL database

Table C.1: All molecules present in the HITRAN and PNNL database.

Molecules
1-Bromopropane
1-Butene
1-Butyne
1-Chloro-1_1_2_2-tetrafluoroethane_(R124A)
1-Chloro-2-methylpropane
1-Chlorobutane
1-Chloropentane
1-Fluorohexane
1-Heptanol
1-Hexanoic_acid
1-Hexanol
1-Hexene
1-Nitropropane
1-Nonene
1-Octene
1-Pentanol_(n-amyl_alcohol)
1-Pentene
1-Propanethiol
1-Propanol
1_1-Dichloro-1-fluoroethane_(R141B)
1_1-Dichloroethane
1_1-Dichloroethene
1_1-Difluoro-2_2-dichloroethane_(R132A)
1_1-Difluoroethane_(Freon-152A)
1_1-Dimethylhydrazine
1_1_1-Chlorodifluoroethane_(Freon-142B)
1_1_1-Trichloroethane
1_1_1-Trifluoroacetone
1_1_1-Trifluoroethane_(R143A)

Table C.1: All molecules present in the HITRAN and PNNL database.

Molecules
1_1_1_2-Tetrachloroethane
1_1_1_2_3_3_3-Heptafluoropropane_(HFC227EA)
1_1_2-Trichloroethane
1_1_2_2-Tetrachloro-1-fluoroethane_(F121)
1_1_2_2-Tetrachloroethane
1_2-Dibomotetrafluoroethane_(Freon-114B2)
1_2-Dibromoethane_(EDB)
1_2-Dichloro-1-fluoroethane_(R141)
1_2-Dichloro-1_1_2-trifluoroethane_(F132A)
1_2-Dichloroethane
1_2-Dichloroethene_(Dioform)
1_2-Dichloropropane
1_2-Diclorobenzene
1_2-Difluoro-1_2-dichloroethane_(R132)
1_2-Difluorotetrachloroethane_(R112)
1_2-Dimethoxyethane
1_2-Epoxy-3-chloropropane
1_2-Epoxybutane
1_2-Epoxypropane
1_2_3_4-Tetrahydronaphthaline
1_2_3_4-Tetramethylbenzene
1_2_3_5-Tetramethylbenzene
1_3-Butadiene
1_3-Dichlorobenzene
1_3-Dichloropropane
1_3_5-Trimethylbenzene_(Mesitylene)
1_4-Dichlorobenzene
1_4-Dioxane
2-Bromopropane
2-Butene
2-Butoxyethanol
2-Chloro-1_1_1-trifluoroethane_(R133A)
2-Chloroethanol
2-Chloroethyl_ether
2-Chloropropane
2-Chlorotoluene
2-Chlroro-1_1_1_2-tetrafluoroethane_(HCFC-124)
2-Ethoxyethyl_acetate
2-Ethyl-1-hexanol
2-Ethyltoluene

Table C.1: All molecules present in the HITRAN and PNNL database.

Molecules
2-Fluoroethanol
2-Hexanone
2-Iodo-2-methylpropane
2-Iodopropane
2-Mercaptoethanol
2-Methoxyethanol
2-Methyl-1-butene
2-Methyl-1-pentene
2-Methyl-1-propanal_(Isobutyraldehyde)
2-Methyl-1-propanethiol
2-methyl-1-propanol_(IBA__isobutanol)
2-Methyl-1_3-butadiene_(isoprene)
2-Methyl-2-butene
2-Methyl-2-pentene
2-Methyl-2-propanal_(Isobutenal)
2-Methylfuran
2-Methylpentane
2-Nitropropane
2-Nonanone
2-Pentanol_(sec-amyl_alcohol)
2-Pentylfuran
2-Picoline
2-Propanethiol
2H__3H-Perfluoropentane
2_2-Dichloro-1_1_1-trifluoroethane_(Freon-123)
2_2-Difluorotetrachloroethane_(F112A)
2_2-Dimethyl_butane
2_2_2-Trifluoroethanol
2_2_4-Trimethyl-1_3-pentanediol_isobutyrate_(Texanol)
2_2_4-Trimethyl-2-pentene
2_3-Butanedione
2_3-Dichloro-1-propene
2_3-Dimethylbutane
2_4-Diisocyanatetoluene
2_4_4-Trimethyl-1-pentene
2_6-Dimethylaniline
3-Carene
3-Chlorotoluene
3-Ethyltoluene
3-Methyl-1-butene

Table C.1: All molecules present in the HITRAN and PNNL database.

Molecules
3-Methyl-2-pentanone
3-methylfuran
3-Methylhexane
3-Methylpentane
3-Picoline
3_4-Dichloro-1-butene
4-Chlorotoluene
4-Ethyltoluene
4-Methyl-1-pentene
4-Picoline
4-Vinyl-1-cyclohexene
Acetic_acid
Acetic_anhydride
Acetid_acid_dimer
Acetol
Acetone
Acetone_cyanohydrin
Acetonitrile
Acetylaldehyde
Acetylene
Acetyl_chloride
Acrolein
Acrylol_chloride
Acrylonitrile
Allene_(1_2-propadiene)
Allylamine
Allyl_alcohol
Allyl_bromide
Allyl_chloride
Allyl_fluoride
Allyl_iodide
Allyl_isothiocyanate
alpha-Pinene_(1S)(-)
Ammonia_anhydrous
Amyl_nitrate
Aniline
Arsine
Benzaldehyde
Benzene
Benzenethiol

Table C.1: All molecules present in the HITRAN and PNNL database.

Molecules
Benzonitrile
Benzyl_alcohol
Benzyl_bromide
Benzyl_chloride
beta-Pinene_(1S)_(-)
bis(2-Chloroethyl)_ether
Boron_tribromide
Boron_trichloride
Boron_trifluoride
Bromobenzene
Bromochlorodifluoromethane_(Freon-12B1)
Bromochloromethane
Bromotrifluoromethane_(Freon-13B1)
Butyl_acetate
Butyraldehyde
Butyric_acid
Cadaverine
Carbonyl_fluoride
Carbonyl_sulfide
Carbon_dioxide
Carbon_disulfide
Carbon_monoxide
Carbon_tetrachloride
Chloroacetone
Chloroacetonitrile
Chlorobenzene
Chlorodifluoromethane_(F22)
Chloroethane
Chloroform_(trichloromethane)
Chloromethyl_ethyl_ether
Chloromethyl_methyl_ether
Chloropentafluoroethane_(R115)
Chloropicrin
Chlorosulfonyl_isocyanate_(CSI)
Chlorotrifluoroethylene
Chlorotrifluoromethane_(CFC-13)
Cineole
cis-1_2-Dichloroethylene
cis-1_3-Dichloropropene
cis-2-Pentene

Table C.1: All molecules present in the HITRAN and PNNL database.

Molecules
cis-4-Methyl-2-pentene
Cumene
Cyanogen
Cyanogen_chloride
Cyclodecane
Cycloheptane
Cycloheptene
Cyclohexane
Cyclohexanol
Cyclohexanone
Cyclohexene
Cyclooctane
Cyclopentane
Cyclopentene
Cyclopropane
d-Limonene
Diborane
Dibromomethane
Dichlorofluoromethane_(Freon-21)
Dichloromethane
Dichloromethylphosphine
Dichlorosilane
Diethylamine
Diethylketone
Diethyl_ether
Diethyl_sulfate
Diethyl_sulfide
Difluorodibromomethane_(Freon-12B2)
Difluoromethane_(R-32)
Diiodomethane
Diisopropylamine
Diisopropyl_ether
Diketene
Dimethoxymethane
Dimethylamine
Dimethylcarbamoyl_chloride
Dimethyl_carbonate
Dimethyl_disulfide
Dimethyl_ether
Dimethyl_sulfate

Table C.1: All molecules present in the HITRAN and PNNL database.

Molecules
Dimethyl_sulfide
Dimethyl_sulfoxide
Dipropylene_glycol_methyl_ether
Dipropyl_ether
Ethane
Ethylamine
Ethylene
Ethylenediamine
Ethylene_glycol
Ethylene_oxide
Ethylene_sulfide
Ethyliodide
Ethylvinyl_ether_(EVE)
Ethyl_acetate
Ethyl_acrylate
Ethyl_alcohol
Ethyl_benzene
Ethyl_bromide_(Halon-2001)
Ethyl_butyrate
Ethyl_chloroformate
Ethyl_cyanide_(2-methylacetonitrile)
Ethyl_formate
Ethyl_mercaptan
Ethyl_methyl_ether
Ethyl_nitrite
Ethyl_tert-butyl_ether
Ethyl_trifluoroacetate
Fluoroacetone
Fluorobenzene
Formaldehyde__monomer
Formic_acid_(and_some_dimer)
Formic_acid__dimer
Freon-113_(1_1_2-Trichlorortrifluoroethane)
Freon-114_(1_2-dichlorortetrafluoroethane)
Freon-12_(Dichlorodifluoromethane)
Freon-134a_(1_1_1_2-tetrafluoroethane)
Freon-218_(octafluoropropane)
Freon-C318_(octafluorocyclobutane)
Furan
Furfural

Table C.1: All molecules present in the HITRAN and PNNL database.

Molecules
Furfuryl_alcohol
Germane
Glycoaldehyde
Guaiacol
Hexachloro-1_3-butadiene
Hexachlorocyclopentadiene
Hexafluoroacetone
Hexafluorobenzene
Hexafluoroethane_(Freon-116)
Hexafluoroisobutylene
Hexafluoropropene
Hexamethylphosphoramide
HFC-134_(1_1_2_2-tetrafluoroethane)
Hydrazine
Hydrogen_bromide__anhydrous
Hydrogen_chloride__anhydrous
Hydrogen_cyanide_(prussic_acid)
Hydrogen_fluoride__anhydrous
Hydrogen_iodide__anhydrous
Hydrogen_peroxide
Hydrogen_sulfide
Iron_pentacarbonyl
Isoamyl_alcohol
Isobutane_(2-methylpropane)
Isobutene
Isobutyl_acetate
Isobutyric_acid
Isocumene
Isooctane_(2_2_4-trimethylpentane)
Isopentane
Isopentyl_acetate
Isophorone
Isopropylamine
Isopropyl_acetate
Isopropyl_alcohol
Isovaleraldehyde
m-Cresol
m-Xylene
Methacryloyl_chloride
Methane

Table C.1: All molecules present in the HITRAN and PNNL database.

Molecules
Methanesulfonyl_chloride
Methylamine
Methylchloroformate_(MCF)
Methyldichlorodisilanes_(mixed_isomers)
Methylethyl_ketone
Methylglyoxal
Methyltrichlorosilane
Methylvinyl_ketone
Methyl_acetate
Methyl_acrylate
Methyl_acrylonitrile
Methyl_alcohol
Methyl_benzoate
Methyl_bromide
Methyl_butyl_ether
Methyl_chloride
Methyl_fluoride
Methyl_formate
Methyl_iodide
Methyl_isoamyl_ketone
Methyl_isobutyl_ketone_(MIBK)
Methyl_isobutyrate
Methyl_isocyanate_(MIC)
Methyl_isopropyl_ketone
Methyl_isothiocyanate
Methyl_mercaptan
Methyl_methacrylate
Methyl_nitrite
Methyl_pivalate
Methyl_propionate
Methyl_propyl_ketone
Methyl_salicylate
Monomethyl_hydrazine
Morpholine
Myrcene
n-Amyl_acetate
n-Butane
n-Butylamine
n-Butyl_alcohol
n-Butyl_isocyanate

Table C.1: All molecules present in the HITRAN and PNNL database.

Molecules
n-Decane
n-Heptane
n-Heptene
n-hexadecane
n-Hexane
n-Nonane
n-Tridecane
n-Undecane
Naphthalene
Neopentyl_alcohol
Nickel_carbonyl
Nicotine
Nitric_acid__anhydrous
Nitric_oxide
Nitrobenzene
Nitroethane
Nitrogen_dioxide_(monomer)
Nitrogen_dioxide_and_dinitrogen_tetroxide
Nitrogen_trifluoride
Nitromethane
Nitrosyl_chloride
Nitrous_acid
Nitrous_oxide
N_N'-Dimethyl_formamide_(DMF)
N_N-Diethylaniline
N_N-diethyl_formamide
o-Toluidine
o-Xylene
Octane
Octanoic_acid
p-Xylene
Paraldehyde
Pentafluoroethane_(f125)
Pentane
Perchloromethyl_mercaptan
Perfluorobutane
Perfluoroisobutylene_(PFIB)
Phenol
Phosgene
Phosphine

Table C.1: All molecules present in the HITRAN and PNNL database.

Molecules
Phosphorous_oxychloride
Piperidine
Propane
Propargyl_alcohol
Propargyl_chloride
Propionaldehyde
Propionic_acid_(and_some_dimer)
Propylene
Propylene_glycol
Propylene_sulfide
Propylenimine
Propyl_acetate
Propyne_(methyl_acetylene)
Pyridine
Quinoline
sec-Amylamine
sec-Butylbenzene
sec-Butyl_alcohol
Silane
Silicon_tetrafluoride
Styrene_(monomer)
Styrene_oxide
Sulfuryl_chloride
Sulfuryl_fluoride
Sulfur_dioxide
Sulfur_hexafluoride
Sulfur_trioxide
t-Butyl_alcohol
tert-Amylamine
tert-Amyl_methyl_ether
tert-Butylbenzene
tert-Butyl_methyl_ether
Tetrachloroethylene
Tetrafluoromethane
Tetrahydrofuran_(THF)
Tetrahydrothiophene
Thiophene
Thiophosgene
Thiophosphoryl_chloride
Titanium_tetrachloride

Table C.1: All molecules present in the HITRAN and PNNL database.

Molecules
Toluene
trans-1_2-Dichloroethene
trans-1_3-Dichloropropene
trans-2-Pentene
trans-Crotonaldehyde
Tribromomethane_(bromoform)
Trichloroacetyl_chloride
Trichloroethylene
Trichlorofluoromethane_(Freon-11)
Trichlorofluroethylene
Triethylamine
Trifluoroacetic_acid
Trifluoroacetic_anhydride
Trifluoroacetyl_chloride
Trifluoromethane_(Freon-23)
Trifluoromethylsulfur_pentafluoride
Trifluoronitrosomethane
Trimethylamine
Tungsten_hexafluoride
Valeraldehyde
Valeric_acid
Vinayl_acetate
Vinyl_bromide
Vinyl_chloride
Vinyl_fluoride
Water
Water-d1