In-Situ Raman Spectroscopy Component Identification for Machine Learning Based Decomposition Analysis

Brandon Kern[2], Elizabeth Rasmussen[1,3], Jon Onorato[2,3], Parker Steichen[2]

[1] Department of Mechanical Engineering, [2] Department of Material Science and Engineering, [3] Clean Energy Institute
University of Washington, Seattle, WA, USA

DIRECT
Data Intensive Research
Enabling Clean Technologies

CLEAN ENERGY INSTITUTE
UNIVERSITY of WASHINGTON

eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

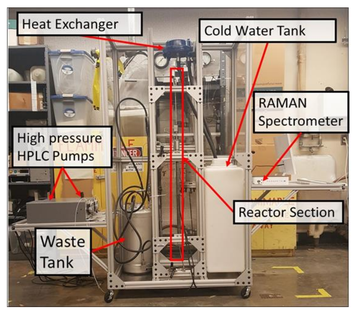
Introduction

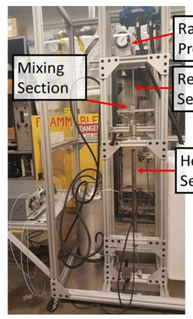
- Using data from a custom built supercritical gasification reactor on campus to analyze formic acid Raman spectra.
- Decomposition of formic acid constitutes the combination of two pathways:

$$HCOOH \rightarrow H_2 + CO_2$$

$$HCOOH \rightarrow H_2O + CO$$

Photographs of UW Gasification Reactor with In-Situ Raman Probe





Motivation

Prior Analysis Method

- Commercial Software suite with yearly license cost of ~ $3,000
- Not Open Source → Research backbone can be out of date
- Very tedious manual processes

PEAXACT 4

MATLAB

Raman Noodles

New Method!

- Free
- Open Source + Customizable → Research backbone easily updated by anyone
- Automated processes

jupyter

python™

GitHub

Goals

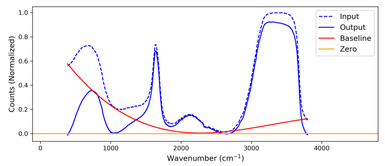
1. Data Mining and Baseline Subtraction
   - Importing open source data sets, create a library of spectra, uniformly format data for analysis
2. Data Visualization
   - Outputting plots of baseline subtraction and peak identification
3. Machine Learning
   - Prepare least squares regression model for calculating kinetic rate decomposition at different resonance times and temperatures

Materials and Methods

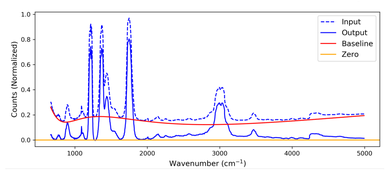
Baseline Subtraction of Raman Signal

Used PeakUtils built in baseline function to perform polynomial fit baseline subtraction on NIST database spectra. Examples of this functionality are shown below.

Water Baseline Subtraction Output



Acetone Baseline Subtraction Output



Component Analysis in a Mixture Raman Signal

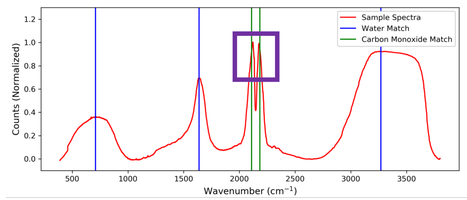
The LMFit package was utilized to identify 5 descriptors per peak in a mixture's Raman signal including: location of the peak, peak height, and peak width. As a mixture has more peaks (components from decomposition) the amount of descriptors increases.

Least Squares Model Equation

$$f(x) = \sum_{i=1}^{n} f_i(x, A, \mu, \sigma)$$

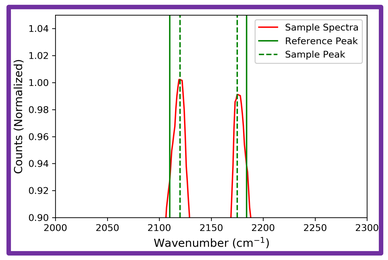
Lorentzian Equation for a Single Peak

$$f_i(x, A, \mu, \sigma) = \frac{A}{\pi}\left[\frac{\mu}{(x-\mu)^2+\mu^2}\right]$$



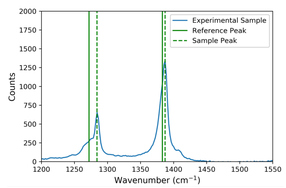
Component Confidence based on Euclidean Distance

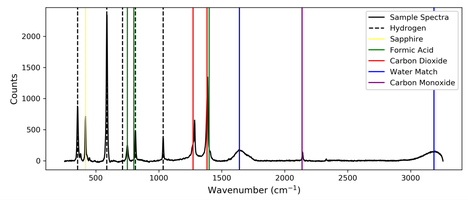
Once components in a 'testing' dataset mixture are identified the next step is to take a 'training' dataset location and share with the user the Euclidean distance between the two datasets. For this software if a peak location is more then ± 10 cm^{-1} from the literature values the confidence that the peak represents the compound is zero. This range was set from experimental considerations.



Results Using Experimental Data

After creating a theoretical NIST 'training' datasets and proving the functionality of the code the final step was to test it on experimental data sets taken in the lab.



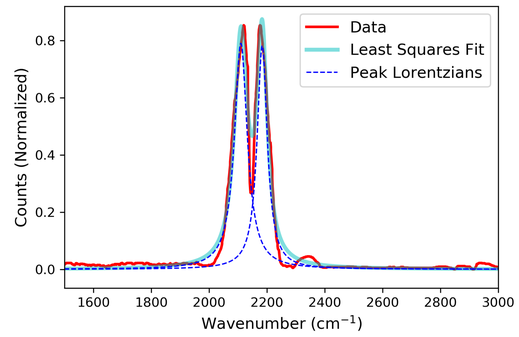


Continuing Work

Machine Learning for Material Decomposition

- We implemented functionality to run a least squares regression that fits Lorentzian curves to the data. The function is given the peak locations determined using scipy.signal.find_peaks.
  1. Expand software to be able to compute decomposition rates across varying parameters such as temperature, resonance time, possibly pressure.
  2. From the defined decomposition rate the software can predict the decomposition rates using machine learning beyond the known data set limits

Carbon Monoxide Spectra Lorentzian Output



Conclusions and Future Work

In conclusion our team successfully created a fast functioning open source code base that saves hours of research time in data cleaning and analysis of Raman Spectra. We have also set a strong base for the next step of our focus which is on calculating decomposition of substances using Lorentzian peak information that will be applied to machine learning optimum temperatures and pressures in a gasification reactor system.

Future work

This work sets up a free and user friendly platform for researchers to be able to analyze their own Raman Spectra.

Acknowledgements

- Dave Beck, Chad Curtis, and Kelly Thornton
- Data sets were taken from publicly available from the NIST WebBook Database and Mendeley Data, "Raman Spectra of Formic Acid Gasification Products in Subcritical and Supercritical Water"
- Only open source packages were used in this work, documentation of all packages used can be found at our GitHub at:
  - https://github.com/raman-noodles/Raman-noodles