IR Spectroscopy Analysis using a Fuzzy Inference System

p15194468

Sam Fuller

2017

Table of Contents

[Introduction 2](#_Toc501740667)

[Literature review 2](#_Toc501740668)

[Fuzzy Inference Systems 2](#_Toc501740669)

[IR Spectroscopy 2](#_Toc501740670)

[The system 2](#_Toc501740671)

[General Approach 2](#_Toc501740672)

[Graphs Analysis 3](#_Toc501740673)

[Input 3](#_Toc501740674)

[Output 3](#_Toc501740675)

[Fuzzy Inference System 4](#_Toc501740676)

[Input 4](#_Toc501740677)

[Rules 4](#_Toc501740678)

[Output 4](#_Toc501740679)

[The system implementation 5](#_Toc501740680)

[Experimental Design and Evaluation 5](#_Toc501740681)

[data\cslow.jp 5](#_Toc501740682)

[data\cshigh.jpg 6](#_Toc501740683)

[data\cswhite.jp 7](#_Toc501740684)

[data\csblack.jpg 7](#_Toc501740685)

[data\ir1-aminobutane.jp 8](#_Toc501740686)

[data\irethanol.jp 9](#_Toc501740687)

[data\irethylethanoate.jp 9](#_Toc501740688)

[data\irpropanone.jp 10](#_Toc501740689)

[data\irlacticacid.jp 10](#_Toc501740690)

[Critical Reflection 11](#_Toc501740691)

[References 11](#_Toc501740692)

[Appendices 12](#_Toc501740693)

# Introduction

Infrared (IR) Spectroscopy uses infrared light to analyse the bond types within a molecule and creates a line graph, for example Appendix 1, of the transmittance of different wavelengths of infrared light. A spectroscopy of nothing would give a line of 100% transmittance because there is nothing to stop the light, bonds in a compound absorb this light and lower the transmittance at certain wavelengths this creates downward peaks in the graph. Each bond types peak will always appear in a similar position, knowing these positions it is possible to figure out the bond types in a molecule and therefore the functional group.

A functional group is “a group of atoms responsible for the characteristic reactions of a particular compound” (Oxford Dictionaries, n.d.). Therefore, functional groups are an important part of chemistry and knowing how the molecule will react. The system will be used to calculate the functional group of a molecule from an inputted IR Spectroscopy graph.

# Literature review

## Fuzzy Inference Systems

Fuzzy inference systems (FIS) they help create a computer with a more human understanding. FIS can be used to determine the degree of membership to a set and then corroborate many inputs to logically find an output. Consequently, fuzzy inference systems can take on a more human understanding of a problem. The similarity to human thought enables fuzzy systems to be coded to more complex problems that perhaps are difficult to solve mathematically.

Fuzzy inference systems are used for many purposes “They can be helpful to achieve classification tasks, offline process simulation and diagnosis, online decision support tools and process control” (Guillaume, 2001). Both classification tasks and online decision support tools relate to a human understanding, where an output requires understanding of continuous data in respect to rules. Process control and simulation may not require a human understanding, process control may be very mathematical, however a fuzzy system is also more readable to a human and therefore for more complex ideas can be simpler to code and easier to debug.

## IR Spectroscopy

IR spectroscopy is a tool used by chemists to determine the bonds present in a molecule and therefore the functional group. Classification systems are currently being used to help identify a compound from data analysis. In Mobaraki’s Structural characterization of carbonyl compounds by IR spectroscopy and chemometric data analysis (Mobarak & Hemmateenejad, 2011) MATLAB is used to create a similar classification system that outputs the functional group. Fuzzy logic is ideal for classification of functional groups because it allows for the input of continuous data that it analyses and outputs a prediction. Ideally a system that classifies functional groups would also use machine learning, to ensure more accurate outputs, it is unlikely the system to be created will perfectly identify functional groups. machine learning would allow minute changes to the system each iteration over a number of iterations a human could not complete.

# The system

## General Approach

For the system to return a functional group, it must perform two key tasks: be able to determine if a given bond peak is present on the graph, be able to correlate these peaks to functional groups. Peaks can not be defined as there or not, small peaks could either be a weak signal from the bond or interference, in addition many bonds peaks share the same wavelength range and can interfere with each other. To aid in identification peaks are often defined as being sharp, broad or even very broad, my system uses the differences in the highest and lowest points of the peak to determine sharper peaks and an average to determine broad peaks. A fuzzy inference system will be used to match the peaks to functional groups. The system takes the strength of five peaks as input and gives five outputs, each output is the percentage of how much the given graph relates to the ‘ideal’ graph for each functional group.

## Graphs Analysis

My system identifies five different peaks from a graph by separating the image down to the range the peak is expected to be in, analysing the peak strength of the range and then feeding the results in to the inference system.

If a sharp peak is expected the strength of the peak takes the value of the lowest transmittance subtracted from the highest transmittance, effectively the difference of the min and the max across the range, a taller peak will give a higher strength. This method allows for broad peaks across this range to be ignored, a broad peak will be consistently low, will not have a high point, and therefore the difference of the min and max will be small.

If a broad peak is expected the strength of the peak is equal to the average transmittance across the range. This method ensures a sharp peak is discounted, because although a sharp peak may have a low minimum transmittance, the average across a broad range is still be high.

### Input

The input for the graph analysis is a graph, Appendix 1 shows a typical IR spectroscopy graphs for 1-aminobutane, an example of an amine. Currently the graphs are very difficult to analyse using code, so the inputs will be cropped and stretched to 4000 x 100 to match the axis of the graph. The inputs to the system will look more like Appendix 10.

### Output

The output for the graph analysis is an array of five values representing the strength of five common peaks:

The single bond between Carbon and Oxygen (CO) with a range of 1000 – 1300 with a sharp peak.

The double bond between Carbon and Oxygen (C//O) with a range of 1640 – 1900 with a sharp peak.

The single bond between Oxygen and Hydrogen appearing in carboxylic acids (OH\*) with a range of 2500 – 3300 and a very broad peak.

The single bond between Nitrogen and Hydrogen (NH) with a range of 3200 – 3500 with a sharp peak.

The single bond between Oxygen and Hydrogen (OH) with a range of 3200 – 3550 with a broad peak.

These specific ranges have been taken from the OCR data sheet (OCR, 2009) for their suitability to the project. The strength of each range is given on a scale of 0 – 100 with 0 being no peak and 100 being the highest possible peak.

## Fuzzy Inference System

The fuzzy inference system assesses the peaks that appear and how strong they are and gives a percentage for each functional group output as to how certain it is that functional group is present.

### Input

The input of the fuzzy inference system is the output of the graph analysis, each of the five outputs has its own input. Appendix 16 shows each of the graphs for all inputs. Each graph for the input uses the same set of membership functions: a low and a high strength function. Both membership functions use a gaussian function, gaussian functions never fully reach 0 similarly even the smallest peak could still be the sign of a given bond. The functions are not symmetrical, the low function having a deviation of 40 and high having a deviation 25, this means medium sized peaks, for example one with a strength of 50, are more low than high, but still count as slightly high. Only two functions are needed here because a peak can either be present or not, it can be less likely to be present or more likely to be present, but a peak can not be both present and not.

### Rules

Refer to Appendix 15 for a list of all the rules as printed by MATLAB. Each functional group gets three rules in a similar format.

Rule 1: if the strength of ALL the relevant peaks is high then the chance of the functional group is high. This rule uses the ‘and’ operator to ensure all the peaks are high; a functional group must contain all of the relevant peaks in order to be present.

Rule 2: if the strength of ANY of the relevant peaks is low then the chance of the functional group is low. This rule uses the ‘or’ operator because if any one of the relevant peaks is low then the functional group is unlikely to be present.

Rule 3: if the strength of ANY peak not relevant to the functional group is high then the chance if the functional group is low. This rule has a weighting of 0.5 this is because although its less likely to be a functional group if other peaks are high, its not necessarily true. For example, ketones and esters both expect a high C//O peak, but if the CO peak is also present its more likely to be an ester, however it is possible to have both the ketone and ester functional group on the same molecule. Therefore, although its less likely to be a ketone, it is not impossible.

The rules are based on common organic chemistry knowledge: Amines contain only an NH bond, Alcohols contain both a CO and an OH bond, Esters contain both a C//O and a CO bond, Alkanes contain only a C//O bond, and carboxylic acids contain a C//O, CO and an OH bond.

The system uses a centroid defuzzification method to find the centre of the curve and return a more accurate result.

### Output

The fuzzy inference system outputs the certainty of a functional group being present on a scale of 0 to 100 with higher being more certain. Appendix 17 shows the graphs for each output, like the input graphs each output has two membership functions a high and a low. This time the high takes up more space in the graph to give more accuracy when there is a high chance of the functional group being present. The chances when the group is present is more note worthy than when the group is not. In addition, it is also likely that some functional groups get triggered falsely and so comparing the relative percentages gives a clearer idea about which functional groups are present and which are side effects of different bond types.

## The system implementation

The system itself contains a single MATLAB script and five MATLAB functions, there is also a data folder containing nine jpg images for testing. The system is run through Main.m and will display two graphs containing the inputs and outputs, and the results of the given test data. One test is commented out for reasons explored in experimental design and evaluation, to view this test the ‘%’ must be removed from in front of the line.

Main.m contains the code used to call the analysis for testing, it contains a matrix of the testing data. The code simply loops through the matrix and runs the inference system for each input.

Analysis.m calls the graph analysis function and the fuzzy inference system itself. The function receives an image as an input, separates it in to the different ranges and analyses each peak. The peak data is stored in peaks and passed to the fuzzy inference system, the outputs of which are neatly printed out using several ‘fprintf’ statements.

rangeMin.m searches through the given range and returns a strength value dependant on the given input. Either through the average, the range or a combination of both.

typeStrength.m is the fuzzy inference system itself, it sets up the inputs, the outputs and rules then the peak strengths are passed to the system for evaluation.

newIn.m and newOut.m are functions to create new inputs and outputs to keep typeStrength.m tidy and nonrepetitive.

# Experimental Design and Evaluation

System testing requires both real life data, to test the system works as expected when given an actual set of data, and custom extreme cases, to test the system when given cases that won’t be given in typical use. It is important to test the system thoroughly to solve any potential issues before handover. Each of the nine test cases with explanations are below:

## data\cslow.jpg

Input:

Appendix 6 – data\cslow.jpg This data is a solid one-pixel thick line towards the bottom of the image.

Reason for test:

This test tests the output of the system when the graph is low, it ensures that the formula for working out broad peaks is correct. We need to ensure the graph analysis is correct before testing for the functional groups.

Expected Output:

The output for bonds CO, C//O and NH are all expected to be 0, there is no difference between the highest and lowest points in a one-pixel thick line. OH\* is expected to be high, this peak is a very broad peak therefore is found using the average, the average should be consistently low therefore the inverse should be high. OH is a broad peak and so is found using both the average and the difference, this should mean this peak is between OH\* and the other peaks.

Actual Output:

current file: data\cslow.jpg

CO = 0 C//O = 0 OH\* = 88 NH = 0 OH = 44

Alcohol = 12% Ester = 12% Carboxylic Acid = 12% Amine = 12% Ketone = 12%

Explanation:

The results were as expected, it is interesting however to note the functional groups are all set to 12, this is believed to be partially the use of the ‘centroid’ defuzzification method and the use of gaussian functions, even if the value is extremely small it still returns a non-zero value calling the rules and giving the output a value. The OH group was exactly half of the OH\* group, although was thought of before testing this is as expected, the OH weights the average and difference equally so should be exactly between a group using only difference and the OH\* group that uses only the average.

## data\cshigh.jpg

Input:

Appendix 7 – data\cshigh.jpg This data is a solid one-pixel thick line towards the top of the image.

Reason for test:

This test tests the output of the system when the graph is high, it ensures that the formula for working out broad peaks is correct. In addition to the previous test, these tests ensure the graph analysis is correct before testing for the functional groups.

Expected Output:

The output should be similar to the previous test, however OH\* and OH should be lower. It is important that the same trend carries across from the previous test, with the CO, C//O and NH bonds being exactly zero, and the OH bond being exactly half the OH\* bond.

Actual Output:

current file: data\cshigh.jpg

CO = 0 C//O = 0 OH\* = 16 NH = 0 OH = 8

Alcohol = 12% Ester = 12% Carboxylic Acid = 12% Amine = 12% Ketone = 12%

Explanation:

This test produces the outputs exactly as expected. The 16 is the inverse of the average (100 – x = 16) meaning the line of pixels is on the 16th line down.

## data\cswhite.jpg

Input:

Appendix 8 – data\cswhite.jpg This is a graph with no dark pixels, all pixels are entirely white.

Reason for test:

To test the output of an empty test.

Expected Output:

The system uses max(vectorOfDarkPixels) if there are no dark pixels it is expected this function will either return zero or null, zero will cause all further values to be zero and the output for every output to be zero, null will cause the system to crash.

Actual Output:

current file: data\cswhite.jpg

Improper assignment with rectangular empty matrix.

Error in Analysis (line 14)

peaks(i) =

rangeMin(in(1:100,(4000-ranges(i,2):(4000-ranges(i,1)))),

ranges(i,3));

Error in Main (line 17)

rules = Analysis(imread(inputs(i,:))); %calls analysis with

the file given

Explanation:

The result was an error, for this reason the input has been commented out in Main.m it can be uncommented to see the effects if needed. However, if an empty file is input to the system no useful information could be drawn from it, effectively the crash is acceptable and tells the user of an improper input in the unlikely case a similar file is used as an input.

## data\csblack.jpg

Input:

Appendix 9 – data\csblack.jpg This is the complete opposite of the previous input with a graph with only dark pixels, all of the pixels are entirely black.

Reason for test:

To test the output when the graph is spread out, not all lines will be exactly one pixel high and it is imperative this does not crash the system.

Expected Output:

The output should not crash, but because the graph has both high and low pixels in every region it will have high outputs for the CO, C//O and NH bonds at exactly 100, OH\* at the average of exactly 50 and OH at the midpoint of 75.

Actual Output:

current file: data\csblack.jpg

CO = 99 C//O = 99 OH\* = 50 NH = 99 OH = 74

Alcohol = 54% Ester = 60% Carboxylic Acid = 35% Amine = 65% Ketone = 65%

Explanation:

The prediction is wrong here, the code works out peaks as the min (1) subtracted from the max (100) the minimum was ignored. This is as designed, however was not considered when modelling a prediction.

## data\ir1-aminobutane.jpg

Input:

Appendix 10 – data\ir1-aminobutane.jpg – This is a modified version of Appendix 1, changed to allow for input to the system.

Reason for test:

To test the system on real life data, and ensure the system outputs a reasonable output for an amine.

Expected Output:

The input is a graph of an amine, the NH peak is expected to be high and the Amine output is expected to be higher than any other functional group.

Actual Output:

current file: data\ir1-aminobutane.jpg

CO = 59 C//O = 56 OH\* = 52 NH = 99 OH = 78

Alcohol = 44% Ester = 41% Carboxylic Acid = 37% Amine = 66% Ketone = 44%

Explanation:

The amine group is more than 20% higher than any other output, this is as expected and shows the system working for an amine input. Although the other functional groups do have a reasonable percentage, the amine percentage is considerably higher than any other.

## data\irethanol.jpg

Input:

Appendix 11 – data\irethanol.jpg – This is a modified version of Appendix 2, changed to allow for input to the system.

Reason for test:

To test the system on real life data, and ensure the system outputs a reasonable output for an alcohol.

Expected Output:

The input is a graph of an alcohol, both the CO and OH peaks are expected to be high and the Alcohol output is expected to be higher than any other functional group.

Actual Output:

current file: data\irethanol.jpg

CO = 89 C//O = 26 OH\* = 59 NH = 31 OH = 88

Alcohol = 68% Ester = 14% Carboxylic Acid = 14% Amine = 17% Ketone = 14%

Explanation:

This is almost the ideal output for an alcohol, all other outputs are very low, there is a small amount of interference from the OH group overlapping the OH\* group, however the systems rules mean the carboxylic acid output won’t be high unless all the inputs are high (CO, C//O and OH\*).

## data\irethylethanoate.jpg

Input:

Appendix 12 – data\irethylethanoate.jpg – This is a modified version of Appendix 3, changed to allow for input to the system.

Reason for test:

To test the system on real life data, and ensure the system outputs a reasonable output for an ester.

Expected Output:

The input is the graph of an ester, so the CO and C//O peaks are expected to be high. This could cause the ketone output to also be high, but the ester output should be the highest.

Actual Output:

current file: data\irethylethanoate.jpg

CO = 90 C//O = 76 OH\* = 31 NH = 62 OH = 49

Alcohol = 35% Ester = 63% Carboxylic Acid = 16% Amine = 49% Ketone = 58%

Explanation:

The Ester the highest, but only slightly higher than the ketone output, this is to be expected, because ketones share a subset of the bonds esters do. The result, although perhaps misleading is still accurate with the highest being the ester output.

## data\irpropanone.jpg

Input:

Appendix 13 – data\irpropanone.jpg – This is a modified version of Appendix 4, changed to allow for input to the system.

Reason for test:

To test the system on real life data, and ensure the system outputs a reasonable output for a ketone.

Expected Output:

The input is the graph of a ketone; therefore, it is expected the C//O bond will have a high peak and the ketone output will be the highest output.

Actual Output:

current file: data\irpropanone.jpg

CO = 10 C//O = 80 OH\* = 7 NH = 25 OH = 20

Alcohol = 12% Ester = 12% Carboxylic Acid = 12% Amine = 14% Ketone = 57%

Explanation:

This is almost the ideal output for a ketone, the C//O bond is higher than any other, and the ketone output is significantly higher than any other, with all the rest being negligible.

## data\irlacticacid.jpg

Input:

Appendix 14 – data\irlacticacid.jpg – This is a modified version of Appendix 5, changed to allow for input to the system.

Reason for test:

To test the system on real life data, and ensure the system outputs a reasonable output for an alcohol and a carboxylic acid.

Expected Output:

The system should output alcohol and carboxylic acid as highest, however ester and ketone may also be very high because they share the same bonds.

Actual Output:

current file: data\irlacticacid.jpg

CO = 69 C//O = 64 OH\* = 54 NH = 12 OH = 52

Alcohol = 39% Ester = 53% Carboxylic Acid = 41% Amine = 12% Ketone = 48%

Explanation:

Unfortunately, here ester and ketone are higher than alcohol and carboxylic acid, this is incorrect however there is very little that can be done with this information alone, ideally the IR spectroscopy would be used in conjunction with another analysis method. Despite the incorrect groups being shown, the system correctly identified the four present bonds and the non-present NH bond.

# Critical Reflection

All tests involving one functional group passed often by a large margin, for these tests it could be said the system worked perfectly. The results for propanone and ethanol were particularly promising with the other functional groups having a negligible similarity. This would suggest for more simple molecules the system does work as intended. When testing lactic acid, it was found that the system thought other groups were also present; despite not being accurate with groups, the system found the bonds that were present very accurately. IR spectroscopy is one of many techniques used in unison to determine the functional group present in a molecule, perhaps with the inputs of the other techniques a more accurate system could be designed.

The system could also be extended to allow for more bonds and groups to be added.

In respect to bonds the CO bond could be split up for different organic compounds similarly to how OH and OH\* have been. Separating the peaks like this will improve the accuracy of the system where the similar input peaks would differ from each other. This could be achieved by changing the peaks matrix to include the new bonds and edit the old ranges, a new input would need to be added in the fuzzy inference system itself this would need to call the newIn.m function. The system passes matrixes around and has functions for creating new inputs, so the changes would be minimal for a significant difference in the system.

In respect to groups amides could be included, a molecule including both the C//O bond and the NH bond. This would allow the system to identify more types of functional group and inherently be more useful to chemists. The implementation of this involves adding a new output using the output function and adding new rules to identify the new compound.

# References

Guillaume, S., 2001. Designing fuzzy inference systems from data: An interpretability-oriented review. *IEEE Transactions on Fuzzy Systems,* 9(3), pp. 426-443.

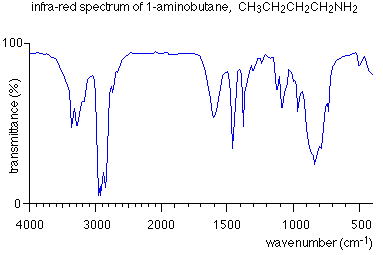
Mobarak, N. & Hemmateenejad, B., 2011. Structural characterization of carbonyl compounds by IR spectroscopy and chemometrics data analysis. *Chemometrics and Intelligent Laboratory Systems,* 109(2), pp. 171-177.

OCR, 2009. *OCR data sheet.* [Online]   
Available at: http://www.ocr.org.uk/Images/74947-datasheet.pdf

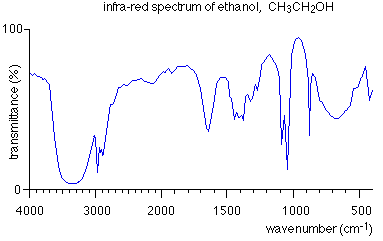
Oxford Dictionaries, n.d. *Functional Group.* [Online]   
Available at: https://en.oxforddictionaries.com/definition/functional\_group

# Appendices

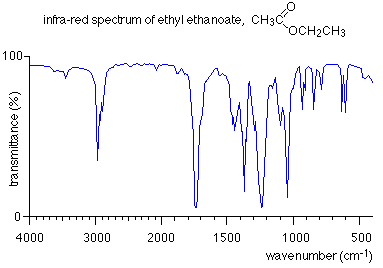
Appendix 1 – Example of IR spectrum results for 1-aminobutane



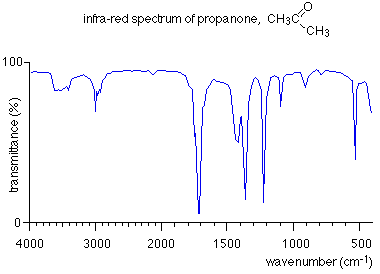
Appendix 2 – Example IR spectrum of ethanol



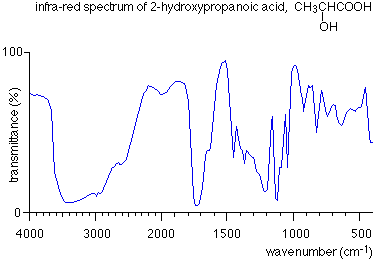
Appendix 3 – Example IR Spectrum of ethyl ethanoate



Appendix 4 – Example IR spectrum of propanone



Appendix 5 – Example IR spectrum of lactic acid (2-hydroxypropanoic acid)



Appendix 6 – data\cslow.jpg – a graph showing a low transmittance across all frequencies

Appendix 7 – data\cshigh.jpg – A graph showing a high transmittance across all frequencies

Appendix 8 – data\cswhite.jpg – An empty graph of only white pixels

Appendix 9 – data\csblack.jpg – An empty graph of only black pixels

Appendix 10 – data\ir1-aminobutane.jpg – A modified graph for 1-aminobutane

Appendix 11 – data\irethanol.jpg – A modified graph for ethanol

Appendix 12 – data\irethylethanoate.jpg – A modified graph for ethyl ethanoate

Appendix 13 – data\irpropanone.jpg – A modified graph for propanone

Appendix 14 – data\irlacticacid.jpg – A modified graph for lactic acid

Appendix 15 – A list of all the rules as printed by MATLAB

1. If (CO is high) and (OH is high) then (Alcohol is high) (1)

2. If (CO is low) or (OH is low) then (Alcohol is low) (1)

3. If (C//O is high) or (OH\* is high) or (NH is high) then (Alcohol is low) (0.5)

4. If (CO is high) and (C//O is high) then (Ester is high) (1)

5. If (CO is low) or (C//O is low) then (Ester is low) (1)

6. If (OH\* is high) or (NH is high) or (OH is high) then (Ester is low) (0.5)

7. If (CO is high) and (C//O is high) and (OH\* is high) then (Carboxylic Acid is high) (1)

8. If (CO is low) or (C//O is low) or (OH\* is low) then (Carboxylic Acid is low) (1)

9. If (NH is high) or (OH is high) then (Carboxylic Acid is low) (0.5)

10. If (NH is high) then (Amine is high) (1)

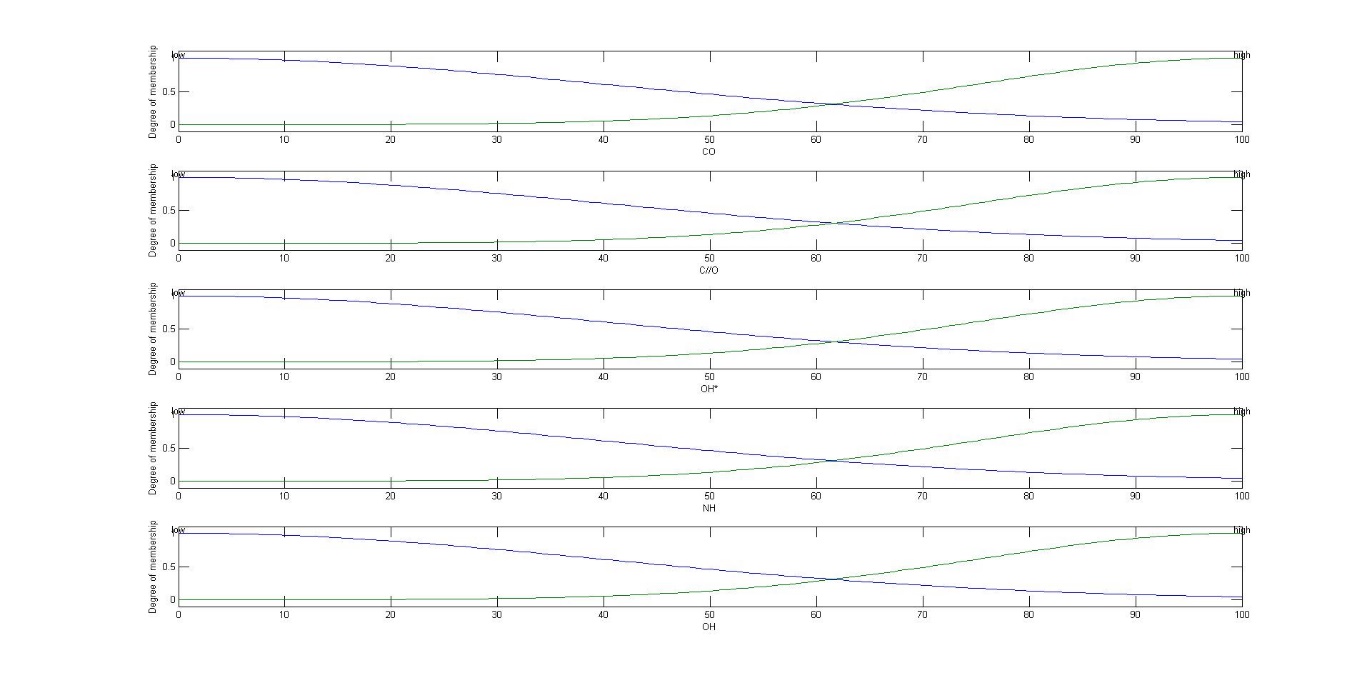
11. If (NH is low) then (Amine is low) (1)

12. If (CO is low) or (C//O is low) or (OH\* is low) or (OH is low) then (Amine is low) (0.5)

13. If (C//O is high) then (Keytone is high) (1)

14. If (C//O is low) then (Keytone is low) (1)

15. If (CO is low) or (OH\* is low) or (NH is low) or (OH is low) then (Keytone is low) (0.5)

Appendix 16 – The graphs for the inputs

Appendix 17 – The graphs for the outputs

